8-30-2023 9:00 AM

State-of-the-art Approaches for Sequencing, Assembling and Annotating Naphthenic Acid Degrading Bacterial Metagenomes

Henry H. Say, *Western University*

Supervisor: Gloor, Gregory B., *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Biochemistry

Abstract

Naphthenic acids (NAs) are the main toxic component of oil refinery wastewater and require special processes to be removed. Harnessing bacterial biodegradation for NA removal has the potential to be effective, yet NA-degrading bacteria and pathways are poorly understood and uncharacterized. To improve our understanding of NA degradation, I characterize the metagenomes of novel NA-degrading bacterial communities seeded in NA-enriched granulated activated carbon (GAC) filters. I demonstrate methods that maximize the throughput of extraction, sequencing, and annotation of novel metagenomes - producing 72 MAGs and other 5432 circular contigs - 226 of which were putative phages. I also include state-of-the-art protein structure prediction and structure homology search tools, which greatly enrich annotations of novel sequences that are below the threshold for homology finding by sequence alone. Overall, these approaches unveiled a diverse and constantly changing consortium of novel bacteria and many potential NA-degrading genes.

Keywords

Metagenome, Oil refinery, Wastewater, Annotation, Structure Prediction, Naphthenic Acid, Bacteria, Bacteriophage, Structure Homology, Granulated Activated Carbon, Tailings, Nanopore, Sequencing,

Summary for Lay Audience

The amount of toxic wastewater accumulated by the oil industry has increased exponentially since it began its operations, and there is no cost-efficient, sustainable, yet effective way of reclaiming this wastewater. The major challenge is naphthenic acids (NAs), which are a natural byproduct of crude oil refining. NAs are complex, difficult to remove without special processes, and highly toxic. In wastewater treatment, however, there is potential for harnessing bacteria that can degrade NAs. These bacteria have been observed before, but there is no known bacteria or community that degrades the full range of NAs efficiently, and not many NA-degrading genes are known. The study of NA-degrading bacteria and how they remove NA will be foundational to creating future bioengineer wastewater treatment systems. In this thesis, I characterized NA-degrading microbial communities living in granulated activated carbon (GAC) filters from an oil refinery wastewater treatment collected over time. I first demonstrated how to collect DNA from these samples and sequence the DNA using Nanopore – a state-of-the-art DNA sequencing technology. This technology allowed me to develop methods to reconstruct entire bacterial genomes from GAC, and I observed a highly diverse bacterial community that is constantly changing over time and is mostly composed of bacteria never sequenced before. Also, I annotate these bacterial DNA sequences to unveil their biological capabilities and discover several genes likely related to NA degradation. Since these bacteria are novel, however, there were still a lot of uncharacterized DNA sequences that could be important. I tested a new annotation strategy on a small subset of viruses in the GAC community called bacteriophage, where I identified unknown sequences by predicting 3D models of the DNA products, the proteins, and compared them against other known 3D models of proteins. This approach drastically improved the ability to identify what the viruses are biologically capable of since protein structure comparisons are generally more reliable than sequence comparisons when inferring the functions of proteins. Though more work is needed to confirm NA-degrading bacteria and genes, this thesis sets a foundation for future analysis of NA-degrading bacteria and pathways.

Co-Authorship Statement

Henry Say performed the research and analyzed the data for the works presented in chapters 2, 3 and 4, with help from authors noted below. Dr. Gregory Gloor and Dr. Martin Flatley conceived the project.

Chapter 2

Dr. Daniel Giguere and Dr. Ben Joris helped conceive the assembly and polishing method. The DNA extraction methods for GAC are based on a DNA extraction protocol created by Daniel Giguere. Ben Joris conceived the secondary assembly pipeline.

Chapter 3

Dr. Gregory Gloor provided input for the analysis.

Chapter 4

Dr. Gregory Gloor conceived the idea to use Colabfold and Foldseek and provided input for the analysis.

Acknowledgments

I would like to thank my supervisor Dr. Gregory Gloor for giving me a shot and introducing me to bioinformatics, a field I'd never heard of until the end of my undergraduate years. When I first joined your lab in my 4th year of undergrad, I never touched a command line in my life. I have learned so much since then in both the wet and dry lab and feel like I was given an opportunity that most people do not come by. I am grateful that you continued to believe in me and provided me with the opportunity to work on this project. It was a pleasure working in your lab and I could not be thankful enough for everything.

I want to thank Dr. Daniel Giguere, who mentored me in the beginning when I first joined Dr. Gloor's lab. His enthusiasm and patience really helped me get through the initial learning curve of Nanopore sequencing and easing me into the world of bioinformatics. I am sorry for the flow cells that did not make it in my undergrad (R.I.P).

Thank you to Dr. Ben Joris, who also helped me along my journey. His work contributed a lot to the success of this project, as his pipeline helped me gather more cool genomes to look at.

Thank you to Dr. Martin Flatley for organizing and making time for the sample collections. You made this project possible in many ways and I appreciate what you have done.

Thank you to my committee members, Dr. David Edgell and Dr. Ken Yeung for your support. I want to thank Dr. David Edgell particularly and everyone in his lab for letting me work in and around your space when I needed to. I was genuinely happy to help with any Nanopore sequencing needs, and I wish everyone the best in their research and future.

I couldn't finish without thanking my parents, who fought so hard and sacrificed so much to give me the life I have today. Last but not least, thank you Ashley. You've done everything and more for me in the past 6 years and I often feel like I can attribute you to all my accomplishments. If all my experiments blew up in my face, I'd still feel like the luckiest person on earth.

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1

1 Introduction

The oil industry is a multi-trillion-dollar industry and plays a crucial role in the global economy and energy needs. Unfortunately, oil has significant environmental impacts that begin from the extraction and processing of crude oil to its usage as an energy source. Much progress has been made in terms of addressing this issue as alternative clean energy sources become increasingly adopted and collective efforts have been made to implement green operating regulations in the oil industry (1,2). However, significant progress can still be made to reduce the production of oil refinery wastewater, which remains an industry-wide challenge.

1.1 The Substantial Challenge of Naphthenic Acid Removal in the Oil Industry

The process of refining crude oil consumes vast amounts of freshwater and outputs toxic wastewater at a rate that outpaces current standard practices to recycle or reclaim this wastewater from its major toxic components (2,3). Oil refinery wastewater (ORW) is extremely heterogeneous, and its components can vary in content and concentration, but consists of inorganic and organic components such as hydrocarbons, phenols, nitrogen, sulfur, salts, heavy metals, and other suspended solids. Notably, ORW naturally contains naphthenic acids (NAs) - a broad family of alkyl substituted acyclic or cycloaliphatic carboxylic acids classically represented by the formula $C_nH_{2n-z}O_2$,where z represents the degree of cyclization multiplied by 2 (4). NAs are the main toxic component of ORW and present significant challenges to refineries. Not only do NAs have corrosive properties contributing to equipment wear and tear, highly cyclic or branched NAs are recalcitrant to natural degradation or standard ORW treatment methods, requiring special processes that have high operational costs and are impractical for widespread use (5,6). These treatment methods aim to either degrade NA into less toxic intermediates with advanced oxidation processes, or to simply capture NAs either on a membrane filter, through coagulation, flocculation or granulated activated carbon beds (7). Advanced oxidation processes for example, can risk degrading NAs into more toxic intermediates that cannot

be broken down further (8). Methods that rely on capturing NAs instead allow for the reclamation of freshwater from wastewater to acceptable standards but generate waste in another form as the capture substrate needs to be continually replaced and replenished. Though these methods can be sufficient for individual small-scale refineries, they are unsustainable as the industry produces ORW generation at a rate that necessitates the storage of wastewater in massive reservoirs called tailing ponds, where wastewater can be stored for many years before being recycled or treated. Figure 1 displays the total estimated area occupied by tailings ponds in the oil sands regions in Alberta, Canada, which was 120 square kilometers in 2020. Over the last few decades, the total volume of tailings has only been increasing as the oil sands operations expand (3).

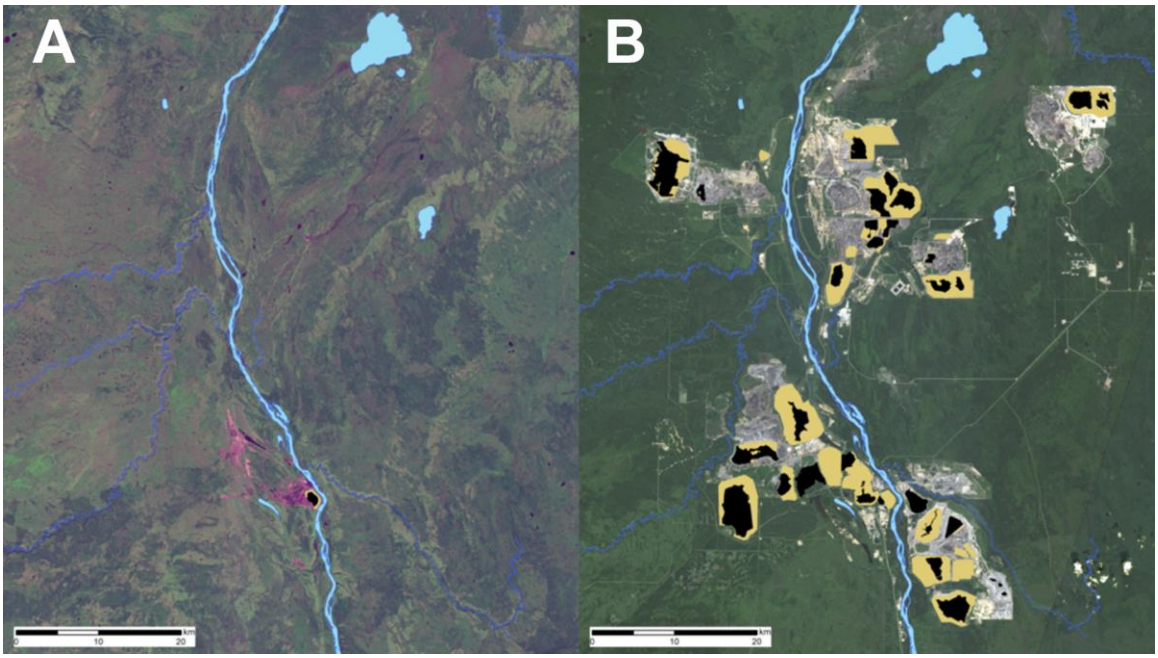

Figure 1. Areas affected by fluid tailings (black) as well as associated tailing features (yellow) at Fort McKay, Alberta, Canada increased from 2.44 km^2 in 1974 (A) to 307.31 km^2 in 2020 (B) (3). Tailing features include dams, berms, end pits that are involved with tailings containment or reclamation.

The storage of ORW in tailings ponds has long-lasting effects to the containment area and is a risk factor for the surrounding environment. Though many sustainable practices have been put in place to limit the environmental impact of ORW outside the

containment area, leakage is the major concern as it could contaminate surrounding ecosystems, and contaminated areas remain toxic for many years since natural aging and biodegradation of NAs is a slow process. NAs are surfactants and if they leak into waterways, soil or groundwater, they cause acute and chronic toxic effects on terrestrial and aquatic wildlife, plants, and microorganisms. It has been reported that NA mixtures from fresh ORW exhibit neurotoxic effects to aquatic and terrestrial organisms through several different mechanisms, leading to deformities and mortality (9–12). These physical properties and toxicity of NAs can also vary according to its structure, where NAs with cyclic and branched features are generally more difficult to degrade due to their greater hydrophobicity and poor bioavailability to NA degraders, while lower molecular weight NAs are more toxic (13,14). Higher molecular weight NAs are less common relative to other NAs, but still play a significant role in the overall toxicity and recalcitrance of NAs (15). Unfortunately, there is still an unmet need for alternative methods to efficiently reclaim process affected water and ultimately eliminate the environmental footprint of crude oil extraction and refining.

1.1.1 Investigating Naphthenic Acid Biodegradation for Wastewater Remediation

One promising strategy of NA removal involves NA-degrading microbes, which would be more scalable, cost-effective, and environmentally friendly as fully biodegraded NAs result in the release of just carbon dioxide, water, and microbial biomass (16). Biodegradation of natural NA mixtures, commercially prepared NAs, or surrogate NAs by indigenous microbial communities have been observed in ORW (6,17–20). Unfortunately, observed natural biodegradation in tailings ponds is a slow process and fractions recalcitrant to natural biodegradation can persist, leaving wastewater in storage for years before seeing significant changes in toxicity (7). One reason for the low rate of biodegradation may be due to the heterogeneity of ORW where NAs are a minor constituent and are typically harder to degrade, such that naturally occurring isolates identified at tailings ponds generally use NAs as a carbon source of last resort. Additionally, the efficacy of biodegradation would be affected by factors that are often changing, such as the composition of the ORW, temperature, pH and nutrient availability

(20). Though some of these factors can be consistent in a facility, the composition of the ORW can differ depending on the petroleum source, leading to significant differences in the composition of the bacterial community across samples collected from different locations or time points.

Ideally, the study of naturally occurring NA bacteria or communities can lead to important discoveries about which species can be better utilized in specially designed biotechnological processes to further improve the efficiency of ORW reclamation from NA. With a properly bioengineered ecological treatment system, NA biodegradation can be a viable solution to reclaiming these large tailings ponds since the rate of biodegradation would scale proportionately to the size or concentration of the NA source it is introduced to. Biological systems would require minimal energy or chemical inputs and would therefore be cheaper. Additionally, these bioengineered systems can be designed to support existing NA degradation methods to improve the rate of NA clearance. In methods that aim to capture rather than degrade NAs, bioengineered NA-degrading bacteria can be used to support NA removal by introducing them into the substrate where NAs are captured and concentrated. The benefits would be twofold: the rate of NA clearance would improve, and less waste would be produced since the substrate could be more readily recycled without having to deal with residual NA contamination. Unfortunately, however, there has yet to be a specific bacteria or combination of bacteria efficient and robust enough to see widespread use in this type of application.

1.1.1.1 Known Naphthenic Acid Biodegraders

Individual bacterial species or communities have been identified to biodegrade mixtures of NA previously, but those that are characterized to degrade individual NAs have only been NA surrogates, including *Acinetobacter anitratum*, *Alcaligenes faecalis* and *Pseudomonas putida*, which were all observed to degrade cyclohexancarboxylic acid (20). Table 1 lists several bacterial strains identified to be capable of degrading NAs. Beyond knowing that these bacteria degrade NAs to an extent by seeing a reduction in the toxicity of an NA sample over time, specific NA-degrading mechanisms, genes, and bacteria are poorly characterized. Furthermore, no single organism or mixture of

organisms has been identified and characterized that can efficiently or fully degrade the full spectrum of NAs (17,18,20). It is more likely that, given the high biodiversity of observed bacterial communities in ORW, multiple bacterial species in a community contributes partially to the full pathway of NA degradation as opposed to the existence of a single species that degrades all NAs.

Table 1 Various examples of NA-degrading bacterial strains identified in the literature. All identified strains listed here degrade NA surrogates.

Strain	NA Surrogate Degraded	Citation
Alcaligenes faecalis	Cyclohexanecarboxylic acid	(21)
Acinetobacter anitratum	Cyclohexanecarboxylic acid	(22)
Aquamicrobium aestuarii	Multiple classical surrogate NAs	(23)
Aquamicrobium terrae	Multiple classical surrogate NAs	(23)
Mycobacterium aurum	(4′-n-butylphenyl)-4-butanoic acid	(24)
	(4′-t-butylphenyl)-4-butanoic acid	
Pseudomonas stuzeri	Multiple classical surrogate NAs	(23)
Pseudomonas vancouverensis	Multiple classical surrogate NAs	(23)
Pseudomonas knackmussi	Multiple classical surrogate NAs	(23)
Pseduomonas putida	Cyclohexanecarboxylic acid	(25)
Pseudomonas turukhanskensis	Multiple classical surrogate NAs	(23)
Sphingopyxis witflariensis	Mostly straight chain NAs and single ring NAs	(23)
Staphylococcus hominis	Multiple classical surrogate NAs	(23)

1.1.1.2 The Current Understanding of Naphthenic Acid Biodegradation is Poor

Studies have been carried out to elucidate NA-degrading pathways as well, but mainly focus on isolates biodegrading NA surrogates rather than individual naturally occurring NAs. This is likely a result of naturally occurring NAs being extremely difficult to isolate from ORW samples, given the heterogeneity of ORW and the low percentage of NAs generally present. By mass, NAs have been observed to comprise up to 3% of crude oil (17). Since other studies have proposed differences in toxicity and degradability between surrogate NAs and naturally occurring NA fractions, it is possible that results from studies that focus on surrogate NA degradation don't fully reflect the true ability of a bacteria to biodegrade NAs in ORW (17,18). Regardless, the current understanding of NA degradation is that NA-degrading organisms mainly use existing fatty acid catabolic pathways, including beta-oxidation, alpha-oxidation and omega-oxidation (21,24,25). Biodegradation of NAs is also thought to be a mainly aerobic process, as rates of NA degradation have been observed to decrease when oxygen levels are low (18).

1.1.2 Scope and Objective

To make significant progress on eventually designing efficient ecological treatment systems, a solid understanding of not only what degrades NAs, but also the genes and pathways involved with NA degradation is fundamental. In this thesis, I am attempting to characterize a NA-degrading bacteria community originating from an oil refinery waste-water treatment plant in Ontario, in which the last step of a multi-step ORW reclamation process is the capture of residual NAs on industrial-scale granulated activated carbon (GAC) beds (Fig. 2).

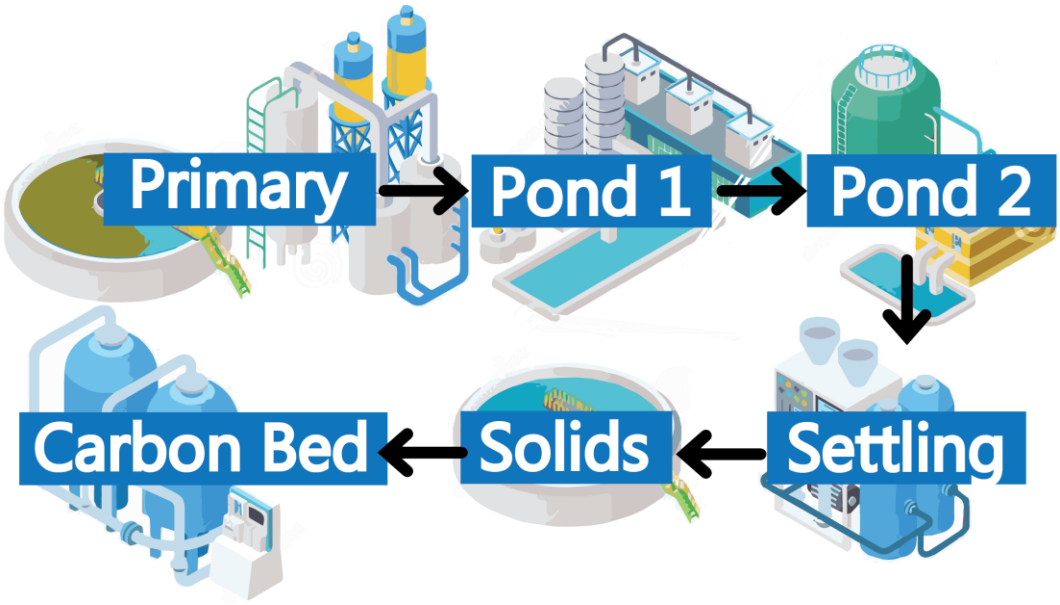

Figure 2. A schematic of the oil refinery wastewater treatment facility from which the GAC samples collected in this thesis originate. The last step in the process are the carbon beds, which contain GAC seeded with a bacterial community. By the last step, most toxic contaminants are removed to safe levels and mostly recalcitrant NAs remain.

This serves as a natural experiment since the is environment highly enriched for naturally occurring NAs, upon which a bacterial biofilm grows where NAs are the sole carbon source. The community is observed to aid in the bioremediation of the NAs, but ultimately the accumulation of biomass reduces filtration capacity and necessitates periodic exchange with fresh GAC. Regardless, the sample of interest has not been studied before, and this presents a unique opportunity to investigate a novel ecosystem for its NA-degrading capabilities to fill some knowledge gaps surrounding NA biodegradation. By sequencing the metagenomes of GAC samples collected over time with next-generation long-read sequencing technologies, I expect to unveil a number of novel organisms capable of degrading NAs, NA-degrading genes, and biodiversity in this ecosystem. This information can be used in synthetic biology applications to design efficient and robust NA ecological treatment systems that address the weakness in efficiency and robustness with naturally occurring NA-degrading bacteria. By discovering NA-degrading genes and pathways across samples, important genes can be

discovered which lay the groundwork for how to engineer a bacterium that can degrade a wider spectrum of NAs, more efficiently and under more conditions. Using long-read sequencing to sequence and analyze GAC metagenomes could reveal NA-degrading genes, which is an approach that has not been utilized often in the literature.

1.2 Characterizing Bacterial Metagenomes with NGS

Long-read sequencing is particularly useful for generating complete metagenomically assembled genomes (MAGS) of difficult-to-assemble species, and this is commonly performed with Oxford Nanopore Technologies' (ONT) Nanopore platform. Nanopore sequencing allows for direct, real-time sequencing of DNA or RNA samples by using an array of protein complexes embedded in a membrane with a resting potential. These protein complexes contain a nanopore, a channel in which one strand of a DNA or RNA molecule is fed through by other proteins responsible for the capture and transfer of DNA across the nanopore (Fig. 3).

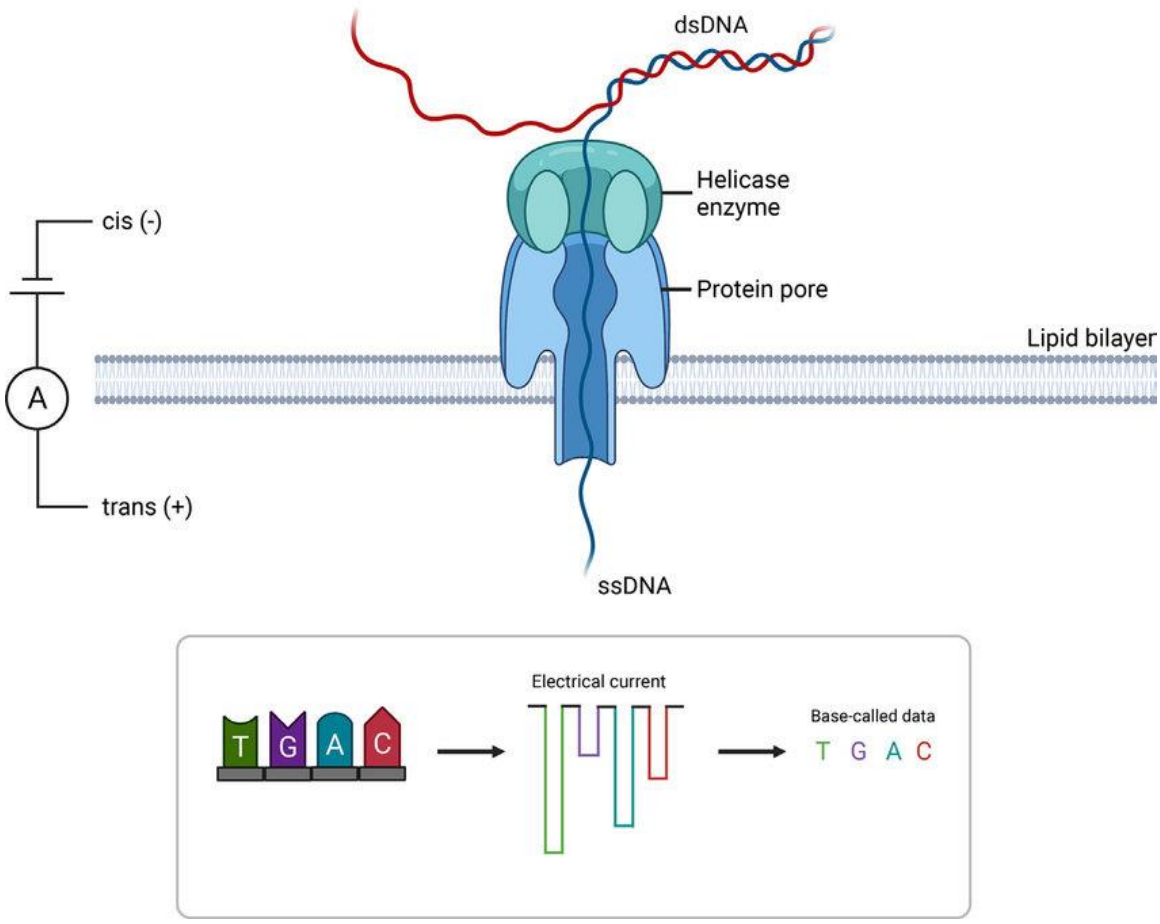

Figure 3. A diagram representing a single DNA molecule being unwound and fed through a nanopore as a single, uninterrupted sequence (26). As the DNA molecule passes through the membrane with a resting potential, characteristic changes in the electrical current occur as a result of the specific base passing through. This electrical current is then translated to basecalled data by basecalling algorithms.

As DNA or RNA molecules pass through, each nucleotide base changes the electric current across the membrane in a characteristic way, allowing for the identification of nucleotide sequences with basecalling algorithms such as Guppy that translate the electric signals (27). The main advantage over other popular short-read sequencing methods, such as Illumina, is that it allows for the collection of more contiguous and complete sequence information, as the sequencing technology behind Illumina imposes technical constraints. Illumina sequencing is a massively parallel sequence-by-synthesis method, in which the prepared DNA fragments are attached to a flow cell and bridge amplified to produce local

clusters of cloned fragments that are used as templates for sequencing-by-synthesis (Fig. 4).

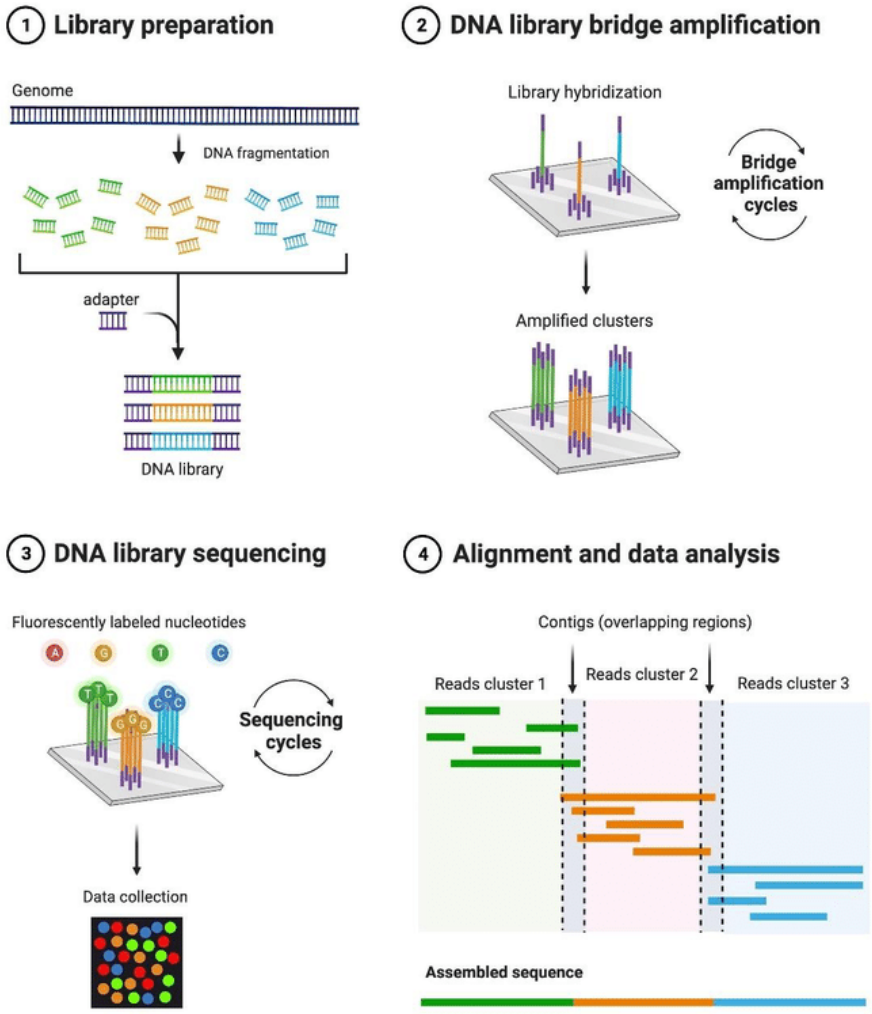

Figure 4. A diagram of Illumina sequencing steps (28). Unlike Nanopore, Illumina requires the initial library to be fragmented prior to attaching its adapters. The adapters are captured onto the flow cell and bridge amplified to produce clusters representing the same sequence. The actual sequencing then occurs in cycles – in which one fluorescently labeled nucleotide is added to each sequence per cycle. The fluorescent signal is measured after each cycle to determine which base comes next in the sequence. The disadvantage of short-read sequencing is that the assembly of contiguous sequences is more difficult given that short-reads contain less overlap

information to confidently determine if two sequences are connected, especially if the read is not long enough to span features such as repeat regions.

The templates are sequenced in cycles, where in one cycle, a base is added to each of the templates in parallel. An optimal cluster produces a clear and detectable signal that is translated into a respective base read, but issues arise when cycles become dephased and molecules in a cluster fail to extend or extend too far. As template length increases, the number of errors can accumulate, and the signal becomes too noisy to be translated reliably (29). Nanopore avoids this read length limitation as DNA or RNA molecules are read as a single, uninterrupted sequence. By virtue of this, the main limitation on read length is the quality of the fragment itself, as nicks in the DNA will interrupt a read and contamination can increase the likelihood of jamming the nanopore. Although Nanopore sequencing has the advantage of producing contiguous sequence information, the read accuracy is generally lower. Read accuracy is measured by quality scores (Q-score) that logarithmically indicate base error probabilities, where ($Q = -10 \log_{10}P$). Raw read quality scores of Q30 and above are typical for Illumina sequencing, but the V10 chemistry based sequencing kits by ONT generally produce raw read scores of Q15. At the time of this writing, however, ONT has updated their kits to V14 with new enzymes, nanopores and updated basecalling algorithms. With their V14 chemistry, ONT claims that raw read accuracies of Q30 or more are possible (30).

1.2.1 Nanopore-Based Metagenomic Assembly

1.2.1.1 Major Considerations

When it comes to assembling high-quality MAGs from metagenomes sequenced with Nanopore, there are several other considerations that must be addressed in parallel. These considerations can drastically impact downstream analyses in a cumulative way and are especially important for environmental samples given their high complexity. First, extracting pure and intact high-molecular-weight DNA is a prerequisite for generating quality, high-length reads with nanopore. For GAC samples in particular, the extraction method must be suitable or optimized for the high concentrations of heavy metals in the sample. GAC is also a physical object that can shear DNA, and a method to isolate DNA

with minimal movement of the granules is necessary. Second, sufficient data must be collected from the sequencing run to be able to characterize species at non-trivial relative abundances. Samples that are extremely biodiverse require much more data to be collected to capture low abundance, but otherwise potentially important species. Third, care must be taken in the assembly to ensure that redundant sequences and low-quality reads are excluded from the assembled genomes. These considerations are important to obtain sufficient sequence information that is also contiguous and high-quality. Without addressing these considerations, the genetic variability and high biodiversity of microbial species expected in environmental samples make it difficult to accurately resolve complete, individual genomes and will convolute any downstream analyses. High read lengths are important as de novo assembly algorithms typically look for overlapping end regions of different DNA fragments to link them into contiguous sequences (31). As such, highly fragmented datasets lead to a variety of issues. A set of reads that are too short to span problematic features such as direct repeat regions may lead to many assemblies that never circularize or are only circularly permuted. Trying to reconstruct complete genomes is also difficult by virtue of not having enough overlap information to confidently determine if two sequences can be linked. A standard way of evaluating the quality of a set of reads in terms of their contiguity is with the read N50 metric (32). The read N50 indicates that half of the total bases sequenced belong to fragments that are at the N50 length or higher, so a dataset with a greater N50 value contains more contiguous information and therefore more accurate reconstructed genomes. Equally important to high read lengths are high-quality reads, which are essential to reduce the number of instances where unrelated sequences are combined to produce chimeric and misleading assemblies. Reads with high base accuracy are especially important since chimeric assemblies are more likely to appear if there are species in the metagenome that are closely related, but not the same.

1.2.1.2 Assembly and Polishing Strategies

Reconstructing genomes from a set of reads can be easily done with a number of tools, each with their own applications. Flye is one de novo assembly algorithm designed specifically for use with long-read nanopore data and for metagenomic assembly and has

been shown to perform better than other popular assemblers such as Canu or Miniasm in producing a greater number of complete metagenomic assemblies with fewer errors and with clean circularization (33). Since nanopore reads are relatively error-prone however, additional consensus-building strategies are employed to increase the confidence of de novo assemblies and improve base accuracy. Drafts initially produced by the assemblers are subject to polishing pipelines in which reads are aligned to the assemblies to generate a consensus and determine if there are any consistencies at each position. Flye includes a post-assembly polishing step - however, post assembly tools such as Racon, Medaka or Minipolish can be used iteratively or in combination for multiple rounds of polishing (34–36). Racon in combination with Medaka was frequently used in the past as a standard for assemblies generated with nanopore data. It is important to recognize, however, that these popular polishers lack intrinsic filtering of alignments, which can be detrimental to an assembly's quality post polish since poorly aligned or low-quality reads could potentially be used to correct base errors. Given the nature of consensus building strategies, it would be ideal use of reads that truly belong to the genome for polishing, otherwise differences between the alignments and assembly may be incorrectly attributed as errors. Popular polishing tools such as Racon or Medaka do not intrinsically filter reads used for polishing, but tools such as Gerenuq can be used to filter alignments based on its alignment score, length and q-score prior to polishing in order to prevent the use of low quality and poorly aligned reads for polishing (34,35,37).

1.2.1.2.1 Secondary Assembly Strategy for Long Read Assemblers

Assemblers like Flye have room for optimization; sequences can usually be assembled and circularize without extra steps, but many fragments can potentially be truly circular when they appear to be linear upon an initial draft. Contigs may fail to assemble and circularize for many reasons, including the heterogeneity of metagenomic datasets and insufficient coverage. Flye does not natively perform iterative assembly strategies unlike short-read assemblers such as IDBA-UD or SPAdes where it has been shown to improve assembly quality beyond the primary assembly, yet Flye stands to benefit from such strategies (38,39). More circularized assemblies could be generated by implementing a

binning strategy in which a second round of assembly is performed on subsets of reads that align to binned uncircularized assemblies produced by the first round. Reads that map end-to-end are included.

1.2.2 Quality Control

Evaluating the quality of resulting assemblies after polishing can be done with additional metrics such as coverage, completeness, and contamination.

1.2.2.1 Completeness and Contamination for MAGs

Assessing the quality of MAGs follows a well-defined set of rules regarding genome gene content and gene redundancy, which are measured by completeness and contamination scores, respectively (40). Completeness and contamination scores are exclusive to genome assemblies and provide information on whether or not a genome has expected gene features. Tools such as CheckM can automatically estimate completeness and contamination scores for expected microbial genomes based on the presence and copy number of single-copy marker genes in an assembly, which are marker genes present in most microbial genomes (41). Completeness is the percentage of expected marker genes present in a query assembly, and contamination is the percentage of foreign DNA and duplicated or fragmented marker genes in an assembly. High-quality MAGs have at least 90% completeness, less than 5% contamination, encode 23S, 16S, 5S rRNAs and encode tRNAs for at least 18 of the 20 amino acids. Furthermore, for a MAG to be considered finished, it also requires a consensus error rate of Q50 or better (40).

Low percentage completeness does not guarantee that an assembly is not a completed genome - candidate phyla radiation (CPR) bacteria genomes, for example, have reduced sizes. CPR bacteria refers to a diverse group of mostly uncultivated bacteria that have been discovered through metagenomic studies of a wide range of ecosystems, including soils, sediments, and aquatic systems (42). CPR bacteria are characteristically missing single-copy marker genes that are considered universal and often have incomplete metabolic pathways. As such, it is common for these species to appear with 60-80% completeness, despite being truly complete (43). As a result of its diminished biosynthetic potential, CPR bacteria typically have obligate symbiotic or parasitic

relationships but nevertheless are believed to have important roles in microbial ecosystems given their high abundance in various microbial communities and diverse gene content contents in accordance with its niche. For example, a group of CPR bacteria collected from aquifer sediment called Zixibacteria was shown to encode genes involved in multiple pathways that are beneficial for a changing redox environment - including fatty acid oxidation, ferric/ferrous iron reduction, anaerobic respiration via nitrite reduction, and fermentation (44). Finally, CPR bacteria have been observed to carry out the horizontal transfer of genes between itself and their hosts, further contributing to the evolution of the community they are in (42).

1.2.2.2 Coverage

Coverage is a metric that represents the consensus built at a given base or base window and describes the number of reads that align to or cover the region. A higher coverage essentially indicates greater confidence in the accuracy of a base or base window and, when evaluated across a genome, can indicate overall assembly quality or highlight problematic or ambiguous regions. The coverage at which a nanopore-based assembly is considered sufficient can vary depending on the type of analysis intended, the sequencing kit used, the pipeline used, and quality of the data. Although there are no universal rules for what the minimum coverage should be, a fold coverage of 30x was considered to be reliable for purposes such as detecting single nucleotide variants for assemblies that were based on ONT's v10 chemistry sequencing kits and were generated and polished with Flye and Medaka, respectively (27).

1.2.3 Identification and Annotation of Novel Microbial Sequences

Beyond generating quality metagenomic assemblies, it can also be challenging to extract useful information from MAGs and other assembled metagenomic fragments, especially from novel or unique ecosystems. Assemblies of this nature are problematic simply because the sequences obtained can be divergent from any sequence in current reference databases. Without any strong similarities to a reference, it becomes difficult to perform any identification, gene prediction, or functional annotation of novel sequences. Addressing this challenge is bottlenecked by the ability of bioinformatic tools to

accurately detect distant homology and the amount of data available in reference databases.

1.2.4 Taxonomic Classification of MAGs

To better understand the microbial biodiversity within GAC samples, it is typical to extract taxonomic information using taxonomic classification tools, which typically compare 16s rRNA percent identity or average nucleotide identity of queries against a database. For microbial metagenomes, it is common to use the Genome Taxonomy Database Toolkit (GTDBTK). Since GAC samples have seldom been studied and likely contain completely novel genomes, it is expected that classifications will not be deep and that few, if any, assemblies would have been identified previously (45). Furthermore, it was expected that few identified genomes would persist across GAC samples collected over time, given the frequent exchange of GAC in filters and the variability of ORW composition. Regardless, certain genes and pathways important to the survival of bacteria in the community are expected to be common across assemblies collected from GAC samples, including naphthenic acid-degrading genes.

1.2.5 Phage Prediction

Investigating phages may provide insight into advantageous genes and community dynamics in ORW environments, as phages often play a role in the horizontal transfer of genes by packaging host genes during their replication cycle and subsequently infecting other bacteria. These advantageous genes could involve NA-degrading genes, which would help bacteria utilize the only available carbon source in GAC, or other genes such as membrane pumps that prevent the accumulation of toxic materials such as heavy metals. Given the diversity of NA's encountered and the variability of NA's available as a carbon source over time, horizontal gene transfer is likely important for the survival of the community. Identifying phages is typically done through the alignment of a sequence to reference databases or through alignment-free methods, which rely on machine or deep learning models. To reap the benefits of both methods, a tool called INHERIT uses both deep representation learning and alignment-based methods, which outperforms other popular tools such as DeepVirFinder or VirSorter (46).

1.2.6 Sequence Homology Based Annotation

Annotations provide valuable insight into the functional capabilities of a community and can provide strong clues to which genes are involved with NA degradation when coupled with downstream analyses such as differential expression. The most popular methods for annotation are based on nucleotide or protein sequence similarity, but when the query sequences are highly novel, it can be difficult to annotate sequences as they may be at or below the threshold for homology finding by sequence alignment. Furthermore, sequence similarity has been shown to be weakly correlated to functional similarity in many cases (47), and this problem could be worsened in environmental samples where many annotated sequences likely straddle this threshold.

In this case, a gene prediction and annotation tool called Bakta was used, as it contains certain advantages over popular sequence-based annotation tools such as Prokka or Prodigal (48). Bakta uses a taxonomy-independent database based on UniProt's entire UniRef protein sequence cluster universe and provides rich database cross-references to databases that include Gene Ontology, Pfam, and KEGG, which is favorable for more comprehensive annotation of unknown MAGs (48–51). Furthermore, Bakta is adept at detecting short open reading frames, which are often overlooked by standard gene prediction tools. For this ecosystem, however, the use of Bakta led to poor annotations as many of the predicted genes remained hypothetical.

1.2.7 Protein Structure Prediction and Homology Based Annotation

Since protein structures are often more conserved than their sequence and are often correlated to their function, the inclusion of structure-based homology searches into the analysis was expected to enrich our annotations and make them more informative (52,53). Compared to sequence-based approaches, structure-based approaches are expected to be more sensitive as more distant functional relationships can be detected. Experimental approaches such as X-ray crystallography and nuclear magnetic resonance spectroscopy are typically used to obtain protein structures, but they are expensive, time-consuming, and simply not feasible for our GAC samples. Therefore, computational

approaches are necessary in this case - but obtaining accurate protein structures computationally has been a fundamental challenge for computational biologists and bioinformaticians for decades.

Generally, two strategies are employed when predicting protein structures from sequences: template-based modeling and de novo modeling (45). Template-based modeling relies on the idea that proteins with similar sequences have similar structures. Established protein structures are used to model the query if their sequence similarity is high, but this strategy is unsuccessful when the sequence is divergent. De novo modeling is a template-free method that relies on general protein folding principles and energetics to search for the lowest energy conformation out of all the possible conformations that a protein sequence can have. Since the conformational search space is so large, querying large datasets of thousands of protein sequences is an extremely time-consuming and computational task. Furthermore, the usefulness of de novo predicted protein structures in biological applications was limited as protein structure prediction tools fell short of the accuracy of experimental structures up until AlphaFold was introduced (46).

The development of AlphaFold represents a significant breakthrough for de novo modeling, making a large leap in accuracy and speed by incorporating attention-based neural networks to predict protein structures with accuracies comparable to experimental methods—a first for any computational approach. It is a machine learning method trained on large, validated protein structure datasets to develop a more complex understanding of protein sequence-to-structure relationships. Since its introduction to the field, AlphaFold has powered its own structure database, which now contains over 200 million publicly available protein structure predictions (46,47). AlphaFold has led to the development of tools that further improve the accessibility, speed, and accuracy of AlphaFold, such as ColabFold, which implements MMSeqs2 in place of AlphaFold's native homology search algorithm to accelerate prediction speeds by 40–60 fold without compromising prediction quality (48). ColabFold aims to make AlphaFold much more accessible, as it simplifies the usage of the tool through command-line interfaces and implements AlphaFold with Google Colab notebooks, a cloud-based computing platform that allows users without expensive hardware to perform structure predictions. With ColabFold, thousands of

protein structures can be predicted in a day, making it much more feasible to predict structures from metagenomes in GAC within a reasonable timeframe while retaining structure accuracies comparable to experimental methods.

Using predicted structures of novel proteins with divergent sequences to perform structure homology searches is expected to be a powerful method for inferring functional similarities between proteins, but there are caveats when it comes to using computationally derived protein structures. The method relies on the assumption that two protein structures that are highly similar share similar functions. As such, the confidence in which we can say that a query protein shares a similar function to another based on structural similarities would only be as good as the quality of the structure prediction itself, which can be evaluated with two confidence scores internal to AlphaFold: the predicted local distance difference test (pLDDT) and the predicted aligned error (PAE) (46). Firstly, pLDDT is an alignment-free score that evaluates the plausibility and consistency of atomic local distances and stereochemical properties in a predicted structure. Generally, regional scores greater than 70 are considered to have good overall backbone predictions, with scores higher than 90 being considered highly accurate. Good overall backbone predictions can still reasonably be used for functional inference based on structure homology to an already annotated protein, as overall folds are generally indicative of function. Secondly, PAE scores are the positional error between residue pairs of the predicted model versus a reference model in angstroms. This metric is generally used to assess the accuracy of the position and orientation of domains, where a smaller score means a closer agreement with position and orientation relative to an established model (47). These two scores can give insights into the confidence of a predicted structure, which affects the confidence of functional annotations when searching for structure homology. If our query-predicted structure shows a high degree of structural similarity and aligns well with an annotated protein, there is a stronger case to be made that these two proteins have functional similarity. It is still important, however, to recognize that structural similarities don't always guarantee similar function since proteins that have different functions but similar structures have been identified previously (49).

1.3 Developing New Approaches for Investigating NA-degrading Communities

Naphthenic acid-degrading bacterial communities have been seldom studied, much less communities seeded in NA-enriched GAC at the genomic level. The objective of this thesis is to characterize the community and potentially identify naphthenic acid-degrading genes or pathways, but this can be challenging as with any environmental sample. Here, I also demonstrate different useful strategies in a custom workflow that addresses the aforementioned challenges with NGS sequencing, assembly, and annotation to extract the most possible information from GAC samples. I assembled multiple high-quality MAGs and sub-genomic-sized assembled circular contigs (ACCs) in two rounds of assembly, successfully circularizing uncircularized contigs from the first round. I characterized the bulk properties of ACCs across 10 GAC samples collected between 2019 and 2022, assigned taxonomic classifications to them, and annotated them with sequence homology-based tools. Since many predicted coding sequences remained hypothetical, I attempted to enrich annotations using a structural homology-based approach. However, given the limitations of computational resources needed for protein structure prediction of an entire metagenome, my current focus for structure-based annotations has shifted to my subset of bacteriophage to demonstrate the principle of the annotation method. I show that structure-based annotation based on ColabFold and Foldseek was able to add substantial annotation power, even for putative bacteriophages, which are notoriously hard to annotate.

Chapter 2

2 Methods

Here, I describe a combination of wet-lab and computational approaches to analyze GAC samples including DNA and RNA extraction, sequencing, assembly, and annotation of high-quality assembled circular contigs (ACCs) from GAC samples (Fig. 5).

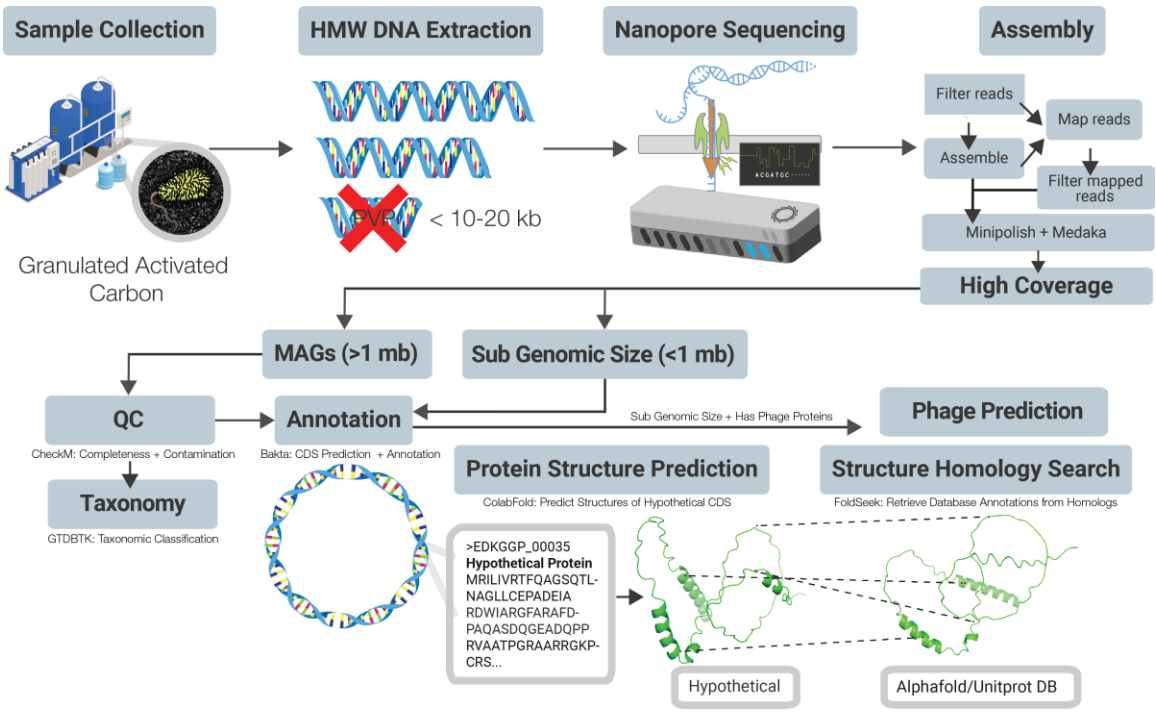

Figure 5. GAC samples were collected from an oil refinery site in Ontario when the filters were being exchanged. Ultra-high-molecular-weight DNA was extracted, and low-molecular-weight fragments were removed by PVP precipitation. A library was prepared from the purified DNA using the Oxford Nanopore LSK109 or LSK110 kit and was sequenced on r9.4.1 flow cells using the Oxford Nanopore MinION platform. The DNA was assembled using a custom workflow that included a post-assembly filtering step using Gerenuq (37) to remove low-quality and poorly mapped reads from the assembly. Only fully closed contigs with an estimated minimum coverage of 10 were kept for further analysis. Gene prediction and initial annotation were done using Bakta (50) and any open reading frame labeled as

hypothetical were kept for structural prediction and annotation via Colabfold (48) and Foldseek (51).

2.1 DNA Extraction

The preparation of ultra-high-molecular-weight DNA is essential for producing quality circularized assemblies. Prior to extraction, GAC samples were collected and frozen at -80 °C. Approximately 10 g of GAC seeded with bacteria was added to a 50 mL falcon tube, followed by 10 mL of lysis buffer (10 mM Tris-HCl, 100 mM NaCl, 25 mM EDTA, 0.5% (w/v) SDS) and 100 µL of lysozyme (25 mg/mL). The sample and buffers were mixed by slowly rotating the tube while trying to minimize granule movement, and then incubated for 1 hour at 37 °C. After 1 hour, 5 µL of RNase A (20 mg/mL) was added, the sample was mixed gently, then incubated for 1 hour and 30 minutes at 57 °C. Finally, 100 µL of Proteinase K (800 units/mL) was added to the tube, gently mixed, and the sample was incubated for another 1 hour and 30 minutes at 57 °C.

The lysate was decanted into a new 50 mL falcon tube, without transferring the carbon granules. Then, 1 volume of 25:24:1 phenol:chloroform:isomayl alcohol was added to the lysate, the mixture was rocked gently for 8 minutes, then spun at 3000 x rcf for 3 minutes. The aqueous phase was transferred to a new 50 mL falcon tube using wide-bore pipette tips. This process was repeated twice with 1 volume of chloroform instead of phenol:chloroform:isomayl alcohol, for a total of 2 chloroform washes.

To precipitate the DNA, 1/10 volume of 3 M sodium acetate at pH 4.5 was added, followed by 2 volumes of ice cold 100% ethanol. DNA that precipitated immediately was spooled out into an Eppendorf tube with a sterile Pasteur pipette that had been melted into a hook. The DNA was washed twice with nuclease free 75% ethanol and resuspended in 500 µL to 1 mL of Tris buffer (10 mM, pH 8) overnight, depending on the size of the pellet.

2.2 DNA Size Selection

Prior to library preparation, 60 µL of the DNA sample was added to 60 µL of 3% PVP 360000 solution (1.2 M NaCl, 20 mM Tris-HCl, pH 8) and mixed thoroughly by

inverting. The mixture was spun at 10000 x rcf for 30 minutes at RT to preferentially precipitate high-molecular-weight DNA. The supernatant was discarded, then the DNA pellet was washed twice with 75% nuclease free ethanol. The DNA was resuspended in 60 µL of Tris buffer (10 mM, pH 8).

2.3 Sequencing Library Preparation

The DNA library was prepared using the ONT LSK109 or LSK110 kit according to the manufacturers protocol, with the following changes. The repair and end prep steps were extended to 15 minutes at 20°C and 15 minutes at 60°C instead of the recommended 5 minutes at each temperature. Additionally, Omega Biotek Mag-Bind beads were used in place of Ampure XP beads. The libraries were sequenced on a MinION, with 9.4.1 flow cells for libraries prepared with the LSK109 kit or LSK110 kit.

2.4 Primary and Secondary Assembly

Raw reads were basecalled with Guppy v6.3.8 using the arguments '-c dna_r9.4.1_450bps_sup.cfg --min_qscore 7'. Basecalled reads were checked for length and q-score using PycoQC (v2.5.2) (52) to determine a cutoff for lower quality data to discard. The minimum length was always greater than 500 nt with a minimum read q-score of 7. Once a cutoff was determined, NanoFilt (v2.8.0) (53) was used to filter the reads with the length and q-score parameters set to the cut off, as well as the argument `--headcrop 50`. The filtered reads were then assembled with Flye (v2.9-b1768) (31) in `--nano-hq –meta` mode. After the initial assembly, additional assemblies were yielded using a secondary assembly pipeline. Briefly, reads for a given sample were aligned to uncircularized contigs obtained from the same sample with Minimap2 v2.24 (54) and were binned using MetaBAT2 v2.12.1 (55). Reads aligned to a bin were filtered using Gerenuq v0.2.3 (37) on default settings to keep only alignments over 1000 bp, with a score of 1 and at least 50% identity. For each bin, Gerenuq filtered reads were passed on to Flye with the genome size set to the total size of the bin. Only assemblies that had an estimated coverage of at least 10 and that were tagged as circularized by Flye were extracted from the assembly graphs as a GFA file and passed on to the polishing pipeline. The naming of the ACCs were retained from the names assigned by Flye.

2.5 Polishing and QC

Before polishing, all metagenomic reads from a given sample were aligned to each assembly obtained from the same sample using Minimap2. Mapped reads were filtered with Gerenuq to keep alignments that are at least 1000 bases long, with a score of at least 1 and at least 90% identity. Then, draft GFA assemblies were polished with Minipolish v0.1.2 (36) using the Gerenuq filtered reads that were converted to fasta format. In order for Minipolish to accept Flye assemblies however, minor changes were made to the file format of Flye's outputs - namely by changing converting the GFA file format to GFA2 with GFAKluge (56) and by adding "l" as a suffix to sequence names. Additionally, Minipolish was run with the '--skip-intial' argument as the initial step requires non-standard data unique to the Miniasm assembler. The polished assemblies were converted to fasta format, then underwent a second round of polishing with Medaka v1.6.1 (35) using the same Gerenuq filtered reads. To assess the coverage of polished assemblies, Gerenuq filtered reads were mapped back to the polished assemblies and put through Mosdepth v0.3.3 (57) to calculate coverage by 1000 base windows. Polished assemblies that were greater than 1 mb in size were checked for completeness and contamination using Checkm v1.2.2 (41), and those that had greater than 90% completeness and less than 10% were considered high-quality MAGs.

2.6 Annotation

Bakta v1.5.1 (50) with the '--complete' argument was used to annotate the polished assemblies. The assemblies were subset further based on Bakta annotations - assemblies that contained ribosomal related proteins were discarded, and those that also had annotated bacteriophage related proteins were passed to INHERIT (58) to identify potential bacteriophage genomes using their pre-trained model. From each bacteriophage predicted by INHERIT, all amino acid sequences from Bakta annotations were passed to Colabfold v1.3.0 (48) with the arguments '--amber --templates --num-recycle 3 --use-gpu-relax --num-models 1' to predict the structure of each protein. For each protein that had a structure prediction with a mean pLDDT score greater than 70, the relaxed model was taken and queried against the AlphaFold/Uniprot database using Foldseek v90b (51) with the 'easy-search' function. Functional annotations including KEGG and GO terms

were retrieved using Uniprot's API by querying only the best Foldseek hits, which are filtered for an e-value greater than 1e-10 and for an LDDT score greater than 0.7, for each predicted protein structure. Pathway information for KEGG KOs retrieved with Foldseek was obtained using the KEGGREST 1.38.0 package (59).

2.7 Singleton and Recurring Assemblies

Singleton and recurring ACCs were determined by performing an all-versus-all BLAST. A pair of assemblies from different samples were counted as identical if they had a percent identity of 98%, a query coverage that is within 10% of the query length, and a query length that is at least 90% of subject length. Each set of ACCs that were considered singleton or recurring were subset further if they were predicted bacteriophage.

To detect pairs of MAGs that have regions of high similarity between samples, an all-versus-all BLAST was performed for all high-quality MAGs, where hits were filtered for 99% identity and a query coverage length of at least 10 kb.

To observe the presence of structural bacteriophage proteins across samples, the total set of protein structures from each recurring predicted bacteriophage ACC was collected and clustered with Foldseek using the greedy set cover algorithm, and alignments where 80% of the sequence is covered by the alignment are kept. The clusters were subset based on keywords in their representative's Foldseek annotations - "head/tail/capsid/plate". For each cluster in which the representative is a putative head, tail, capsid or baseplate related protein, the protein structure of the cluster representative was aligned with its members to obtain a TMScore using Foldseek.

Chapter 3

3 Characterization of Assembled Circular Contigs from GAC Metagenomes

This chapter follows up on the ACCs generated from the methods outlined in the previous chapter. As these GAC samples have not been sequenced before, it is important to analyze the quality of the ACCs and perform bulk characterization on the ACCs collected to unveil the microbial diversity within this ecosystem and their potential roles in the degradation of naphthenic acid. In this chapter, I assess the quality and analyze these ACCs to identify genomes, perform annotations, and discuss other broad features or dynamics of the community that may be relevant to understanding NA degradation in GAC samples.

3.1 Sequencing and Assembly Results

In total, 10 samples of the microbial biofilm growing on the granulated activated carbon (GAC) beds were collected from the refinery's wastewater treatment plant between 2019 and 2022. The first sample, GAC0, was collected in 2019, but the day and month were not noted. Ultra-high-molecular-weight DNA was isolated from these samples and sequenced on r9.4.1 flow cells using ONT's MinION platform, then base-called and assembled as outlined in the Methods and summarized in Figure 5. ACCs that were fully closed, non-repetitive circular DNA sequences with an estimated coverage of greater than 10 were selected for polishing. Across the 10 GAC samples collected, 112 ACCs greater than 1 mb were assembled, although only 80 were more than 90% complete and less than 5% uncontaminated, according to CheckM (Appendix A). Out of the complete and uncontaminated MAGs, 72 were considered high-quality as they encoded 16S, 23S, and 5S rRNA genes, as well as the full complement of tRNAs. The remaining 8 were missing 5S rRNA genes but fulfilled all the other criteria for high-quality MAGs by being complete, uncontaminated, and encoding all tRNAs. In total, 5432 ACCs less than 1 mb in size were collected. 423 ACCs smaller than 100 kbs had at least one putative bacteriophage-related protein in their Bakta annotations, and these were passed on to the INHERIT package to score their likelihood of being true phages. Of the 423 ACCs, 227

bacteriophages were predicted. Table 1 summarizes the collection dates and statistics from the sequencing and assembly results for each individual GAC sample. In each of the 10 samples, multiple high-quality MAGs were collected, and the secondary assembly pipeline yielded additional high-quality MAGs for all samples except GAC1 and GAC9. For ACCs smaller than 1 mb, the secondary assembly pipeline also yielded additional assemblies across all samples. Although it is generally expected that a higher quantity of more contiguous reads would lead to more MAGs being produced, the trend is not fully clear, as some samples like GAC7 with a lower quantity of quality and contiguous reads produced a greater quantity of high-quality MAGs than in GAC9, for example. Thus, the minimum amount of data needed, and the quality needed to capture the full biodiversity of GAC samples are not yet established.

Table 2. Sequencing and Assembly Statistics of the 10 GAC samples collected from the oil refinery wastewater treatment facility. All samples were sequenced using ONT's MinION platform with r9.4.1 flow cells, and the number of gigabases (gbs) collected, the mean q-score and rN50 of basecalled reads are shown here. The assembly statistics demonstrate the number of assemblies produced in the primary and secondary assembly steps (P+S). Metagenomic assembled genomes (MAGs) are assembled circular contigs (ACCs) that are estimated to be complete (>90%) and uncontaminated (<5%) by CheckM.

Sample	Date Collected	Gbs	Mean Q-Score	rN50	MAGs (P+S)	ACCS < 1 mb (P+S)	Putative Phage (P+S)
GAC0	xx/xx/2019	14.54	13.5	8.1	2 + 1	195 + 101	5 + 1
GAC1	27/01/2020	9.88	13.8	4.5	1 + 0	314 + 31	10 + 0
GAC2	14/02/2020	19.96	13.6	20	6 + 5	453 + 76	27 +15
GAC3	02/03/2020	22.71	14.3	17.9	10 + 6	624 + 114	37 + 1
GAC4	06/10/2020	26.65	13.2	12.4	2 + 7	405 + 69	9 + 0

GAC5	03/05/2021	15.06	13.3	12.3	3 + 4	561 + 98	35 + 3
GAC6	07/05/2021	19.95	13.4	9.5	4 + 7	405 + 72	14 + 4
GAC7	14/05/2021	24.40	13.3	6.4	2 + 10	746 + 110	32 + 5
GAC8	10/08/2021	35.20	14.5	30.1	5 + 3	445 + 38	16 + 2
GAC9	21/04/2022	33.52	13.6	9.7	2 + 0	494 + 81	24 + 0

3.2 Most Putative Bacteriophage have Comparable Standard Deviations and Coverages to High-quality MAGs

Currently, the scope of assembly quality evaluation is limited to MAGs and putative phages. The coverage for the 72 high-quality MAGs and 227 putative phages was calculated per base with Mosdepth (57).

High-quality MAGs greatly varied in mean coverage—from 6-fold to over 500-fold. Although some MAGs have lower than ideal coverages, a relatively low or uneven coverage does not guarantee that a MAG is of poor quality when it is complete and uncontaminated. Complete and uncontaminated MAGs have almost all expected genes and no foreign genes, indicating that the MAG is likely representative enough to be used for taxonomic and functional characterization.

Since standard metagenomic quality control metrics like completeness and contamination are not available for the predicted bacteriophage, I assessed the assembly quality of putative phage by examining the relationship between the coverage depth and uniformity of high-quality MAGs to compare with putative bacteriophage ACCs. In Figure 6, I plotted the per-base mean coverage and the per-base standard deviation for quality MAGs and putative phage ACCs. There was a linear relationship in which the SD/cov ratio was less than 1 for almost all high-quality MAGs. For most of the putative phages, the vast majority also had an SD/cov ratio less than one, which was consistent with the coverage characteristics observed for high-quality MAGs. However, there was a subset of putative

phage ACCs that had higher than expected SD/cov ratios, which warrants manual inspection for validation. Some high-standard deviation phages can be validated by having full-length alignments, as represented by enlarged symbols in Figure 6A or denoted by ** in Figure 6B.

Figure 6A reveals the evenness of coverage calculated in 1% bins for high-quality MAGs collected across all samples, and few MAGs have very even coverage. MAGs such as 5322 from GAC2, however, have very high and even coverage distributions, and these are candidates for being finished, high-quality MAGs. ONT claims that with their “nano-hq” basecalling option, consensus accuracies of Q50 can be consistently achievable with 100-fold coverage with their basic recommended pipeline. Given that basecalling was also performed with the “nano-hq” mode and that the pipeline I used includes additional filtering steps as outlined in the methods section, it follows that MAGs with high and even coverage likely reach consensus accuracies of that level.

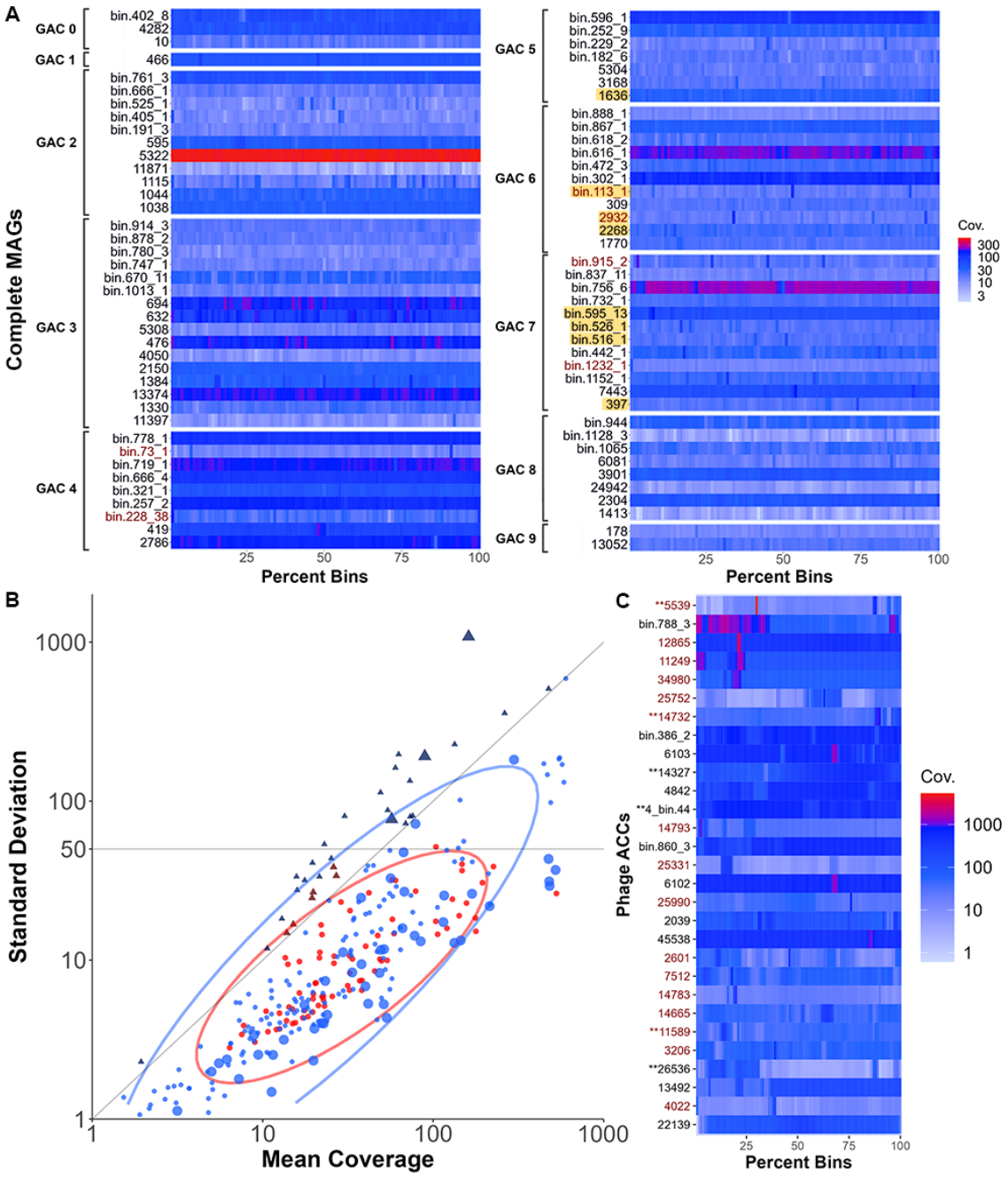

Figure 6. Evaluating the quality of high-quality MAGs and putative phages using a coverage-based metric. The mean coverage in 1% bins was calculated for MAGs with higher than 90% completeness and less than 5% contamination (A). MAGs highlighted in yellow represent those with missing ubiquitous bacterial genes outlined by Bowers et al. that are required to be present for a MAG to be

considered high-quality (40). MAGs with standard deviation (SD) versus coverage (cov) ratios greater than 1 are colored in red. Plotting the SD of the per-base cov as a function of the mean per-base coverage reveals a linear relationship for high-quality MAGs (red) in which almost all MAGs have ratios less than 1, and most predicted bacteriophage (blue) ACCs follow this trend (B). The ovals also show the 95% confidence interval for each group. However, several phages had SD:cov ratios higher than 1 (triangle). A plot of mean coverages was also calculated in 1% bins for putative bacteriophage ACCs with standard deviations higher than 50, ordered from highest SD to lowest (C). These heatmap visualizations help identify potential problematic regions that lead to high standard deviations and inflated mean coverages. Most bacteriophage ACCs in this subset have one or more coverage spikes that span 1-5% of the genome. Conversely, there are few bacteriophage ACCs with relatively even coverage for most of the genome but have drops in coverage over broad regions. ACCs denoted with ** contain at least one read that aligns over its full length with at least 90% sequence identity to the reference. Labels in red indicate those ACCs with SD/cov ratios greater than 1.

Figure 6B shows the coverage distribution along putative phages with standard deviations over 50. The ACCs are ordered from highest to lowest ratio, and this shows where there are small regions along the sequence that have much higher or lower mean coverages, leading to a high standard deviation. These could represent chimeric assemblies between closely related bacteriophages, direct repeats, or circular permuted sequences where the end sequences were not perfect direct repeats.

In general, lower coverages over broad or even smaller regions may be a result of the depletion of DNA fragments smaller than 10 kb during the DNA extraction and sequencing library prep steps, which was originally done since analyzing bacterial genomes and plasmids was the initial goal. Without the size selection step, there may have been more full-length reads from phage to help validate suspicious ACCs, and there may have been a higher quantity of phage ACCs recovered.

Regardless, for some assemblies highlighted in Figure 6C such as ACC 5539, there exists at least 1 full-length read alignment with a minimum 90% sequence identity that can be used to provide some validity to the assembly. However, for other ACCs, it is not yet clear if the coverage spikes or drops indicate misassemblies or not. It was concluded that the majority of the putative phage ACCs were assembled and polished to a standard generally consistent with that observed for high-quality MAGs. Furthermore, in the absence of a reference sequence and other information, the SD:cov ratio can be used as a proxy to help identify poorly assembled sub-genomic circles from a large dataset.

3.3 Sequence Homology Based Annotations

3.3.1 Gene Prediction and Annotation with Bakta

All ACCs assembled across all GAC samples were annotated with Bakta, a state-of-the-art sequence-based annotation tool. Across all ACCs assembled, a total of 966992 coding sequences were predicted, and 592730 remained hypothetical. For high-quality MAGs, 285982 coding sequences were predicted and 96671 remained hypothetical. For phages, there were 18099 predicted coding sequences (CDS) with 16606 remaining hypothetical.

3.3.2 Hydrocarbon Degradation Genes in High-Quality MAGs

Common hydrocarbon degradation genes were searched for in high-quality MAGs using the built-in CANT-HYD HMMs in HMMER. MAGs with hits to hydrocarbon degradation genes with a minimum e-value of 1e-10 are summarized in Appendix B. In total, 78 MAGs across all sequenced GAC samples contained at least 1 hydrocarbon degradation gene, and there was at least 1 MAG with an identified hydrocarbon degradation gene in each sample. 123 out of 823 CDS that were identified as hydrocarbon degradation genes with CANT-HYD were previously hypothetical, according to Bakta. Notably, acyl-CoA dehydrogenase was found in ACCs across all samples except GAC1. This enzyme is part of the fatty acid beta-oxidation pathway and was shown previously to have the highest level of overexpression relative to other dehydrogenases in Pseudomonas fluorescens Pf-5 when grown on 4′-n-butylphenyl-4-butanoic acid, which is a surrogate NA (19). This suggests that acyl-CoA dehydrogenase plays a significant role in degrading branched and aromatic NAs, which are features that

typically increase the recalcitrance of NAs to biodegradation. Glu and Leu dehydrogenases were also found in ACCs across 3 and 4 GAC different samples, respectively. In the same study by McKew et al., these amino acid degrading enzymes were also upregulated and are likely to be involved with NA degradation (19). Other hydrocarbon degradation genes include 3-octa-prenyl-4-hydroxybenzoate carboxy-lyase, which is found in every sample except GAC9. This enzyme has not been implicated in NA degradation before, but is only known to catalyze a decarboxylation reaction in ubiquinone biosynthesis in *E.coli* (60). The protein also requires manganese as a cofactor, which is common in ORW. Since it is known that existing pathways in NA-degrading bacteria can be adopted for NA biodegradation and that the substrate of 3-octa-prenyl-4-hydroxybenzoate carboxy-lyase shares structural similarities to known NAs (61), it can be reasonably speculated that the enzyme could be important to the degradation of NAs with carboxy groups attached to aromatic groups. Regardless, given that NA is the most abundant carbon source in GAC, it is possible that many of these hydrocarbon degradation genes could also be involved in NA degradation – especially genes that appear in multiple samples.

3.3.3 GO Terms in High-Quality MAGs

GO terms were collected from Bakta annotations, and GO labels were retrieved using the GO BioLink API. The top 10 GO terms across GAC samples are represented by Figure 7A, and GO terms related to metabolic processes are also shown in Figure 7B. A majority of the terms are related to basic functioning in bacteria, with the top two terms being related to the ribosome. Metabolic GO terms were underrepresented, with only 6 total metabolic GO terms across all samples: 3 involved carboxylic acids; 1 was for amino acids, 1 for carbohydrates, and 1 for one-carbon molecules.

3.3.3.1 No Annotated Genes Attached to GO Terms Were Known NA-degrading Genes in High-quality MAGs

Genes associated with lipid or amino acid metabolism have been identified in the past as likely being involved with naphthenic acid degradation. Alpha, beta, and gamma oxidation pathways have been shown to be upregulated in NA-degrading bacteria, and as

a result, it is expected that metabolic GO terms related to fatty acid metabolism would be found (19,20). However, the 3 annotations containing carboxylic metabolic GO terms were all from malate dehydrogenase genes, which is a widely distributed enzyme known to be involved with the conversion of malate to oxaloacetate by reducing NAD to NADH in the citric acid cycle. Whether or not it is involved with NA degradation has not been experimentally determined, although it has been noted that microbial malate dehydrogenases demonstrate much more diversity among prokaryotes relative to malate dehydrogenases between eukaryotes (62). Furthermore, malate dehydrogenases catalyze an oxidation reaction much like another protein on the same pathway, succinate dehydrogenase, which has been observed to be overexpressed in a NA-degrading bacterial strain (19). Although they are both mainly known for being key enzymes in the Kreb cycle pathway, existing evidence hints that it is not implausible for malate dehydrogenases in GAC bacteria to have evolved to degrade NAs, especially since existing pathways are often also used in NA biodegradation (17,63).

Other metabolism-related GO annotations include an aspartate carbamoyltransferase catalytic subunit as part of amino acid metabolism, 3-hexulose-6-phosphate synthase in carbohydrate, and one-carbon metabolism. It might be important to recognize that all metabolic GO annotations were retrieved from ACCs that have relatively low coverage and occur in only one sample with low relative abundance, suggesting that these genes were not particularly significant for fitness in ORW or that they were erroneous. Regardless, none of these enzymes with attached GO terms from Bakta annotations have

been implicated in NA degradation previously but may still be of interest.

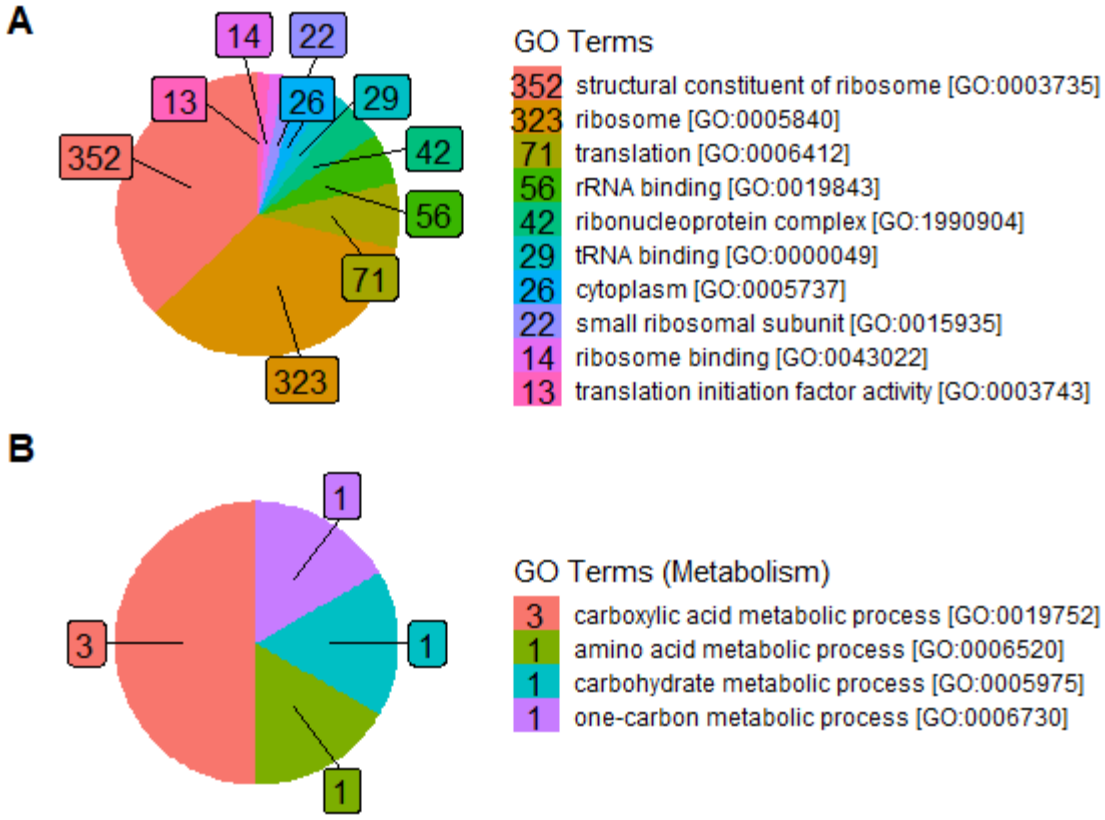

Figure 7. Top 10 GO terms (A) and all metabolic related GO processes (B) retrieved from Bakta annotations of high-quality MAGs. Metabolic GO terms originated from contigs with relatively low coverage.

It may not be sufficient to rely on GO terms from Bakta annotations when it comes to narrowing down the search for likely NA-degrading genes. Oddly, proteins such as Acyl-CoA dehydrogenase were confirmed to be present in multiple ACCs with CANT-HYD HMMs and with Bakta, but Bakta does not attribute GO terms to the annotations despite the existence of GO terms for it (GO:0003995). This could indicate that the annotations are based on homology to protein sequences that are poorly characterized and have yet to be assigned a GO term.

3.3.4 KEGG Pathways Reveal Potential NA-degrading Genes

To observe which KEGG pathways are present in MAGs from each sample, KEGG KOs were collected from Bakta annotations, and all related KEGG pathways were retrieved

(Appendix C). A pathway's completeness was estimated by the proportion of KOs observed in MAGs versus the total KOs in a KEGG pathway. All pathways and modules were estimated to have low completeness across all samples.

KEGG pathways in Bakta annotations of high-quality MAGs that may have a role in hydrocarbon or NA degradation include the following broadly defined pathways, ordered from highly ubiquitous across samples to uncommon: “Microbial metabolism in diverse environments”, “Carbon metabolism”, “Fatty acid metabolism”, “Degradation of aromatic compounds”, “2-oxyocarboxylic metabolism” and “D-amino acid metabolism”. Fatty acid metabolism related genes appear in all samples but GAC0 and GAC01. Comparatively, fatty acid metabolism genes were found in all samples except GAC1 with CANT-HYD.

However, there were many unexpected pathways that were common across samples but should only appear in eukaryotes, including “GABAergic synapse”, “Diabetic cardiomyopathy” and “Ribosome biogenesis in eukaryotes”. This could indicate that there are several Bakta annotations that are spurious, perhaps due to high sequence divergence.

3.4 GAC Samples Contain a Diverse, Everchanging Consortium of Bacteria

High-quality MAGs were assigned taxonomic classifications with GTDB-TK, and the classifications were diverse but mostly dominated by one particular phylum. Most MAGs had classifications to the genus level, while all had classifications to at least the order level. Only 9 had classifications at the species level. Figure 8 shows the relative abundances of MAGs within each sample, which were calculated based on their mean coverages relative to the total coverage of MAGs in a sample after polishing. As such, many species that were lower in abundance or simply missed from the assembly pipeline are not shown, and therefore, the full biodiversity of the samples is likely not represented. Based on the ACCs that could be readily reconstructed from the method used in this thesis, MAGs from the 10 GAC samples consistently appear to be dominated mainly by those belonging to the phylum Pseudomonadota. A class of Pseudomonadota,

Gammaproteobacteria, was observed to be the most common in the first 4 samples collected but became less common as time went on. Alphaproteobacteria, another class of Pseudomonadota, conversely surged in abundance in GAC5 after being relatively rare in the first 4 samples and unretrievable in the first 2. Pseudomonadota have been observed previously to be common in ORW, which makes these observations unsurprising (63). Other bacteria belonging to phylum previously observed in ORW that were also identified here include Bacteroidota and Actinobacteriota. Pseudomonadota, Bacteroidota, and Actinobacteriota have been reported to contain members involved with anaerobic degradation of hydrocarbons and degradation of NAs in ORW moving bed biofilm reactors (63).

Regardless, no previously identified NA-degrading bacterial strains have been sequenced from these samples. At least one MAG (bin.1128_3) is related at the family level to *Sphingopyxis witflariensis* and another (bin.402_8) at the order level to NA-degrading *Pseudomonas* strains - but any resemblance to identified NA-degrading bacterial strains was mainly at the class level. Several previously identified strains such as ones belonging to the genus *Pseudomonas, Sphingopyxis, Aquamicrobium,* or *Bosea* all belong to the phylum Proteobacteria where the prior 2 belong to the class Gammaproteobacteria and the latter 2 to Alphaproteobacteria.

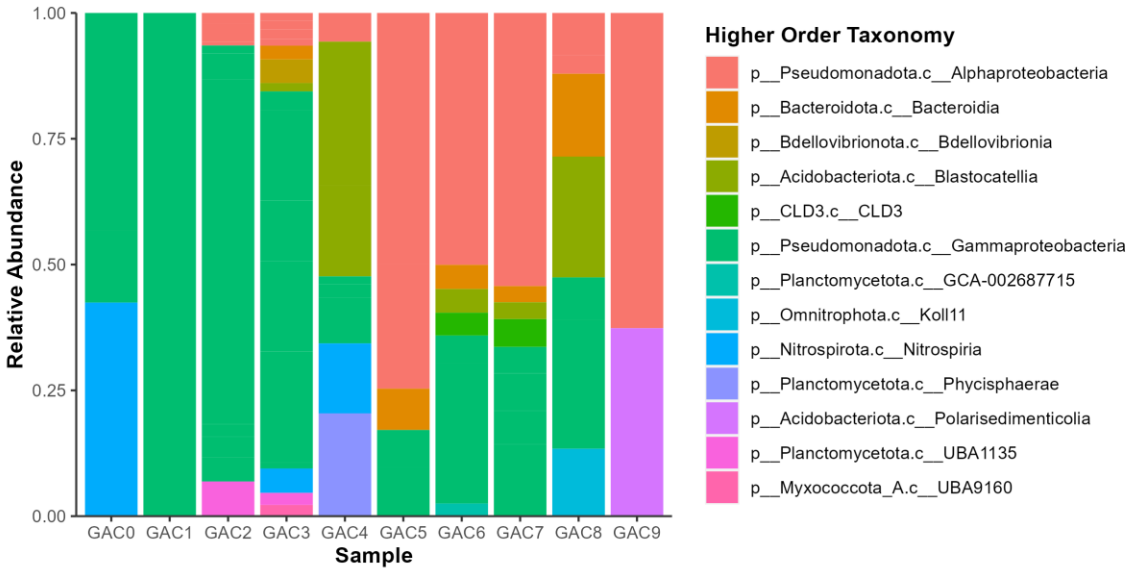

Figure 8. High-quality MAGs generated from nanopore-sequenced GAC metagenomes were taxonomically classified by GTDBTK, and their relative abundances within samples were calculated based on their mean coverage divided by total coverage. The reconstructed MAGs are dominated by those belonging to the Pseudomonadota phylum, with Gammaproteobacteria being the most prominent class in samples collected early on, while Alphaproteobacteria became more prominent in later samples. It is important to recognize, however, that these only represent MAGs that could be reconstructed using the outlined methods – meaning that this is likely not fully representative of the full biodiversity of bacteria in GAC.

3.4.1 MAGs Identified at the Species Level Were Found in Other Wastewater Metagenomes – but not Oil Refinery Wastewater

9 high-quality MAGs were identified at the species level with GTDBTK and information regarding the isolated species was collected directly from the GTDB (64). 8 unique species were identified and collected previously from other wastewater metagenomes in separate studies.

OLB10_sp001567275 are chemolithoautotrophs that originate from a partial-nitritation anammox (PNA) reactor in a wastewater treatment facility for the potato processing industry. PNA bacteria can anaerobically convert ammonium and nitrite to dinitrogen gas

without a carbon source, making this process ideal for treating sludge with low dissolved oxygen and high ammonium levels (62). JAJVID01 sp022072205 was another identified species that is also an anammox organism, but its sample of origin was not disclosed.

Accumulibacter similis was found in activated sludge from a lab-scale Enhanced Biological Phosphorus Removal (EBPR) enrichment culture. *Ferruginibacter sp017303335* was also sequenced from an EBPR bioreactor. EBPR is part of activated sludge systems to remove phosphate in wastewater treatment facilities by phosphate accumulating organisms (PAO), in which the PAO would accumulate phosphate as polyphosphate for energy storage. *Accumulibacter* are also often identified in EBPR systems and are typically the most dominant in those systems (65). Notably, PAOs typically use carbon sources and only accumulate phosphate during times of low nutrient availability. Coincidentally, the ACC (GAC2-1038) identified in this study to be this putative PAO does contain multiple hydrocarbon-degrading genes according to CANT-HYD, including Acyl-CoA dehydrogenase, 4-hydroxy-3-polyprenylbenzoate decarboxylase and molybdopterin-dependent oxidoreductase. This could hint that this organism is utilizing these genes to use NAs as an energy source since it's the sole carbon source in GAC, and its survival is further propagated by having polyphosphate as an alternate energy source. Unfortunately, this particular species has only appeared in one GAC sample. There are 3 other related ACCs identified to be in the same genus, however.

Maganitrophus morganii is another chemolithoautotroph from a manganese oxidizing enrichment culture inoculated with an iron oxide mat. This is part of the *Candidatus Maganitrophus* genus that is characterized by performing Mn(II) oxidation, whereas most species identified previously to be manganese chemolithotrophs use Mn(III/IV). Members from this genus are found in freshwater and marine environments globally and since manganese can be found in ORW, it is not surprising to find them in GAC filters. In fact, this species was identified in two different samples – GAC0 and GAC4.

CAINVI01 sp016713765 was previously collected from an activated sludge metagenome in a wastewater treatment facility in Denmark but is not well characterized (66).

Macondimonas sp021783685 originated from an acid mine tailings metagenome. Although this strain is not characterized well, another study identified this genus as phylogenetically narrow and noted how it is highly abundant in crude oil and coastal marine ecosystems. Furthermore, they are known petroleum hydrocarbon degraders and nitrogen fixators (67). In this study, multiple hydrocarbon degradation genes were identified in both ACCs under this taxonomic classification using CANT-HYD. This species was observed in GAC6 and was relatively higher in abundance but could not be reconstructed in any other sample.

Other than *Macondimonas sp021783685*, the nature of the sample from which most of these species originate seems to suggest that multiple species observed in GAC metagenomes may not necessarily be NA-degrading, but rather are chemolithotrophs – bacteria that rely on oxidizing inorganic compounds that also exist in ORW such as ammonium, nitrate, sulfur, manganese, methane, etc. Given the toxicity and nutrient scarce environment of GAC filters, there may be other genes in these species that allow them to tolerate the toxic environment in ORW rather than survive off of NAs as a carbon source. This could include membrane transporter proteins that pump out heavy metals or simply NAs, since they can potentially be toxic to bacteria. Long-chain fatty acid transport proteins and ABC transporters have been observed to increase in relative expression with exposure to NAs (19). Interestingly, GAC are at the last stage of filtration and GAC is not suitable for the removal of dissolved inorganics. The removal of dissolved inorganics is typically performed earlier in the process, meaning that these chemolithotrophs are likely surviving on trace amounts. Despite this, most MAGs identified at the species level are relatively abundant relative to other MAGs in their respective samples.

3.5 Genome-Sized ACC Sequences Were Essentially Unique Across Samples

To identify if any of the same ACC sequences larger than 1 mb appear independently over time, an all-versus-all BLAST was performed to look for MAGs that had 98% sequence identity, had sequence lengths within 10% of each other, and had 90% query coverage. Only 1 was identified to be the same in both GAC3 and GAC8 – but not in any

of the GAC samples collected between those two samples. It is possible that the abundance of these species was too low to be captured between those points where they were identified, but given that the filtering criteria used captured ACCs that do not change over time, it is possible that events where certain genetic elements were gained or lost over time were missed. Given the variability and heterogeneity of the environment from which these MAGs originate, it would be unsurprising to see a high turnover rate of bacterial species across samples or a high rate of genetic diversification events.

3.5.1 Large Regions of High Sequence Similarity Still Exist Between Few Genome-Sized ACCS

When performing an all-versus-all BLAST for ACCs having at least 99% identity but only a minimum alignment length of 10000 bases, 12 groupings of related MAGs across samples were identified. Interestingly, the ACCs within groups have the same taxonomic assignment and are very consistent in size, deviating by only a few thousand bases at most (Fig. 9). These related ACCs share large regions of very high similarity, but not along the entirety of their sequence. This could imply that various genetic elements in these species are being exchanged or replaced over time, suggesting that there are conserved core genomes that cause the observed sizes to be relatively stable. Furthermore, tiles that contain two size values indicate that there were hits between a primary ACC and a secondarily assembled ACC within a sample. Since these pairs do not align in their entirety, it is possible that they could represent closely related but distinct genomes that coexist in the same sample.

It might be worth noting that of the 12 groupings, 5 belonged to the class Gammaproteobacteria, 4 to the class Alphaproteobacteria, 1 to Blastocatellia and 1 to the species *Maganitrophus morganii* (Appendix D). For whatever reason, Gammaproteobacteria (LNEJ01, SBBG01, CAKKSB01, Accumulibacter) and Alphaproteobacteria (Sphingobium, UBA9219, JACADY01, UBA11222) appear to be well adapted to wastewater environments relative to other bacteria in the community, as they make up the majority of recurring MAGs and represent the majority of GTDB classifications in general. All ACCs grouped here contain hydrocarbon degradation genes according to CANT-HYD except 2, which is shown in Appendix D.

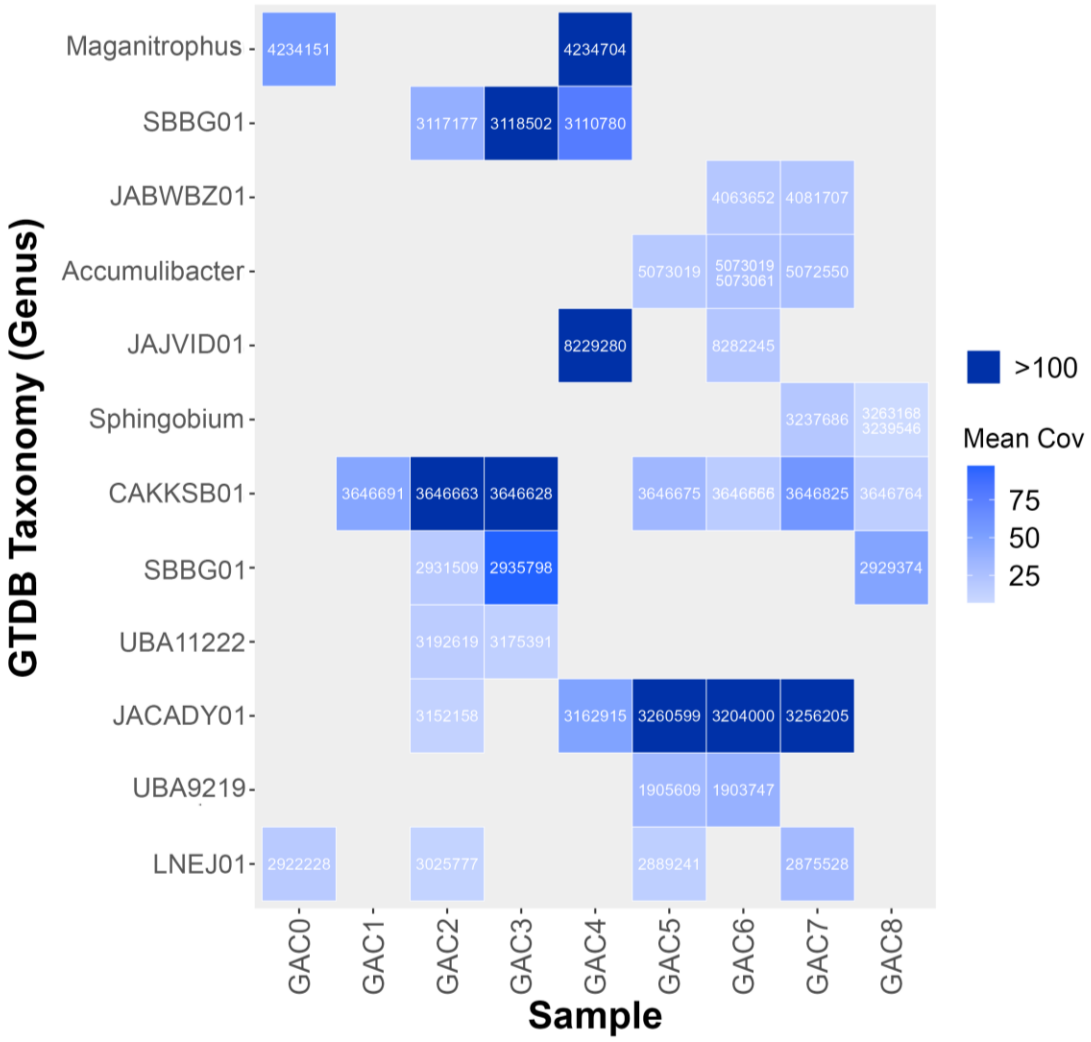

Figure 9. All ACCs larger than 1 mb were grouped by performing an all-versus-all BLAST, where reciprocal best hits of 99% identity and a minimum alignment length of 10000 bases were retrieved. All ACCs within each group shared the same taxonomy classification by GTDBTK, and ACC sizes are shown within each square. The lack of high similarity along the full length of genomes between samples yet consistent taxonomic classification suggests the existence of a core genome and accessory genome for these MAGs, in which the latter is often exchanged throughout time. ACCs belonging to each taxonomic group here are summarized in Appendix D.

Regardless, most groups persist only in 2-3 consecutive samples; LNEJ, JACADY01, and CAKKSB01 are the only groups that appear in multiple samples widely separated by time. MAGs belonging to CAKKSB01 appear in almost every sample, except GAC0 and GAC4. SBBG01 appears in GAC2 and GAC3 but does not appear until GAC8, which is approximately 2 years later. LNEJ01 was observed in GAC0 and GAC7, which are separated by approximately 2 years, but only appeared between those samples in GAC2 and GAC5. Given that MAGs belonging to CAKKSB01 persisted for the longest and had a relatively high abundance, it may be important to search for both the core and accessory genomes of ACCs in this group and compare them to other genomes in the sample. Its core genome likely contains genes that allowed for its long-term survival, and its accessory genome could potentially have genes that were beneficial for survival at each point in time.

Looking into MAGs with higher mean coverage would be of interest, as high coverage indicates more accurate assemblies and can also be a proxy for relative abundance within samples. The higher or lower presence of a particular MAG suggests better or worse fitness at a certain point in time, and when paired with metadata regarding the ORW composition at each timepoint, it could indicate pathways relevant to those environmental changes. CAKKSB01, for example, did persist the longest but was mainly dominant in GAC3 and became less prominent as time went on. This could indicate some compositional change in the crude oil source that took place between the collection of GAC3 and GAC5.

3.6 The Size Distribution of Smaller ACCs in GAC Metagenomes are Bi-Modal

Examining the size distribution of all ACCs reconstructed across all samples showed that there was no density in the interval greater than 100 kbs and that most ACCs fall under the size range of 100 kbs or less. Interestingly, there was a bimodal size distribution with the highest densities at approximately 10 and 42 kbs, with a smaller peak at 60 kbs – and this was consistent across ACCs generated from each individual sample (Fig. 10).

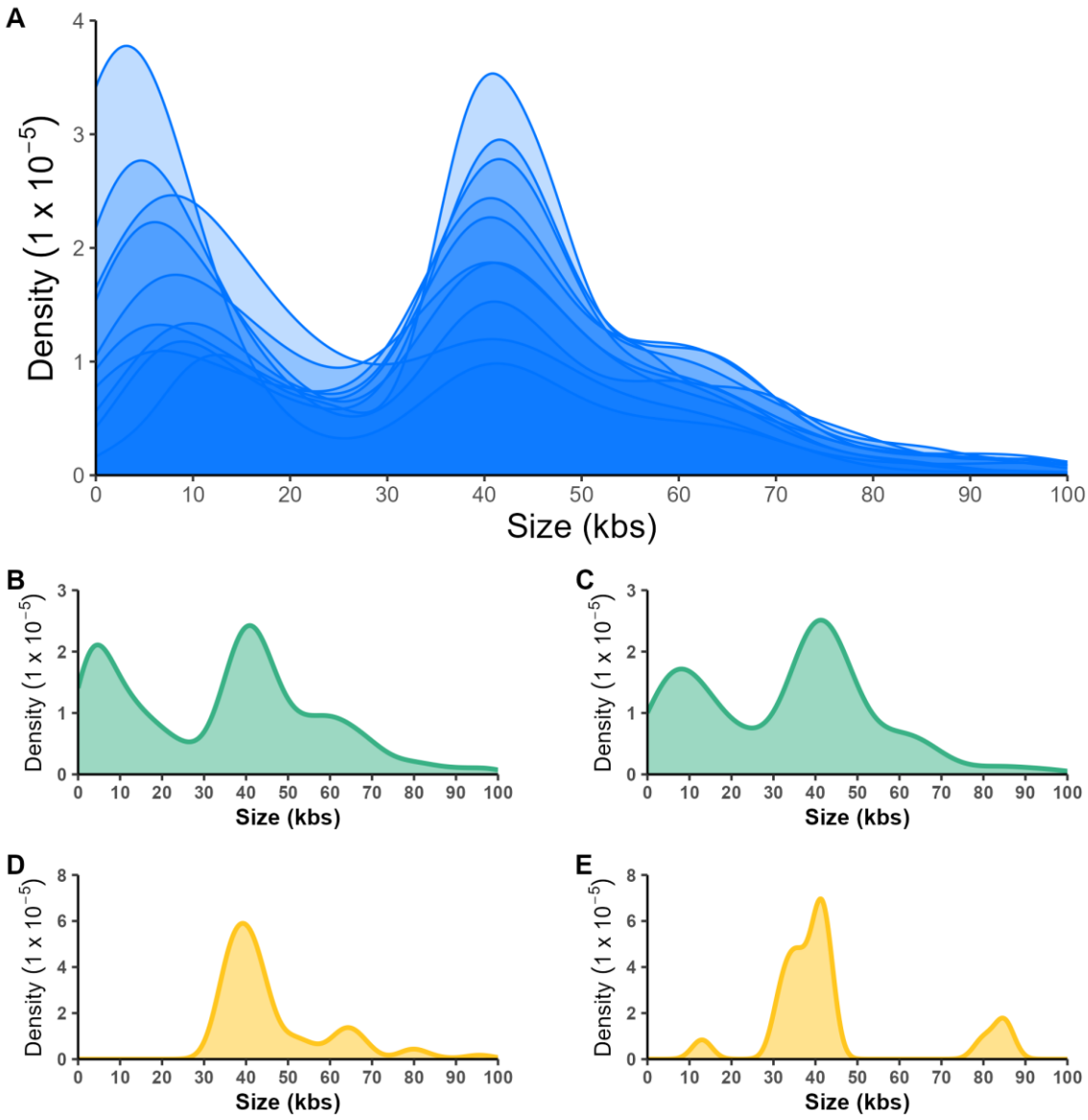

Figure 10. Size distributions of assembled circularized contigs (ACCs) were obtained from the metagenomes of 10 GAC samples. Each density curve represents the subset of draft assemblies under 100 kbs that are circularized in each of the 10 GAC samples. Each curve follows a bimodal distribution across the 10 GAC samples that were sequenced, with the majority of contigs in this range appearing at approximately 8 and 42 kbs. (B) ACCs that appeared only in 1 sample follow a similar bimodal distribution. Most polished assemblies fall into this category. (C) A subset of polished ACCs that appeared in more than 1 sample follows the same bimodal distribution but with a higher density at the 42 kbs size. (D) The

distribution of predicted bacteriophage ACCs that only appeared in a single sample differed in that bacteriophage in this subset were mainly 40 kbs and up. (E) Predicted bacteriophage that appeared in more than one sample were mainly those in the 42 kbs range, but also in the 85 and 12 kbs ranges.

This observation remains true even after subsetting small ACCs further into those that are seen only in one sample, and those seen in more than one sample. The determination of whether ACCs across samples were the same was done by performing an all-versus-all BLAST to find reciprocal best hits with a percent identity of at least 98%, 90% query coverage or more, and a query length that was at least 90% of the subject length. The resulting two subsets of singleton and recurring small ACCs had the same bimodal distribution with peaks at approximately 10 and 42 kbs, although slightly higher densities were seen in the 42 kbs peak for recurring ACCs.

In contrast, ACCs that were predicted to be phages had different size distributions. The subset of 227 circularized ACCs that contained at least one putative bacteriophage protein annotation by Bakta and that were predicted to be bacteriophage by INHERIT demonstrated different size distributions as shown in Figure 10D and E for those observed in one sample and those observed in multiple samples. The greatest density of putative bacteriophage in the single sample category was approximately at 40 kbs, with smaller peaks at 65, 80 and 95 kbs. The size distribution of putative bacteriophages in the recurring category had a major density peak at 42 kbs, with smaller peaks at 12, 35 and 85 kbs.

The consistency in size distribution for most subgenomic-sized ACCs across samples could either be due to a particular set of selectively advantageous genes, or even type of plasmid mobility. Previous studies that have also observed bi-modal distribution of plasmids within species observed correlations between mobility type and plasmid sizes (68). Additionally, it is possible that size constraints can result from bacteriophage transduction of plasmids which are limited by size compatibility - and similar to the observation of bacteriophage and plasmid size distributions made in the study by (68). Interestingly, the largest peak in our putative bacteriophage ACCs matches the largest of

the two peaks observed for all ACCs at approximately 42 kbs. Though it is not yet clear, this may be a mechanism that explains the bimodal distribution of ACCs up to 100 kbss seen here.

Chapter 4

4 Improving Annotations with a Structure-Based Approach

Discovering functional information about novel MAGs and other ACCs from GAC metagenomes is imperative to discovering NA-degrading genes, although this is challenging when we are working with the sequences only, and they are too novel. As mentioned in the last chapter, a large proportion of CDS predicted with sequence-based annotation tools remained hypothetical when annotated with Bakta. This is especially the case for putative phages, which is unsurprising as phages are notoriously difficult to annotate. In this chapter, I discuss a way in which we can improve our annotations greatly by using protein structure homology detection withs state-of-the-art tools.

Since protein structures are more conserved than sequence, it was expected that more informative annotations could be generated by identifying many of these hypothetical CDS' using structure homology. Hypothetical CDS were first passed to Colabfold (48) to predict their protein structures. Under the assumption that proteins with similar folds have similar functions, these structures were queried against the AlphaFold structure database to look for structural homology, and annotations of a query's best hit would be inherited.

Due to the sheer number of hypothetical CDS, limitations on computational resources, and lack of time, only CDS from putative phages were used as proof of principle for this annotation method.

4.1 Colabfold Produces High Confidence Structures

For this subset of predicted CDS' from novel putative phages, Colabfold was able to produce structures with most having mean pLDDT scores in the acceptable range of above 70 (Fig. 11). Mean PAE scores for predicted structures were higher than expected since a score of 5 or below is considered very confident, but since Foldseek considers

local similarities, correct domain positioning and orientations are likely to be less important for detecting significant structural similarities (51).

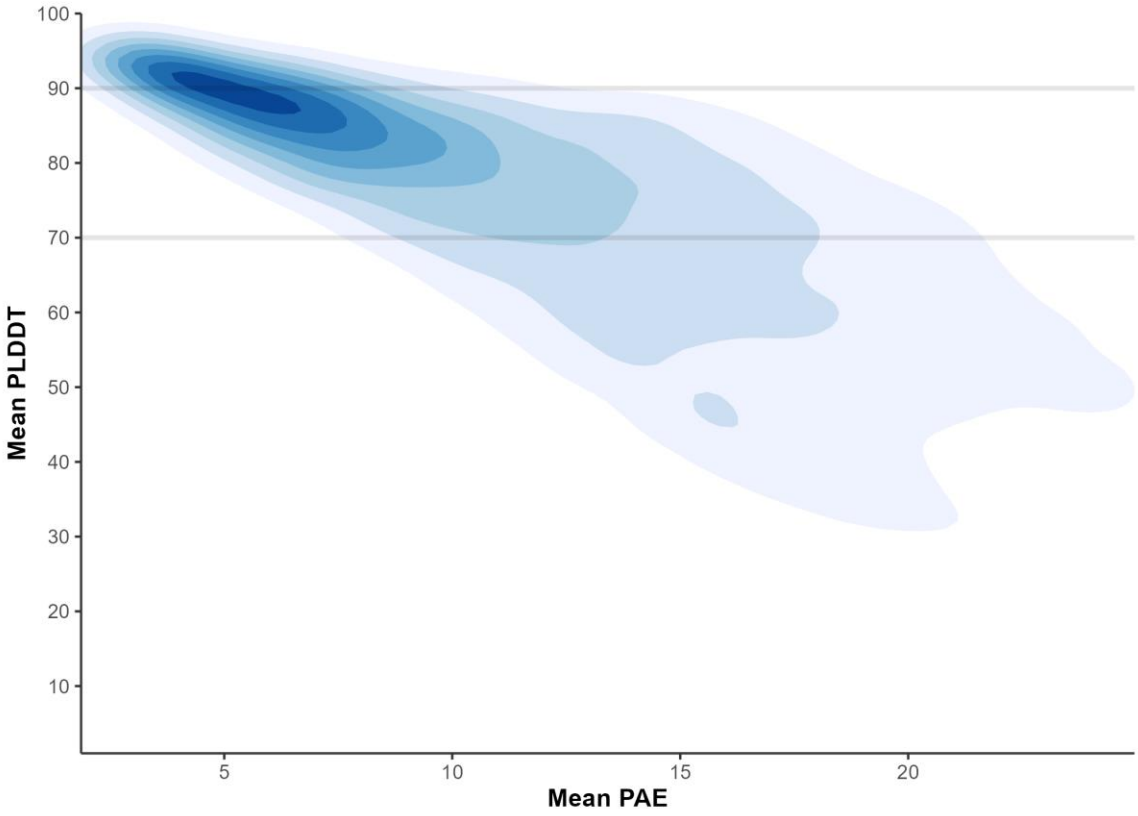

Figure 11. The majority of protein structures predicted with Colabfold from putative phage CDS' have pLDDT scores in the acceptable ranges of what is considered confident, which is a pLDDT score above 70. An average pLDDT score above 90 signifies a highly confident model. Though PAE scores for a large group of predicted structures are higher than 5, which is beyond the threshold for what is considered high confidence, it is not expected to significantly affect homology detection since Foldseek considers local similarities rather than simply global similarities.

Though it seems that most protein structure predictions were generally good, it is important to note that the evaluation of structure prediction quality with summarized pLDDT scores may mask inaccuracies in local regions of a structure with a higher mean score, or conversely, be too sensitive to small, disordered regions that significantly lower

the mean score. It would be more rigorous to evaluate the per-residue scores of each protein structure prediction to identify if regions are likely to be true errors, but mean pLDDT scores can still be a sufficient indicator of the quality of the overall fold. Overall, a majority of the structures produced by Colabfold were considered confident and usable for structure homology searching.

Quality predicted structures, which I considered to be structures with a mean pLDDT greater than 70, were then used to assess structural homology vs. the entire universe of known and predicted protein structures with Foldseek. Structural homology searching is much more sensitive than sequence homology searching, as only the fold and not the sequence need to be conserved (69). If a queried predicted structure's best hit has a high confidence homology to an annotated protein in the AlphaFold database where the e-value is less than 1e-10, the annotation was inherited by the predicted protein.

4.2 Foldseek Detects Structural Homology for Hypothetical Proteins and Enriches Annotations

Using this approach, a substantial increase in identifiable proteins and annotations with GO or KEGG terms was observed for all CDS in all putative phage ACCs across all samples when compared to sequence homology annotation alone (Fig. 12). In total, 5726 out of the 16606 hypothetical putative phage CDS had confident hits to structures in the AlphaFold database. While Bakta provided an annotation for an average of 10%-15% of CDS in all samples, the structural homology search approach gave confident predictions for approximately 30% to 40% of CDS (Fig. 12A). Furthermore, this approach revealed KEGG and GO terms that were otherwise nonexistent with Bakta annotations alone (Fig. 12B). Structure prediction and homology search with Colabfold and Foldseek enrich these annotations, providing an average of approximately 10% of CDSs with KEGG or

GO functional terms for each putative phage.

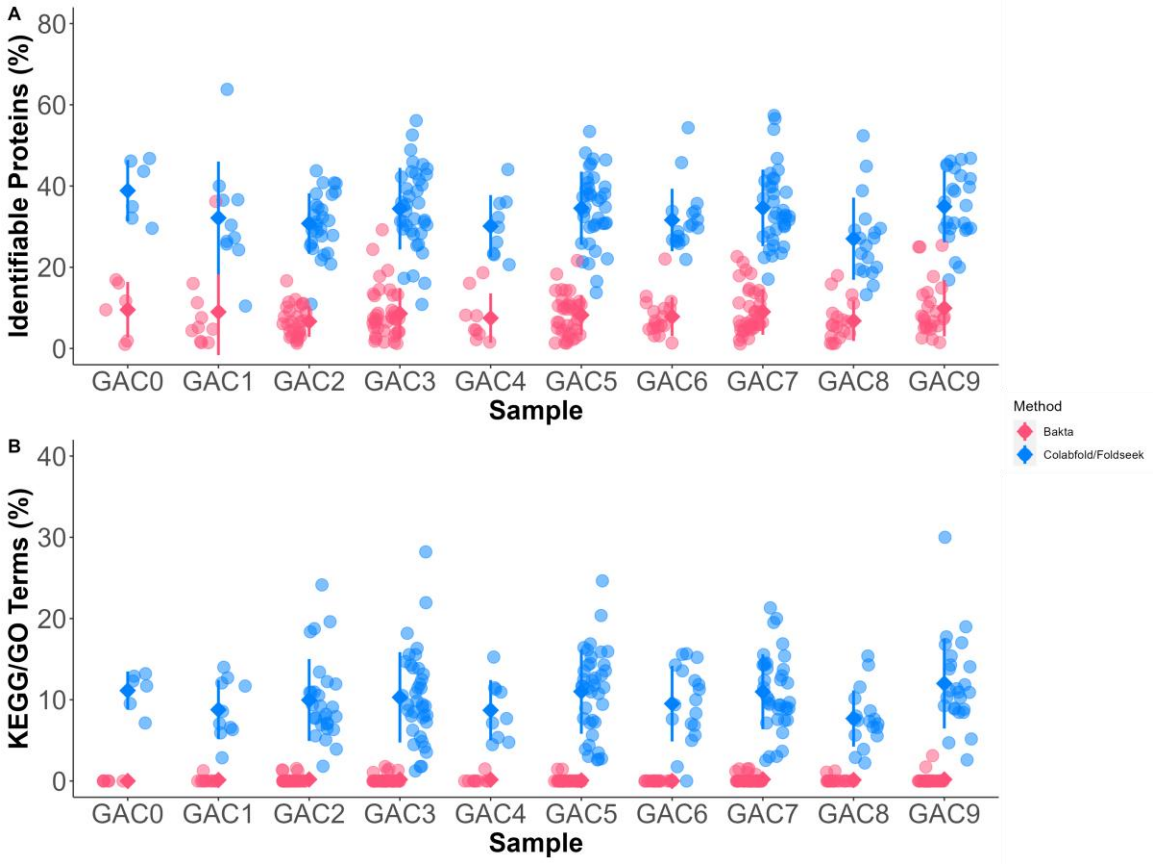

Figure 12. Using protein structure prediction and structure homology search methods (Colabfold and Foldseek) consistently increases the proportion of predicted CDS' that are annotated in bacteriophage across all samples versus sequence homology alone (Bakta). (A) A large proportion of proteins that remained hypothetical with Bakta were able to be identified using Colabfold and Foldseek. (B) The proportion of annotations that also include KEGG or GO terms increased consistently across all predicted bacteriophage across each sample. Whereas Bakta annotations had no KEGG/GO terms for most bacteriophage ACCs, Foldseek led to KEGG or GO annotations in all bacteriophage ACCs.

4.2.1 Structure Homology Approach Was More Performant Than Sequence Homology Across All CDS Sizes

The structural-based annotation approach identified more putative homologs than sequence-based annotation for all size classes of CDS up to 5.5 kb, with the greatest

number of putative homologs being identified in CDS between 500 bases and 1 kb (Fig. 13A). The count and proportion of CDS annotated by either method is summarized across 500 base bins in Figure 13A and 13B respectively. The sequence homology approach was able to annotate between 5% and 25% of predicted CDS' up to 5 kbs in size, whereas the structural homology method was able to annotate between 5% and 70% of CDS' up to 5.5 kb in size. Bakta was most efficient in annotating CDS of sizes 1 to 1.5 kbs and sizes 4.5 to 5 kbs – where about 25% of proteins in these size ranges were identifiable by this method. On the other hand, the structural homology approach was particularly efficient for CDS between 0.5 and 3 kbs, where between 50 and 70% of CDS in this size range were annotated by this approach. Although the proportion of CDS annotated by Foldseek was drastically lower outside this size range, there were still improvements in the proportion of annotated CDS across all size ranges in comparison to the sequence homology approach, except for the 4.5 to 5 kbs size range. Regardless, CDS identification using Bakta predicted CDS of up to 12 kbs in size, although both methods struggled to produce many annotations for CDS sizes 5 to 5.5 kbs and above.

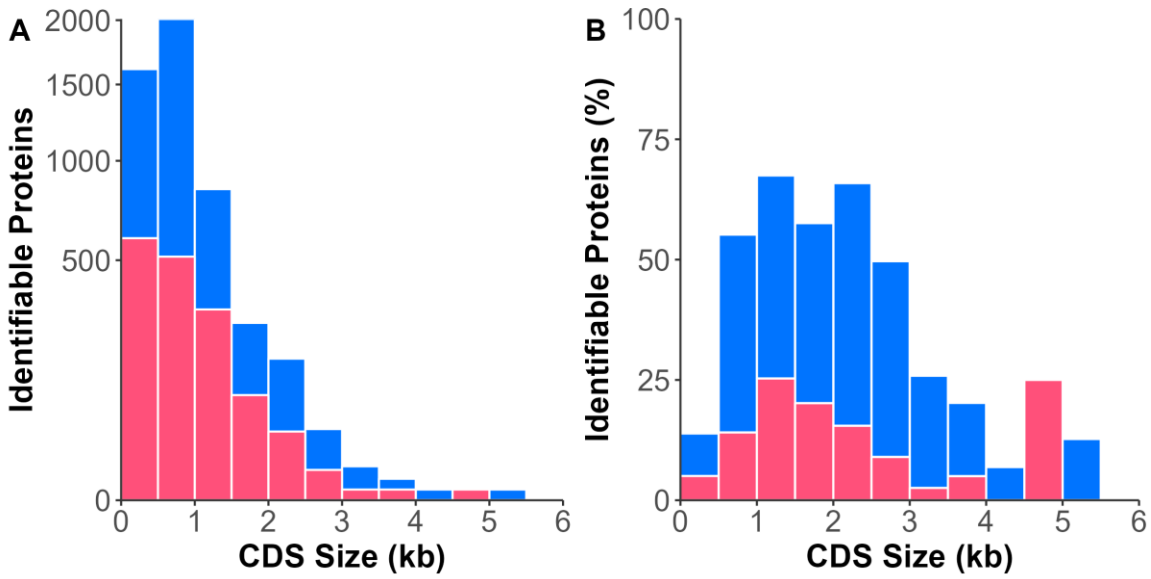

Figure 13. Including protein structure predictions and homology search based methods result in a larger proportion predicted CDS' of up to 5.5 kbs being annotated in predicted bacteriophage. (A) There is a universal increase in the proportions of CDS' annotated of sizes up to 5.5 kbs when using structure

homology. CDS' between 5.5 and 12 kbs were also predicted by Bakta but could not be annotated with either method. (B) The number of CDS' predicted and annotated is related to CDS size. While the number of identifiable proteins more than doubled when including structure homology methods for CDS' of sizes less than 5 kbs, an overwhelming proportion of proteins from CDS less than 0.5 kbs CDS size range remained unannotated.

4.2.2 GO and KEGG Terms

With the improvement in annotations, several functions and pathways present in putative bacteriophage ACCs in this sample were identified, many of which were functions and pathways expected to be found in bacteriophage (Fig. 14).

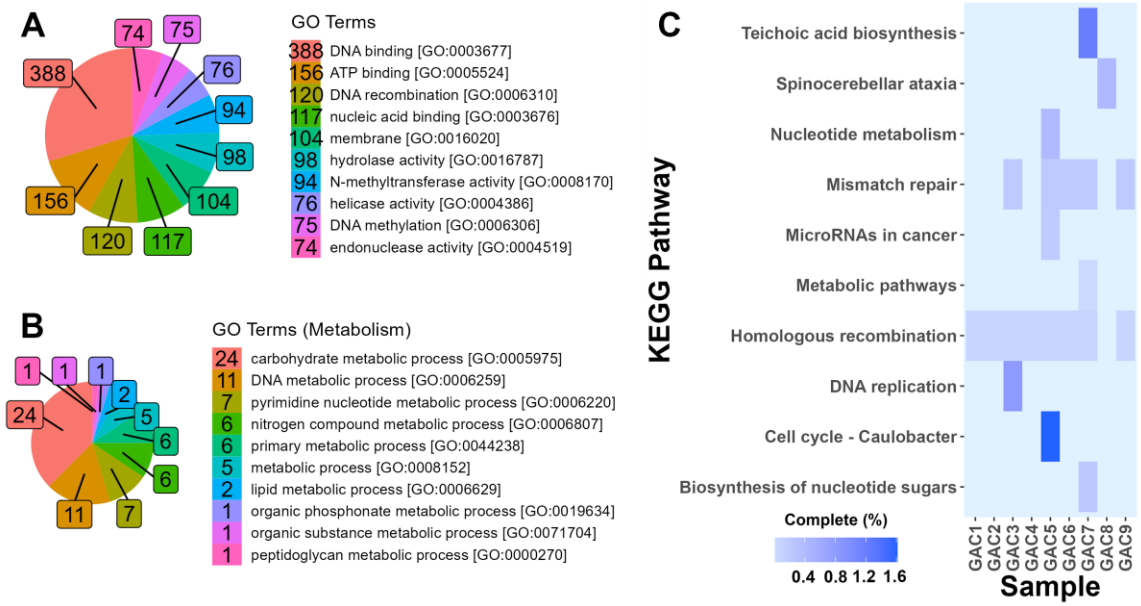

Figure 14. Foldseek revealed many functional pathways and processes present in putative bacteriophage ACCs, many of which were expected bacteriophage pathways. (A) The top 10 GO terms across all putative bacteriophage included mainly GO terms relevant to the bacteriophage life cycle, including DNA related functions and functions related to host infection and interaction. (B) Proteins involved in a variety of metabolic processes were identified across all the putative bacteriophages. Though carboxylic acid and lipid metabolic processes represent the minority, they may potentially be related to NA degradation based on past proposed

mechanisms of NA degradation. (C) The presence of KEGG pathways were identified from Foldseek annotations for the entire set of predicted protein structures in all putative bacteriophage in each sample. Some pathways common to bacteriophage, including mismatch repair and homologous recombination, were expected and present in most samples. All the pathways annotated in our samples were incomplete - KOs in putative bacteriophage Foldseek annotations across each sample represented only a small fraction of the total unique KOs from each KEGG pathway.

As expected for phages, the most common GO terms included processes related to DNA metabolism, repair, recombination, and replication processes, in addition to host infection or interaction such as hydrolase activity or membrane-related activity (Fig. 14A and Fig. 14B). KEGG pathways that appeared in Foldseek annotations also included expected pathways in most samples, including homologous recombination and mismatch repair pathways (Fig. 14C) .

The number of metabolic GO terms produced was diverse, but their relation to naphthenic acid degradation was not yet clear (Fig. 14B). The most common metabolic GO term was for carbohydrate-related metabolic processes. There were only 2 GO annotations in lipid metabolic processes that are unlikely to be involved with NA degradation, as the GO annotation is attached to glycerophosphoryl diester phosphodiesterase (GP-PDE) and a GP-PDE subunit, and this enzyme targets ester bonds between glycerol and phosphate. GP-PDE enzymes are known to be evolutionarily conserved proteins that are ubiquitous among eukaryotes and prokaryotes. In bacteria, they are typically involved with phospholipid membrane remodeling in Gram-positive bacteria, but also in bacterial pathogenicity. (70). GP-PDEs are also implicated in the removal of organophosphate esters in wastewater treatment plants (71), but these compounds are not typical in ORW. Regardless, many of these observations regarding the functional capabilities of our bacteriophage ACCs would not have been possible with a sequence homology-only approach.

4.2.3 PFAM Concordance

PFAM is a database of protein families that is useful for the identification of conserved domains. In Figure 15, I examined the concordance of PFAM annotations between the sequence and structural homology search approaches, including data from all samples. All PFAM annotations generated by either the sequence or structural homology approach were compared, and the latter was able to identify 583 PFAM annotations whereas the prior found only 296, with 183 (26%) of these being in common. Thus, the total PFAM identifications found by the structural approach was not a strict superset of all PFAM identifications found by sequence homology (Fig. 15A). A more rigorous way of comparing the annotation methods was to examine the overlap when both approaches provided an annotation for the same CDS shown in Figure 15B. Of the 194 CDS where both approaches predicted a CDS, 141 or 72% were concordant for PFAM identifications. Determining which approach was more accurate cannot be done in the absence of orthogonal evidence, but given the extremely high sequence divergence between the data collected here and the sequence databases it is possible that many of the sequence-only annotations may be spurious or of low confidence.

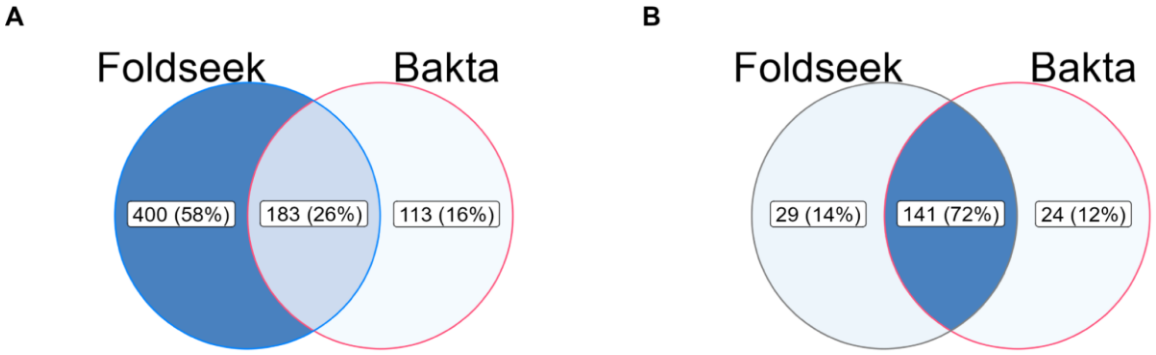

Figure 15. PFAM annotations were compared between Foldseek and Bakta. (A) All PFAMs annotations of predicted bacteriophage were pooled from all samples and compared for matches. (B) Pairwise comparisons were performed between Foldseek and Bakta PFAM annotations for each predicted CDS, to see if results from both methods agree. Matches are counted if at least one PFAM obtained from one method agrees with the other for a given CDS that was predicted by Bakta.

4.2.4 Skipping AMBER Relaxation in Structure Prediction

In generating the protein structures, there is an optional step to refine the predicted structure in which additional molecular dynamic simulations are performed by Colabfold based on AMBER force fields (72). AMBER describes a set of parameters needed to generate plausible bond lengths, angles, and side chain formations. These changes are often minor and target amino acid side chain positioning. Though useful in relaxing protein structures to their lowest energy state and eliminating side chain clashes, it is highly computational, even with GPU acceleration. To see how AMBER relaxation could affect homology detection with Foldseek, the best hits with relaxed and unrelaxed versions of a predicted structure were compared. Approximately 40% of best hits differed between unrelaxed and relaxed versions of a predicted protein, although unrelaxed and relaxed proteins were always assigned to the same PFAM. This suggests that there are several structures in which AMBER relaxation causes enough structural variation to alter homology detection slightly, but a good backbone prediction is enough to infer related functions. Therefore, a feasible strategy to reduce computational time could be to first perform structure predictions without the AMBER relax step, then subsequently perform relaxation to refine only proteins that are part of protein families with our functions of interest.

4.3 Detecting Phage Structural Proteins in Recurring Phages

All proteins of the bacteriophage head, tail, capsid or baseplate categories that the Foldseek annotations identified were clustered from all putative bacteriophage by structural alignments (Fig. 16). Near perfect TM scores within most clusters show that the same putative best structural homolog was often seen in samples widely separated by time, suggesting that similar bacteriophage genomic sequences were being captured on different collection dates. Unfortunately, however, relating bacteriophage in GAC samples to known families of bacteriophage based on the cluster representative was not possible due to the lack of bacteriophage specific taxonomic information in the UniProt entries, and the high sequence divergence precluded phylogenetic inference. However, future investigations considering factors such as the morphology of bacteriophage

structures, bacteriophage hosts and functional proteins to classify our bacteriophage can be done with the annotation methods highlighted in this thesis.

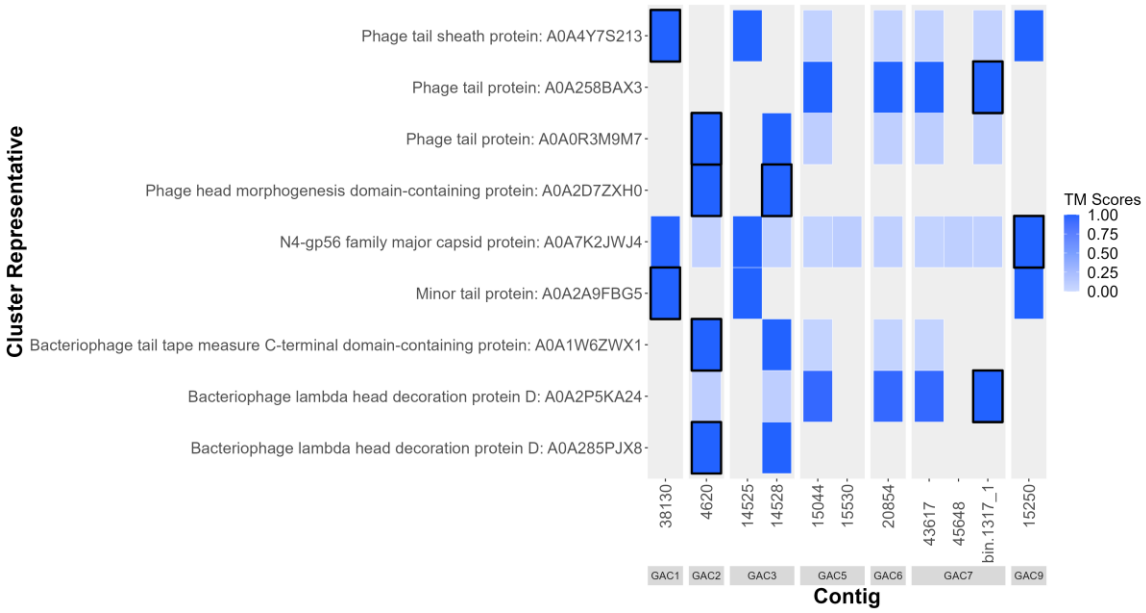

Figure 16. Structural bacteriophage proteins, which are often used for classification of bacteriophage, were identified via Foldseek and were shared between putative bacteriophage across samples. All predicted proteins structures belonging to recurring putative bacteriophage ACCs were clustered using Foldseek, and each cluster containing representatives that were annotated by Foldseek to be head, capsid, baseplate or tail related proteins had its representative (outlined black) structurally aligned to its members to generate a TM-score. Most clusters have members that strongly resemble its representative, having near perfect alignment scores close to 1 for the entire cluster. Clusters are named according to the UniProt recommended name for the representative protein structure, followed by its accession number.

Chapter 5

5 Discussion

In the previous chapters, I was able to sequence and characterize some features of MAGs and putative phages of GAC metagenomes – including taxonomic classifications, some interesting community dynamics and features, and some potential NA-degrading genes from annotations. I also discussed a new method for enriching annotations by using protein structure prediction and structural homology tools on novel sequences that are divergent from known sequences in current sequence databases. However, there are still many potential optimizations in the methods used and many unanswered questions from the data collected, which is what will be discussed in this chapter.

5.1 Further Work is Needed to Confirm NA-degrading Genes and Other Important Genes

Communities seeded in GAC from this oil refinery are diverse, with a seemingly high turnover rate of bacterial species, where very few predicted ACCs persist across samples. Based on the mean coverages of MAGs across all samples, the community appears to be mostly dominated by bacteria belonging to the Pseudomonadota phylum, where its subclasses Gammaproteobacteria were prominent in samples up to GAC3 and Alphaproteobacteria were more dominant in all samples afterwards. This is an observation that is often shared by other studies done on wastewater communities (63). These bacterial species that persist across samples or are exceedingly abundant relative to others in the community are likely to be important to NA degradation and long-term survival in ORW. Since ORW composition can be variable, changes in community biodiversity likely reflect those changes in composition. It is difficult, however, to make any inferences about specific factors that influenced the changes observed in the community since no data regarding the ORW composition was accessible for each time point. If changes in the toxicity or NA concentrations were correlated to the increase or decrease in abundance of a certain species across samples, a stronger case could potentially be made that a certain species is involved with NA degradation. Understanding which genes make a particular species robust to those changes in ORW is

also important if the goal in the future is to engineer ecological treatment systems for ORW. Genes apart from NA degradation to look out for can include membrane transporters or pumps, which were previously seen to be overexpressed in NA-exposed bacteria and were proposed to prevent NAs and other inorganic contaminants in ORW from accumulating in bacteria to toxic levels (19).

Although CANT-HYD identified a number of hydrocarbon degradation genes in high-quality MAGs including in those that persist across multiple samples, only two genes were manually confirmed to be previously identified NA degradation genes. Since there is a lack of specific databases or automated ways to help identify NA degradation genes in our samples, there may be more genes annotated that are actually involved with NA degradation but were not pointed out. Conversely, there are also some genes previously implicated with NA degradation that were absent in GAC samples. One majorly upregulated gene in *P. fluorescens* degrading a model NA is acyl-CoA thioesterase II, which does not appear at all from the CANT-HYD and Bakta annotations of all MAGs collected here (19). Either the enzyme remains unannotated by sequence homology methods and could potentially appear after enriching annotations with structure prediction and homology search, or the gene is simply not necessary or present in the samples collected here. Regardless, these GAC samples are exposed to a large variety of naturally occurring NAs and it would be expected that likely NA-degrading genes identified from a single isolate growing on NA models or commercial mixtures would be present at the very least in a bacterial community that degrades NA. As such, the lack of previously identified genes such as acyl-CoA thioesterase II could indicate some differences between model NAs and naturally occurring NAs.

ACCs that were smaller than 1 mb or were not identified as phages represented most ACCs collected from GAC, and these are yet to be analyzed. The remaining ACCs could likely be extrachromosomal elements such as plasmids that often carry advantageous genes, including potential NA-degrading genes. Finally, it is expected that many phages remain unidentified since phage prediction with INHERIT was only done on ACCs that had phage proteins in their Bakta annotations. This was initially done as a rough way to create a small subset of data to test the protein structure prediction and homology search

methods. Given the high sequence divergence observed in these samples and consequently poor Bakta annotations in many small ACCs, it is likely that a lot of phage proteins in ACCs went unannotated, as well as any potential linear phage chromosomes. In the future, it might be viable to use both INHERIT to predict if an ACC is a phage sequence and its structure homology-enriched annotations to confirm the presence of proteins unique to phages, such as the head, tail, baseplate, or capsid.

5.1.1 Comprehensive Annotations Supports Future Discovery of NA-degrading Genes

Initial annotations revealed many pathways in assembled MAGs, but there was a large proportion of unannotated sequences that could potentially be important to NA degradation or fitness in ORW. I was able to demonstrate a strategy to predict the protein structure of coding sequences unannotated by a sequence homology-based method using Colabfold and subsequently produce annotations for them by searching for structure homology with Foldseek.

Being able to generate more comprehensive annotations not only improves our general understanding of biological processes that occur in GAC samples but also opens the way to further analysis necessary to confidently predict NA-degrading genes. Comprehensive annotation data paired with metatranscriptomic data can allow for gene expression analysis, in which highly expressed genes could indicate genes important to the survival of bacteria in GAC. It is expected that a number of novel upregulated genes will be identified in GAC from this refinery relative to past studies since GAC is collected directly from ORW and is enriched in naturally occurring NAs, which contrasts the use of single species isolates growing on NA models for most studies.

However, there is the challenge of selecting a proper baseline of expression since ORW can be very dynamic and heterogeneous, and the proportion of NA types in ORW is not something that can be measured. One solution could be to use samples in which the NA concentrations are the lowest as the baseline and attempt to observe how gene expression changes over time with respect to NA concentrations.

5.2 Strategies for Reducing the Compute Time of Structure Prediction

Computationally predicting protein structures from sequences is the most time-consuming step of the pipeline, as Colabfold produces approximately 100 structures a day, depending on the size of the proteins being predicted. Over the 10 GAC samples collected in this study, 592730 structure predictions would be needed if every hypothetical CDS was to have its structure predicted – meaning that it would take years to process these samples.

5.2.1 Reducing Dataset Size by Clustering

Reducing the size of the dataset could be done by clustering protein sequences with 50% identity and at least 80% aligned residues, then subsequently predicting the structures of the representative sequences as a general model for its representatives. Though it is known that proteins with similar sequences often adopt similar structures, the inverse is not always true. Previous studies have identified that as sequence similarity decreases between proteins within the same family, the structural similarities exponentially get worse at the 50% threshold (73). Other studies have identified that protein pairs with sequence identities as low as 35-40% are still very likely to be structurally similar, and anything below that is referred to as the “twilight zone”, where protein pairs almost never share similar structures (74).

MMseqs2 (75) with the arguments “--min-seq-id 0.5 -s 7.5 -c 0.8 --cov-mode 1” was used to cluster all hypothetical CDS by at least 50% sequence identity, 80% aligned residues and a sensitivity level of 7.5. The sensitivity level is the average length of the lists of similar k-mers per query sequence position, so higher levels help identify sequence pairs with lower sequence identities in the prefiltering steps. Performing this clustering reduced the dataset by more than half – from 592730 to 248028 CDS.

Since there is still ambiguity in this sequence identity to structure relationship, clustering sequences under the assumption that members within the same cluster will have generally the same structure could still lead to a lot of false positives and negatives. Regardless, given the size of the dataset and the significant size reduction of the dataset after

clustering, it may be worth it to use this strategy simply to narrow down and prioritize proteins to predict and study.

5.3 Capturing Potential Missing Genomes

Many high-quality MAGs collected across GAC samples have particularly low mean coverages, which reduces the confidence of the assembly and consequently its annotations. Despite being complete and uncontaminated, those metrics do not account for chimeric assemblies containing closely related sequences. Manual inspection would be ideal to confirm the quality of low-coverage genomes, as gaps in coverage could indicate points where these sequences could have joined.

Apart from manual inspection, resequencing GAC samples to improve coverage and potentially uncover more genomes would be beneficial. Subsequent sequencing runs can be further optimized with adaptive sequencing to extract only the most necessary information. ONT's MinKNOW program allows for the depletion or enrichment of certain reads, so reads that belong to MAGs or assemblies that have already been assembled with sufficient coverage can be discarded. This would theoretically improve the throughput for reads belonging to genomes not yet recovered and would help with the recovery of additional genomes.

Furthermore, ONT offers other library preparation kits that may be highly beneficial to the recovery of additional genome assemblies. The latest v14 library preparation kits allow for raw read accuracies of Q20 and higher, which is a large improvement from the Q13 scores seen with the v10 chemistry kits used in this thesis. ONT also offers an ultra-long DNA sequencing kit that adopts their new v14 chemistry but also produces read N50's of over 50 kb. With the updated kits, the recovery of MAGs and other ACCs from GAC samples is likely to be much greater due to the greater base accuracy and more contiguous sequencing data.

It is also important to recognize that the number of phage genomes collected may be underrepresentative due to the inclusion of a size selection step after DNA extraction,

which depletes fragments smaller than 20 kilobases. This could have an impact on the coverage and recovery of phage genomes or other ACCs smaller than 20 kb.

Furthermore, the ACCs determined to be phages were limited to phage genomes that are circular or are circularly permuted, and the distinction between the two has yet to be made for the putative phages collected from GAC here. Phages can also contain linear genomes, and as such, many phages could potentially have gone unacknowledged.

5.3.1 "Incomplete" Genomes Could be Complete CPR Genomes

Although 80 out of the 112 ACCs greater than 1 mb were considered complete and uncontaminated, 17 ACCs had completeness values between 60 and 80% - which can be characteristic of CPR genomes. Furthermore, no ACC had completeness levels lower than 60%, or contamination levels higher than 5.22%. Though it is entirely possible that these contigs could simply be incomplete, circularly permuted fragments, it is also possible that these ACCs are CPR genomes, which characteristically have reduced genome sizes and many absent "universal" single copy marker genes. It would be unsurprising as well to find CPR genomes as they are relatively common in metagenomic samples.

Apart from identifying CPR by comparing its sequence against a reference database, CPR genomes could potentially be predicted by searching for both the presence and absence patterns of certain single-copy genes. CPR genomes have also been observed to have a lower GC ratio, as well as lacking certain ribosomal proteins such as uL1, bL9 and/or uL30 which are essentially universal for non-CPR genomes. In looking for potential CPR genomes, ACCs smaller than 1 mb must be included as well. CPR genomes have been observed to range from 0.3 to 1.7 mb, so there could be more than the 17 potential CPR genomes identified here (76).

Regardless, with CPR genomes typically having reduced metabolic capacities and thus relying on host organisms for exchanging metabolites, they may still play another important role in NA degradation. They are known to be involved with biogeochemical cycling and evolution within similar ecosystems, as a previous metagenomic study done

on CPR in activated sludge systems identified carbon cycling genes and horizontal transfer genes within CPR genomes (42). Since they adopt symbiont lifestyles, their role within the system or metabolic capabilities likely directly depends on their host.

5.4 Identifying and Pairing Phage to MAGs

Temperate lysogenic phages can integrate their DNA into bacterial chromosomes as prophage without immediately undergoing the lytic cycle (77). These genes could confer new functions, referred to as lysogenic conversion functions, and persist for multiple generations until the prophage is triggered to either enter the lytic cycle and kill the host, enter the chronic cycle. Prophage can be triggered as a response to changes in the environment, and as previously mentioned, ORW can be highly variable. Thus, phages can have a large impact on the dynamics of the community and may be important to continue to study. The existence of prophages could potentially be captured by mapping phage genomes back to MAG genomes. This way, it becomes possible to identify temperate phages, pinpoint their hosts and detect genes that are being horizontally transferred.

5.5 Comparison to Other Structure-Based Annotation Tools

In bridging the sequence to function gap, using the structure prediction and structure homology search tools Colabfold and Foldseek may be viable to enrich poor annotations of predicted gene sequences that are divergent from sequence databases, which is typical of environmental metagenomic samples. Even though Colabfold can mostly produce confident structures from this dataset, it is difficult to assess the accuracy of these annotations without orthogonal evidence – and while structure often correlates to function, inheriting annotations from a structure homolog is not perfect. Still, sequence similarity has been shown to be weakly correlated to functional similarity (78), while protein structures are often more conserved and correlated to function (69,79). In addition, benchmarking of structural alignments based on AlphaFold’s predicted structures against sequence alignments show that the prior is much more accurate in detecting homology, especially when the sequence identities between sequences fall

under 40% (80). Structural alignments appear to be the next step in homology detection, and it would be interesting to compare the results of the structure-based methods used here to enrich annotations, with other structure-based methods for annotation.

One such tool that is older but was widely used to perform sequence to structure functional annotation is DeepFRI, which uses a Graph Convolutional Network (GCN) with language model features (81). Very briefly, a language model is trained on domain sequences from Pfam to identify sequence features or patterns in relation to its functions and input them to a GCN. The GCN uses the language model to understand how those sequence features are related to each other in terms of their positions in a protein's 3D structure. By considering the physical proximity of residues in the 3D structure and propagating information between nearby residues, the GCN can identify long distance relationships between residues. The model then uses a gradient-weighted Class Activation Map which identifies specific residues in the protein's structure that are crucial for predicting its functions. Though DeepFRI was a relatively popular tool, the largest concern now would be that the sequence to structure steps could be highly inaccurate relative to AlphaFold-based tools, given that DeepFRI has not been validated in CASP and was only benchmarked against other sequence-based annotation methods. Thus, functional annotations from DeepFRI may be better than other sequence-based methods but are unlikely to be better than state-of-the-art annotations based on tools utilizing AlphaFold and structural aligners such as Foldseek.

5.6 Summary and Conclusion

Despite the scale at which wastewater is produced by the oil industry, its impact, and the promise of bioecological treatment systems for the remediation of wastewater, very little research on NA-degrading bacterial communities has been done previously, especially on the genomic level. The overall goal of this thesis project was to utilize state-of-the-art sequencing technologies and bioinformatic tools to better understand what a microbial community living in NA-enriched GAC filters looks like and how the community might be degrading NAs. Here, I demonstrated methods to extract, sequence, taxonomically assign, and annotate metagenomes from 10 GAC samples collected over two years. Based on its taxonomic classifications, I observed a diverse microbial community with a

lot of novel sequences and a high degree of species turnover across samples. Though no previously identified NA-degrading species were sequenced here, the 9 MAGs identified at the species level suggest the presence of chemolithotrophs, indicating that there is a population that does not degrade NA but is able to survive off other trace inorganic elements. Furthermore, most MAG sequences were unique across samples, but 10 groups of MAGs shared high sequence similarities in large regions and the same taxonomic classification across samples, which suggests the existence of a core genome and an accessory genome that changes often. Almost all MAGs that do persist across samples contain a number of hydrocarbon degradation genes identified by CANT-HYD that could potentially be involved with NA degradation, although KEGG or GO terms from Bakta annotations revealed no potential NA-degrading genes. Only 2 putative NA-degrading genes that have been identified before were seen in MAGs here, but it is expected that with the collection of metatranscriptome information, for example, more potential NA-degrading genes can be identified with gene expression analysis. There may also be more previously identified NA-degrading genes in GAC samples since the presence of these genes was confirmed by manually searching the literature since there is no collection or database of NA-degrading genes.

Another goal of the thesis was to enrich poor annotations to obtain more comprehensive overviews of the metabolic capabilities of the community, which would also support future analysis of gene expression. Typical annotation methods rely on sequence homology, but this can be difficult with novel sequences. Here I show on a subset of putative phages that annotations can be significantly improved on highly divergent sequences by predicting the structures of CDS and detecting structure homology. For putative phages, the proportion of CDS remaining hypothetical decreased approximately 5-fold on average, and at least 1 GO or KEGG term was produced for each phage. Most GO and KEGG terms produced were also expected to be phage-related processes. Predicting protein structures also allowed for the clustering of structural phage proteins, which helps support the fact that related phages are being captured independently across samples.

Using this strategy to annotate the rest of the ACCs collected and the MAGs would allow for the identification of many of the remaining hypothetical CDS, which represent more than half of all total CDS. However, there is much room for improvement. The estimated computational time of predicting structures for all the hypothetical CDS with Colabfold is approximately 2 years, but this could be reduced with less potential impact on annotation quality by clustering sequences by 50% identity and skipping AMBER relaxation. Other optimizations and improvements could be made by exploring other complementary function prediction tools, such as DEEPFRI.

Overall, this thesis demonstrates methods to characterize metagenomes from GAC, a unique environmental sample important to understanding NA biodegradation in ORW remediation. Functional information collected from these communities will be the foundation of ecological treatment systems that may one day be the industry standard for ORW remediation, given their potential for being the most cost-effective, efficient, and scalable treatment option.

References

1. Green SJ, Demes K, Arbeider M, Palen WJ, Salomon AK, Sisk TD, et al. Oil sands and the marine environment: current knowledge and future challenges. Front Ecol Environ. 2017;15(2):74–83.

2. Finkel ML. The impact of oil sands on the environment and health. Curr Opin Environ Sci Health. 2018 Jun 1;3:52–5.

3. Chow-Fraser G, Rougeot A, Gagnon E, Cheng R. 50 Years of Sprawiling Tailings - Mapping decades of destruction by oil sands tailings. Ross A, Gray P, editors. Environ Def Can CPAWS North Alta. 2022 May;

4. Brient JA, Wessner PJ, Doyle MN. Naphthenic Acids. In: Kirk-Othmer Encyclopedia of Chemical Technology [Internet]. John Wiley & Sons, Ltd; 2000 [cited 2023 Apr 19]. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/0471238961.1401160802180905.a01

5. Chakravarti R, Patrick BN, Barney M, Kusinski G, Devine TM. Toward the Mechanism of Corrosion in Crude Oil: A Study Using Vibrational Spectroscopic Techniques at Elevated Temperatures. Energy Fuels. 2013 Dec 19;27(12):7905–14.

6. Misiti T, Tezel U, Pavlostathis SG. Fate and effect of naphthenic acids on oil refinery activated sludge wastewater treatment systems. Water Res. 2013 Jan 1;47(1):449–60.

7. Quinlan PJ, Tam KC. Water treatment technologies for the remediation of naphthenic acids in oil sands process-affected water. Chem Eng J. 2015 Nov;279:696–714.

8. Afzal A, Drzewicz P, Pérez-Estrada LA, Chen Y, Martin JW, Gamal El-Din M. Effect of Molecular Structure on the Relative Reactivity of Naphthenic Acids in the UV/H_2O_2 Advanced Oxidation Process. Environ Sci Technol. 2012 Oct 2;46(19):10727–34.

9. Marentette JR, Frank RA, Bartlett AJ, Gillis PL, Hewitt LM, Peru KM, et al. Toxicity of naphthenic acid fraction components extracted from fresh and aged oil sands process-affected waters, and commercial naphthenic acid mixtures, to fathead minnow (Pimephales promelas) embryos. Aquat Toxicol Amst Neth. 2015 Jul;164:108–17.

10. Bartlett AJ, Frank RA, Gillis PL, Parrott JL, Marentette JR, Brown LR, et al. Toxicity of naphthenic acids to invertebrates: Extracts from oil sands process-affected water versus commercial mixtures. Environ Pollut Barking Essex 1987. 2017 Aug;227:271–9.

11. Wang J, Cao X, Sun J, Chai L, Huang Y, Tang X. Transcriptional responses of earthworm (Eisenia fetida) exposed to naphthenic acids in soil. Environ Pollut. 2015 Sep 1;204:264–70.

12. Hughes SA, Mahaffey A, Shore B, Baker J, Kilgour B, Brown C, et al. Using ultrahigh-resolution mass spectrometry and toxicity identification techniques to characterize the toxicity of oil sands process-affected water: The case for classical naphthenic acids. Environ Toxicol Chem. 2017 Nov;36(11):3148–57.

13. Frank RA, Kavanagh R, Kent Burnison B, Arsenault G, Headley JV, Peru KM, et al. Toxicity assessment of collected fractions from an extracted naphthenic acid mixture. Chemosphere. 2008 Jul;72(9):1309–14.

14. Pourrezaei P. Physico-Chemical Processes for Oil Sands Process-Affected Water Treatment [Internet] [Ph.D.]. [Canada -- Alberta, CA]: University of Alberta (Canada); [cited 2023 Jul 16]. Available from: https://www.proquest.com/docview/1353365453?pq-origsite=gscholar&fromopenview=true

15. Hsu CS, Dechert GJ, Robbins WK, Fukuda EK. Naphthenic Acids in Crude Oils Characterized by Mass Spectrometry. Energy Fuels. 2000 Jan 1;14(1):217–23.

16. Grady CPL, Daigger GT, Love NG, Filipe CDM. Biological Wastewater Treatment. CRC Press; 2011. 994 p.

17. Clemente JS, Fedorak PM. A review of the occurrence, analyses, toxicity, and biodegradation of naphthenic acids. Chemosphere. 2005 Jul;60(5):585–600.

18. Clemente JS, MacKinnon MD, Fedorak PM. Aerobic Biodegradation of Two Commercial Naphthenic Acids Preparations. Environ Sci Technol. 2004 Feb 1;38(4):1009–16.

19. McKew BA, Johnson R, Clothier L, Skeels K, Ross MS, Metodiev M, et al. Differential protein expression during growth on model and commercial mixtures of naphthenic acids in Pseudomonas fluorescens Pf-5. MicrobiologyOpen. 2021 Jul 19;10(4):e1196.

20. Whitby C. Chapter 3 - Microbial Naphthenic Acid Degradation. In: Advances in Applied Microbiology [Internet]. Academic Press; 2010 [cited 2023 Jun 11]. p. 93–125. (Advances in Applied Microbiology; vol. 70). Available from: https://www.sciencedirect.com/science/article/pii/S0065216410700034

21. Blakley ER. The microbial degradation of cyclohexanecarboxylic acid by a β-oxidation pathway with simultaneous induction to the utilization of benzoate. Can J Microbiol. 1978 Jul;24(7):847–55.

22. Rho EM, Evans WC. The aerobic metabolism of cyclohexanecarboxylic acid by Acinetobacter anitratum. Biochem J. 1975 Apr 1;148(1):11–5.

23. Arslan M, Müller JA, Gamal El-Din M. Aerobic naphthenic acid-degrading bacteria in petroleum-coke improve oil sands process water remediation in biofilters: DNA-

stable isotope probing reveals methylotrophy in Schmutzdecke. Sci Total Environ. 2022 Apr 1;815:151961.

24. Johnson RJ, West CE, Swaih AM, Folwell BD, Smith BE, Rowland SJ, et al. Aerobic biotransformation of alkyl branched aromatic alkanoic naphthenic acids via two different pathways by a new isolate of Mycobacterium. Environ Microbiol. 2012;14(4):872–82.

25. Blakley ER, Papish B. The metabolism of cyclohexanecarboxylic acid and 3-cyclohexenecarboxylic acid by Pseudomonas putida. Can J Microbiol. 1982 Dec;28(12):1324–9.

26. Beckett A, Cook K, Robson S. A pandemic in the age of next-generation sequencing. The Biochemist. 2021 Dec 16;43.

27. Khrenova MG, Panova TV, Rodin VA, Kryakvin MA, Lukyanov DA, Osterman IA, et al. Nanopore Sequencing for De Novo Bacterial Genome Assembly and Search for Single-Nucleotide Polymorphism. Int J Mol Sci. 2022 Aug 2;23(15):8569.

28. Esposito A, Esposito M, Ptashnik A. Phylogenetic Diversity of Animal Oral and Gastrointestinal Viromes Useful in Surveillance of Zoonoses. Microorganisms. 2022 Sep 10;10:1815.

29. Slatko BE, Gardner AF, Ausubel FM. Overview of Next Generation Sequencing Technologies. Curr Protoc Mol Biol. 2018 Apr;122(1):e59.

30. Oxford Nanopore Technologies [Internet]. 2021 [cited 2023 Jun 14]. The power of Q20+ chemistry. Available from: https://nanoporetech.com/q20plus-chemistry

31. Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. Nat Methods. 2020 Nov;17(11):1103–10.

32. Alhakami H, Mirebrahim H, Lonardi S. A comparative evaluation of genome assembly reconciliation tools. Genome Biol. 2017 May 18;18:93.

33. Wick RR, Holt KE. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. F1000Research. 2019;8:2138.

34. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 2017 May;27(5):737–46.

35. Medaka [Internet]. Oxford Nanopore Technologies; 2018. Available from: https://nanoporetech.github.io/medaka/index.html

36. Wick R. rrwick/Minipolish: Minipolish v0.1.3 [Internet]. Zenodo; 2020 [cited 2023 Jul 16]. Available from: https://zenodo.org/record/3752204

37. Bacheli A. Gerenuq [Internet]. 2020 [cited 2021 Apr 4]. Available from: https://github.com/abahcheli/gerenuq

38. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012 Jun 1;28(11):1420–8.

39. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol. 2012 May;19(5):455–77.

40. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol. 2017 Aug 1;35(8):725–31.

41. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015 Jul;25(7):1043–55.

42. Wang Y, Zhang Y, Hu Y, Liu L, Liu SJ, Zhang T. Genome-centric metagenomics reveals the host-driven dynamics and ecological role of CPR bacteria in an activated sludge system. Microbiome. 2023 Mar 22;11(1):56.

43. Lui LM, Nielsen TN, Arkin AP. A method for achieving complete microbial genomes and improving bins from metagenomics data. PLOS Comput Biol. 2021 May 7;17(5):e1008972.

44. Castelle CJ, Banfield JF. Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. Cell. 2018 Mar 8;172(6):1181–97.

45. Deng H, Jia Y, Zhang Y. Protein structure prediction. Int J Mod Phys B. 2018 Jul 20;32(18):1840009.

46. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug;596(7873):583–9.

47. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 2021 Nov 17;50(D1):D439–44.

48. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nat Methods. 2022 Jun;19(6):679–82.

49. Orengo CA, Todd AE, Thornton JM. From protein structure to function. Curr Opin Struct Biol. 1999 Jun;9(3):374–82.

50. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. Microb Genomics. 2021 Nov;7(11):000685.

51. Kempen M van, Kim SS, Tumescheit C, Mirdita M, Söding J, Steinegger M. Foldseek: fast and accurate protein structure search [Internet]. bioRxiv; 2022 [cited 2023 Apr 26]. p. 2022.02.07.479398. Available from: https://www.biorxiv.org/content/10.1101/2022.02.07.479398v1

52. Leger A, Leonardi T. pycoQC, interactive quality control for Oxford Nanopore Sequencing. J Open Source Softw. 2019 Feb 28;4:1236.

53. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics. 2018 Aug 1;34(15):2666–9.

54. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018 Sep 15;34(18):3094–100.

55. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ. 2019 Jul 26;7:e7359.

56. Dawson ET, Durbin R. GFAKluge: A C++ library and command line utilities for the Graphical Fragment Assembly formats. J Open Source Softw. 2019;4(33):1083.

57. Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics. 2018 Mar 1;34(5):867–8.

58. Bai Z, Zhang Y zhong, Miyano S, Yamaguchi R, Fujimoto K, Uematsu S, et al. Identification of bacteriophage genome sequences with representation learning. Bioinformatics. 2022 Sep 15;38(18):4264–70.

59. Tenenbaum D, Volkening J, Maintainer BP. KEGGREST: Client-side REST access to the Kyoto Encyclopedia of Genes and Genomes (KEGG) [Internet]. Bioconductor version: Release (3.17); 2023 [cited 2023 Jul 16]. Available from: https://bioconductor.org/packages/KEGGREST/

60. Leppik RA, Young IG, Gibson F. Membrane-associated reactions in ubiquinone biosynthesis in Escherichia coli. 3-Octaprenyl-4-hydroxybenzoate carboxy-lyase. Biochim Biophys Acta. 1976 Jul 15;436(4):800–10.

61. Cunha RD, Ferreira LJ, Orestes E, Coutinho-Neto MD, de Almeida JM, Carvalho RM, et al. Naphthenic Acids Aggregation: The Role of Salinity. Computation. 2022 Oct;10(10):170.

62. Takahashi-Íñiguez T, Aburto-Rodríguez N, Vilchis-González AL, Flores ME. Function, kinetic properties, crystallization, and regulation of microbial malate dehydrogenase. J Zhejiang Univ Sci B. 2016 Apr;17(4):247–61.

63. Ahad JME, Pakdel H, Gammon PR, Siddique T, Kuznetsova A, Savard MM. Evaluating in situ biodegradation of 13C-labelled naphthenic acids in groundwater near oil sands tailings ponds. Sci Total Environ. 2018 Dec;643:392–9.

64. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. Bioinformatics. 2022 Dec 1;38(23):5315–6.

65. Bunce JT, Ndam E, Ofiteru ID, Moore A, Graham DW. A Review of Phosphorus Removal Technologies and Their Applicability to Small-Scale Domestic Wastewater Treatment Systems. Front Environ Sci [Internet]. 2018 [cited 2023 Jul 19];6. Available from: https://www.frontiersin.org/articles/10.3389/fenvs.2018.00008

66. Singleton CM, Petriglieri F, Kristensen JM, Kirkegaard RH, Michaelsen TY, Andersen MH, et al. Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing. Nat Commun. 2021 Mar 31;12:2009.

67. Karthikeyan S, Rodriguez-R LM, Heritier-Robbins P, Kim M, Overholt WA, Gaby JC, et al. "Candidatus Macondimonas diazotrophica", a novel gammaproteobacterial genus dominating crude-oil-contaminated coastal sediments. ISME J. 2019 Aug;13(8):2129–34.

68. Ares-Arroyo M, Coluzzi C, P.C. Rocha E. Origins of transfer establish networks of functional dependencies for plasmid transfer by conjugation. Nucleic Acids Res [Internet]. 2022 Nov; Available from: https://doi.org/10.1093/nar/gkac1079

69. Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. Curr Opin Struct Biol. 1996 Jun;6(3):377–85.

70. Wang F, Lai L, Liu Y, Yang B, Wang Y. Expression and Characterization of a Novel Glycerophosphodiester Phosphodiesterase from Pyrococcus furiosus DSM 3638 That Possesses Lysophospholipase D Activity. Int J Mol Sci. 2016 May 30;17(6):831.

71. Pantelaki I, Voutsa D. Occurrence and removal of organophosphate esters in municipal wastewater treatment plants in Thessaloniki, Greece. Environ Res. 2022 Nov;214(Pt 2):113908.

72. Salomon-Ferrer R, Case DA, Walker RC. An overview of the Amber biomolecular simulation package. WIREs Comput Mol Sci. 2013;3(2):198–210.

73. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J. 1986 Apr;5(4):823–6.

74. Rost B. Twilight zone of protein sequence alignments. Protein Eng Des Sel. 1999 Feb 1;12(2):85–94.

75. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017 Nov;35(11):1026–8.

76. Tsurumaki M, Saito M, Tomita M, Kanai A. Features of smaller ribosomes in candidate phyla radiation (CPR) bacteria revealed with a molecular evolutionary analysis. RNA. 2022 Aug;28(8):1041–57.

77. Howard-Varona C, Hargreaves KR, Abedon ST, Sullivan MB. Lysogeny in nature: mechanisms, impact and ecology of temperate phages. ISME J. 2017 Jul;11(7):1511–20.

78. Clark WT, Radivojac P. Analysis of protein function and its prediction from amino acid sequence. Proteins Struct Funct Bioinforma. 2011;79(7):2086–96.

79. Illergård K, Ardell DH, Elofsson A. Structure is three to ten times more conserved than sequence--a study of structural response in protein cores. Proteins. 2009 Nov 15;77(3):499–508.

80. Rajapaksa S, Konagurthu AS, Lesk AM. Sequence and structure alignments in post-AlphaFold era. Curr Opin Struct Biol. 2023 Apr;79:102539.

81. Gligorijević V, Renfrew PD, Kosciolek T, Leman JK, Berenberg D, Vatanen T, et al. Structure-based protein function prediction using graph convolutional networks. Nat Commun. 2021 May 26;12(1):3168.

Appendices

Appendix A: CheckM Results for all ACCs greater than 1 mb across all GAC samples. Completeness (Comp) percentages greater than 90 and contamination (Contam) percentages less than 5 are considered complete MAGs.

Sample	Contig Name	Marker Lineage	Comp	Contam
GAC8	bin.944	p__Bacteroidetes(UID2605)	100	0.48
GAC7	bin.756_6	c__Alphaproteobacteria(UID3305)	99.13	0
GAC2	1038	c__Betaproteobacteria(UID3971)	98.93	0.03
GAC5	bin.596_1	c__Alphaproteobacteria(UID3305)	98.7	0
GAC8	bin.321_1	c__Alphaproteobacteria(UID3305)	98.69	0
GAC8	419	c__Gammaproteobacteria(UID4267)	98.59	0.56
GAC6	bin.618_2	p__Bacteroidetes(UID2605)	98.57	0.71
GAC6	bin.616_1	c__Alphaproteobacteria(UID3305)	98.57	0
GAC2	595	c__Gammaproteobacteria(UID4267)	98.51	0.56
GAC6	2268	c__Betaproteobacteria(UID3971)	98.5	0
GAC8	2304	c__Gammaproteobacteria(UID4267)	98.47	0.52
GAC6	bin.472_3	c__Betaproteobacteria(UID3971)	98.29	0
GAC5	5304	p__Bacteroidetes(UID2591)	98.28	0
GAC3	13374	c__Gammaproteobacteria(UID4274)	98.22	1.42
GAC3	2150	k__Bacteria(UID3187)	98.21	0.89
GAC7	bin.526_1	o__Sphingomonadales(UID3310)	98.16	0.94
GAC5	1636	c__Gammaproteobacteria(UID4274)	98.16	1.42
GAC1	466	c__Gammaproteobacteria(UID4274)	98.14	1.42
GAC2	5322	c__Gammaproteobacteria(UID4274)	97.87	1.42
GAC6	2932	c__Gammaproteobacteria(UID4274)	97.79	1.25
GAC8	bin.1128_3	o__Sphingomonadales(UID3310)	97.75	0.94
GAC6	bin.302_1	c__Gammaproteobacteria(UID4201)	97.7	0
GAC8	bin.778_1	k__Bacteria(UID3187)	97.67	3.64
GAC3	1330	o__Cytophagales(UID2936)	97.6	0.6
GAC8	6081	c__Gammaproteobacteria(UID4274)	97.56	1.18
GAC3	694	c__Gammaproteobacteria(UID4267)	97.5	0.56
GAC7	bin.442_1	c__Gammaproteobacteria(UID4267)	97.47	0.79
GAC7	7443	c__Gammaproteobacteria(UID4274)	97.47	1.42
GAC2	1044	c__Gammaproteobacteria(UID4267)	97.42	1.7
GAC3	bin.747_1	c__Alphaproteobacteria(UID3305)	97.4	0.43
GAC5	bin.229_2	c__Gammaproteobacteria(UID4267)	97.39	0.67
GAC0	4282	k__Bacteria(UID3187)	97.22	2.73
GAC7	bin.1152_1	c__Betaproteobacteria(UID3971)	97.14	0
GAC7	bin.595_13	c__Gammaproteobacteria(UID4267)	97.12	3.55
GAC7	bin.915_2	c__Betaproteobacteria(UID3888)	97.09	0

GAC3	476	c__Gammaproteobacteria(UID4267)	97	0.44
GAC5	3168	c__Betaproteobacteria(UID3971)	96.8	0
GAC0	10	c__Gammaproteobacteria(UID4267)	96.65	0.79
GAC6	bin.113_1	c__Gammaproteobacteria(UID4274)	96.65	1.42
GAC8	2786	k__Bacteria(UID3187)	96.58	1.71
GAC8	3901	k__Bacteria(UID3187)	96.58	5.22
GAC2	bin.525_1	c__Gammaproteobacteria(UID4267)	96.56	0.91
GAC0	bin.402_8	c__Gammaproteobacteria(UID4443)	96.05	0.37
GAC8	bin.228_38	c__Gammaproteobacteria(UID4267)	96.03	2.17
GAC2	bin.191_3	c__Alphaproteobacteria(UID3305)	95.85	1.09
GAC3	bin.780_3	f__Rhodobacteraceae(UID3340)	95.83	0
GAC8	bin.666_4	k__Bacteria(UID3187)	95.73	3.85
GAC6	1770	k__Bacteria(UID3187)	95.73	3.85
GAC2	bin.761_3	k__Bacteria(UID2565)	95.64	3.76
GAC2	bin.666_1	c__Alphaproteobacteria(UID3305)	95.61	0.43
GAC3	bin.1013_1	k__Bacteria(UID2495)	95.6	0.1
GAC3	bin.670_11	f__Xanthomonadaceae(UID4214)	95.59	1.15
GAC5	bin.252_9	c__Alphaproteobacteria(UID3305)	94.98	0.44
GAC3	1384	k__Bacteria(UID3187)	94.84	0.91
GAC6	bin.867_1	c__Alphaproteobacteria(UID3305)	94.76	0.44
GAC7	bin.732_1	k__Bacteria(UID2495)	94.38	1.1
GAC3	632	c__Gammaproteobacteria(UID4267)	94.25	0.83
GAC7	bin.837_11	k__Bacteria(UID3187)	94.18	1.71
GAC8	24942	k__Bacteria(UID2495)	93.96	1.1
GAC2	bin.405_1	c__Alphaproteobacteria(UID3305)	93.92	0
GAC5	bin.182_6	c__Alphaproteobacteria(UID3337)	93.63	0
GAC8	1413	o__Sphingomonadales(UID3310)	93.43	0.94
GAC7	bin.1232_1	p__Bacteroidetes(UID2591)	93.32	0.66
GAC6	309	k__Bacteria(UID2495)	93.16	1.1
GAC8	bin.73_1	c__Gammaproteobacteria(UID4267)	93.15	3.79
GAC6	bin.888_1	k__Bacteria(UID2565)	93.11	3.41
GAC3	bin.878_2	c__Deltaproteobacteria(UID3216)	92.9	1.29
GAC3	4050	c__Alphaproteobacteria(UID3305)	92.61	1.3
GAC7	397	c__Alphaproteobacteria(UID3305)	92.59	0.44
GAC7	bin.516_1	c__Alphaproteobacteria(UID3305)	92.42	0.44
GAC2	1115	c__Gammaproteobacteria(UID4267)	92.39	0.4
GAC8	bin.257_2	k__Bacteria(UID3187)	92.26	0.85
GAC3	5308	k__Bacteria(UID3187)	92.14	4.32
GAC3	bin.914_3	k__Bacteria(UID2565)	91.98	2.27
GAC3	11397	c__Gammaproteobacteria(UID4266)	91.82	0
GAC9	13052	c__Alphaproteobacteria(UID3305)	91.74	0.88
GAC8	bin.1065	k__Bacteria(UID2565)	91.51	3.23
GAC2	11871	c__Alphaproteobacteria(UID3305)	91.21	0

GAC9	178	k__Bacteria(UID3187)	90.55	4.27
GAC8	bin.719_1	k__Bacteria(UID2565)	90.52	0
GAC2	1852	c__Deltaproteobacteria(UID3216)	89.84	1.29
GAC0	4784	c__Deltaproteobacteria(UID3216)	89.77	0.65
GAC5	bin.531_1	f__Rhodocyclaceae(UID3972)	88.07	0
GAC2	bin.547_3	k__Bacteria(UID2565)	88	1.14
GAC3	7970	k__Bacteria(UID3187)	87.87	0.23
GAC3	bin.90_1	k__Bacteria(UID2565)	87.29	0
GAC3	4180	k__Bacteria(UID2565)	86.33	2.33
GAC7	245	k__Bacteria(UID2565)	85.76	3.41
GAC8	bin.941	k__Bacteria(UID3187)	85.18	0.85
GAC9	17507	k__Bacteria(UID203)	84.61	0
GAC8	5287	k__Bacteria(UID1452)	84.56	0.33
GAC6	bin.663_1	k__Bacteria(UID2565)	84.02	0
GAC8	bin.279	k__Bacteria(UID3187)	82.46	0
GAC3	4966	k__Bacteria(UID1452)	82.14	1.1
GAC8	7304	k__Bacteria(UID203)	81.4	0
GAC9	5178	k__Bacteria(UID3187)	81.33	0.85
GAC2	690	k__Bacteria(UID3187)	80.37	0.96
GAC7	288	k__Bacteria(UID2565)	79.69	2.27
GAC8	4444	k__Bacteria(UID203)	79.44	0
GAC8	5562	k__Bacteria(UID2565)	77.7	0.57
GAC7	14392	k__Bacteria(UID203)	77.43	0
GAC5	3791	k__Bacteria(UID203)	77.27	0
GAC6	3344	k__Bacteria(UID203)	75.71	0
GAC3	3758	k__Bacteria(UID2329)	75.04	0.16
GAC5	bin.353_1	k__Bacteria(UID2565)	74.36	0.57
GAC8	5819	k__Bacteria(UID203)	73.12	1.72
GAC9	3209	k__Bacteria(UID2495)	72.22	0
GAC3	bin.771_10	c__Gammaproteobacteria(UID4267)	71.27	0
GAC8	4053	k__Bacteria(UID1452)	70.83	1.85
GAC5	3906	k__Bacteria(UID1452)	70.83	0.93
GAC3	2748	k__Bacteria(UID1452)	70.3	0
GAC6	3120	k__Bacteria(UID1453)	68.95	0.85
GAC7	11047	k__Bacteria(UID1453)	68.95	0.85
GAC8	bin.560_1	k__Bacteria(UID1452)	68.48	0
GAC8	4482	k__Bacteria(UID1452)	68.48	0
GAC3	14643	k__Bacteria(UID1452)	67.73	1.98
GAC8	15073	k__Bacteria(UID1452)	66.01	0
GAC7	1113	k__Bacteria(UID2565)	64.27	1.14
GAC3	13560	k__Bacteria(UID1453)	64.25	0
GAC3	4258	k__Bacteria(UID1453)	64.07	0
GAC3	4773	k__Bacteria(UID1452)	63.86	0

GAC5	bin.222_2	k__Bacteria(UID2565)	63.69	0
GAC8	1565	k__Bacteria(UID2565)	61.94	2.27

Appendix B: CANT-HYD hydrocarbon degradation genes were found in various high-quality MAGS across GAC samples. The number of MAGs containing a given hydrocarbon degradation gene is shown across each sample. Blank entries have no annotated genes.

Hydrocarbon Degradation Gene	GAC									
	0	1	2	3	4	5	6	7	8	9
(2Fe-2S)-binding_protein			1	1	1		1			
2-halobenzoate_1,2-dioxygenase_large_subunit					2			1	1	
3-ketosteroid-9-alpha-hydroxylase_oxygenase_subunit									1	
3-octaprenyl-4-hydroxybenzoate_carboxy-lyase	1	1	4	5	4	3	5	3	2	
3-phenylpropionate/cinnamic_acid_dioxygenase_subunit_alpha			1							
3-phenylpropionate/cinnamic_acid_dioxygenase_subunit_beta			1	2						
4-(Gamma-L-glutamylamino)butanoyl-[BtrI_acyl-carrier_protein]_monooxygenase_BtrO					1		1			
4Fe-4S_Mo/W_bis-MGD-type_domain-containing_protein	1		3	2	2			4		
4-hydroxy-3-polyprenylbenzoate_decarboxylase			1			1	2	1		
4-hydroxyacetophenone_monooxygenase			1	1	1	1	1	2		
4-hydroxybenzoate_decarboxylase_subunit_C			1	1	2	1	1	2	1	
6-phosphogluconate_dehydratase_Phosphogluconate_dehydratase_protein			1							
Acryloyl-CoA_reductase_(NADH)				1	1			1		
Acyl-CoA/acyl-ACP_dehydrogenase				1						
Acyl-CoA_dehydrogenase	1		4	5	7	2	4	6	3	1
Acyl-CoA_dehydrogenase,_short-chain_specific			1							
Acyl-CoA_dehydrogenase_3				1						
Acyl-CoA_dehydrogenase_AcdA			1							
Acyl-CoA_dehydrogenase_domain-containing_protein									1	
Acyl-CoA_dehydrogenase_family_protein			1		1	2	2	2		
Acyl-CoA_dehydrogenase_related_to_the_alkylation_response_protein_AidB			1			1	2	1		
Acyl-CoA-dh-2_domain-containing_protein					1					
Acyl-coenzyme_A_dehydrogenase									1	
Alkane_1-monooxygenase			1	2	1	2	3	2		
Alkanesulfonate_monooxygenase								2	1	
Alkylation_response_protein_AidB-like_acyl-CoA_dehydrogenase					1					
Anthranilate_1,2-dioxygenase_large_subunit				1	1					

Anthranilate_1,2-dioxygenase_large_subunit/terephthalate_1,2-dioxygenase_oxygenase_component_alpha_subunit	1						
Aromatic_ring_hydroxylation_dioxygenase_C	1	1					
Aromatic_ring-hydroxylating_dioxygenase_subunit_alpha	2		1	1			
Aromatic-ring-hydroxylating_dioxygenase			1				
Aromatic-ring-hydroxylating_dioxygenase_subunit_beta	1		1				
Assimilatory_nitrate_reductase_catalytic_subunit		1	2		1	1	2
Assimilatory_nitrate_reductase_large_subunit						1	
Bac-luciferase_domain-containing_protein	2	2	1			1	
Baeyer-Villiger_monooxygenase			1				
Benzene_1,2-dioxygenase	1	1					
Benzene_1,2-dioxygenase_subunit_alpha	1						
Benzene_1,2-dioxygenase_subunit_beta	1						
Benzylsuccinate_synthase_alpha_subunit		1					
Biotin_biosynthesis_cytochrome_P450	1	1				1	
Biotin_sulfoxide_reductase	1						
Biphenyl_2,3-dioxygenase_subunit_alpha						1	1
Biphenyl_dioxygenase_subunit_beta	2	1				1	1
Choline_monooxygenase		1					
Cyclohexanone_1,2-monooxygenase			1				
Cyclohexanone_monooxygenase	2	2	1			1	
Cyclohexanone_monooxygenase/acetone_monooxygenase			1			1	
Cyclopentanone_1,2-monooxygenase	1		2	1	1	1	
Cytochrome_P450	3	3	2	1	1	2	
D504	1	1					
D513	1	1					
D516		1					
Dibenzothiophene_monooxygenase	1		1				
Dioxygenase		1					
DszA					1		
DszC					1		
F420-dependent_glucose-6-phosphate_dehydrogenase	1						
FAD-containing_monooxygenase_EthA	3	2	3	2	3	2	
FA-desaturase_domain-containing_protein		1		1	1		
FdhF/YdeP_family_oxidoreductase					1	1	
Flavin-containing_monooxygenase_FMO			1				
formate_dehydrogenase	1	1	1				
Formate_dehydrogenase,_alpha_subunit_(FdhA1)	1	2	2		1		2
Formate_dehydrogenase,_nitrate-inducible,_major_subunit	1						
Formate_dehydrogenase_H	1				1		
formate_dehydrogenase_subunit_alpha	2	2	2	1	2	2	
Glutaryl-CoA_dehydrogenase		1	1		1		

hypothetical_protein		5	2	4	2	4	5	2	
Isovaleryl-CoA_dehydrogenase		2			1		1	1	
Linalool_8-monooxygenase		1			1				
LLM_class_F420-dependent_oxidoreductase		1							
LLM_class_flavin-dependent_oxidoreductase				1			1		
Long-chain_alkane_monooxygenase							1		
L-prolyl-[peptidyl-carrier_protein]_dehydrogenase									1
Menaquinone_biosynthesis_decarboxylase	1			1		1			
Methane_monooxygenase/ammonia_monooxygenase_subunit_A						1			
Methane_monooxygenase/ammonia_monooxygenase_subunit_B				1		2	1		
Methane_monooxygenase/ammonia_monooxygenase_subunit_C				1		1	1		
Molybdopterin_oxidoreductase			1						
Molybdopterin_oxidoreductase,_Psr/Psh_family,_PsrA-like_catalytic_subunit									1
Molybdopterin_oxidoreductase_family_protein		1							
Molybdopterin-containing_oxidoreductase_catalytic_subunit			1						
Molybdopterin-dependent_oxidoreductase	1	2		1	1	2	2		
Monooxygenase		1	2	1		1	2	1	
NAD(P)/FAD-dependent_oxidoreductase		3	2	1			1		
NADH-quinone_oxidoreductase_subunit_G				1	1		1		
Neopentalenolactone_D_synthase			1				1		
Nitrate_reductase	1	5	6	2	2	3	3	4	
nitrate_reductase_(quinone)		1		1		2			
Nitrate_reductase_alpha_chain			1				4	1	
nitrate_reductase_catalytic_subunit_NapA		1							
Nitrilotriacetate_monooxygenase			1						
Nitrilotriacetate_monooxygenase_component_A_(NTA_monooxygenase_component_A)_(NTA-MO_A)		1					1		
Ortho-halobenzoate_1,2-dioxygenase_alpha-ISP_protein_OhbB				1					
Oxidoreductase_alpha_(Molybdopterin)_subunit				1		1			
Particulate_methane_monooxygenase_A-subunit				1			1		
p-cumate_2,3-dioxygenase_system,_large_oxygenase_component		1					1		
p-cumate_2,3-dioxygenase_system,_small_oxygenase_component							1		
p-cumate_dioxygenase		1	2						
Perchlorate_reductase_subunit_alpha				1					
Periplasmic_nitrate_reductase				1					
Phenazine_N-monooxygenase_PhzNO1				1			1		

Phenol_2-monooxygenase,_oxygenase_component_DmpN			1				
Phenol_hydroxylase			1				
Phenoxybenzoate_dioxygenase_subunit_alpha	1						
Phenylacetone_monooxygenase	1	1	1				
Phenylpropionate_dioxygenase_or_related_ring-hydroxylating_dioxygenase,_large_terminal_subunit	1		1	1	1	2	
propane_2-monooxygenase						1	
Putative_ABC-type_transport_system_involved_in_lysophospholipase_L1_biosynthesis		1					
Putative_dimethyl_sulfoxide_reductase_chain_YnfE				1	2	1	
Putative_dimethyl_sulfoxide_reductase_chain_YnfF	1		2				
Putative_flavoprotein_involved_in_K+_transport	1	1					
Putative_Nitrate_reductase						1	
Putative_oxidoreductase			1				
Putative_oxidoreductase_YoaE			1		1		
putative_vanillate_O-demethylase_oxygenase_subunit_oxidoreductase_protein	1						
Pyrimidine_monooxygenase_RutA			1			1	
Pyrogallol_hydroxytransferase_large_subunit							1
Respiratory_nitrate_reductase_2_alpha_chain	1						
Respiratory_nitrate_reductase_alpha_chain		1	1				
Respiratory_nitrate_reductase_subunit_alpha						1	
Rieske_2Fe-2S_domain-containing_protein						2	
Rieske_domain-containing_protein	4	4	5	2	3	3	
Ring-hydroxyl-A_domain-containing_protein	1						
Ring-hydroxylating_dioxygenase_subunit_beta	2	2					
Ring-hydroxylating_oxygenase_subunit_alpha	1						
Salicylate_5-hydroxylase,_large_oxygenase_componen		1					
SidA/IucD/PvdA_family_monooxygenase							1
Tert-butanol_monooxygenase_/_tert-amyl_alcohol_desaturase_oxygenase_subunit			1				
TIGR03619_family_F420-dependent_LLM_class_oxidoreductase	1						
Tnp-DNA-bind_domain-containing_protein			1				
Toluene-4-monooxygenase_system,_hydroxylase_component_subunit_alpha			1				
Toluene-4-monooxygenase_system,_hydroxylase_component_subunit_beta			1				
Toluene-4-monooxygenase_system_protein_B			1				
Toluene-4-sulfonate_monooxygenase_system_iron-sulfur_subunit_TsaM1	1		2	2	2	3	

Trimethylamine-N-oxide_reductase_(Cytochrome_c)	1	2	1
UbiD_family_decarboxylase	2	2	
UbiD2:_3-octaprenyl-4-hydroxybenzoate_carboxy-lyase			1
Vanillate_O-demethylase_monooxygenase_subunit			2


KEGG Pathway
Vancomycin resistance
Two-component system
Thyroid hormone synthesis
Thermogenesis
Teichoic acid biosynthesis
Sulfur relay system
Sulfur metabolism
Spinocerebellar ataxia
Sphingolipid signaling pathway
Selenocompound metabolism
RNA polymerase
RNA degradation
Ribosome biogenesis in eukaryotes
Ribosome
Retrograde endocannabinoid signaling
Retinol metabolism
Renin-angiotensin system
Renal cell carcinoma
Quorum sensing
Proximal tubule bicarbonate reclamation
Protein export
Proteasome
Primary immunodeficiency
PPAR signaling pathway
Porphyrin metabolism
Plant-pathogen interaction
Phosphotransferase system (PTS)
Phosphatidylinositol signaling system
Peroxisome
Pathways of neurodegeneration - multiple diseases
Nucleotide metabolism
Nucleotide excision repair
Non-homologous end-joining
Nitrogen metabolism
Mismatch repair
Mineral absorption
MicroRNAs in cancer
Microbial metabolism in diverse environments
Metabolic pathways
Lysosome
Lipopolysaccharide biosynthesis
Lipid and atherosclerosis
Legionellosis
Insulin signaling pathway
Homologous recombination
HIF-1 signaling pathway
Glucagon signaling pathway
GABAergic synapse
Fluid shear stress and atherosclerosis
Flagellar assembly
Fatty acid metabolism
Exopolysaccharide biosynthesis
DNA replication
Diabetic cardiomyopathy
Degradation of aromatic compounds
D-Amino acid metabolism
Central carbon metabolism in cancer
Cell cycle - Caulobacter
Cationic antimicrobial peptide (CAMP) resistance
Carbon metabolism
Caprolactam degradation
Biosynthesis of secondary metabolites
Biosynthesis of nucleotide sugars
Biosynthesis of cofactors
Biosynthesis of amino acids
Biofilm formation - Vibrio cholerae
Biofilm formation - Pseudomonas aeruginosa
Biofilm formation - Escherichia coli
Bile secretion
beta-Lactam resistance
Base excision repair
Bacterial secretion system
Bacterial chemotaxis
Autophagy - yeast
Antifolate resistance
Amoebiasis
Aminoacyl-tRNA biosynthesis
Alcoholic liver disease
African trypanosomiasis
ABC transporters
2-Oxocarboxylic acid metabolism
GAC0
GAC1
GAC2
GAC3
GAC4
GAC5
GAC6
GAC7
GAC8
GAC9
Sample
Complete (%)
0.04
0.03
0.02
0.01

Appendix C: KEGG Pathways retrieved from Bakta annotations of MAGs collected in each sample. Completeness of a pathway is determined by the number of KOs in a sample divided by the total number of KOs in a pathway.

Appendix D: An all-versus-all BLAST was performed with ACCs larger than 1 mb. ACCs that aligned to each other with 99% sequence identity and a minimum alignment length of 10000 bases were grouped. Each ACC was taxonomically classified by GTDBTK, and any ACC that had at least one hydrocarbon degradation (HD) gene according to CANT-HYD is indicated. ACCs that are underlined are complete (>90%) and uncontaminated (<5%) MAGs that are missing one or more ubiquitous bacterial rRNA or tRNA genes that are required to be classified as high-quality. No incomplete ACC larger than 1 mb had any hits.

ACC	Sample	HD Gene	Full Taxonomy (GTDBTK)
10	GAC0	Y	d__Bacteria;p__Pseudomonadota;c__Gammaproteobacteria;o__Ga0077554;f__Ga007554;g__LNEJ01;s__
bin.525_1	GAC2	Y	
bin.229_2	GAC5	Y	
bin.442_1	GAC7	Y	
1115	GAC2	Y	d__Bacteria;p__Pseudomonadota;c__Gammaproteobacteria;o__Ga0077554;f__Ga007554;g__SBBG01;s__
632	GAC3	N	
2304	GAC8	Y	
466	GAC1	Y	d__Bacteria;p__Pseudomonadota;c__Gammaproteobacteria;o__CALZJG01;f__CALZJG01;g__CAKKSB01;s__
5322	GAC2	Y	
13374	GAC3	Y	
1636	GAC5	Y	
bin.113_1	GAC6	Y	
2932	GAC6	Y	
7443	GAC7	Y	
6081	GAC8	Y	
bin.526_1	GAC7	Y	d__Bacteria;p__Pseudomonadota;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__Sphingobium;s__
1413	GAC8	Y	
bin.1128_3	GAC8	N	
bin.666_4	GAC4	Y	d__Bacteria;p__Acidobacteriota;c__Blastocatellia;o__RBC074;f__RBC074;g__JAJVID01;s__JAJVID01 sp022072205
1770	GAC6	Y	
3168	GAC5	Y	d__Bacteria;p__Pseudomonadota;c__Gammaproteobacteria;o__Burkholderiales;f__Rhodocyclaceae;g__Accumulibacter;s__
2268	GAC6	Y	
bin.472_3	GAC6	Y	
bin.1152_1	GAC7	Y	
309	GAC6	Y	d__Bacteria;p__CLD3;c__CLD3;o__SB21;f__SB21;g__JABWBZ01;s__
bin.732_1	GAC7	Y	
595	GAC2	Y	d__Bacteria;p__Pseudomonadota;c__Gammaproteobacteria;o__Ga0077554;f__Ga007554;g__SBB
476	GAC3	Y	

419	GAC4	Y	G01;s__
bin.778_1	GAC4	Y	d__Bacteria;p__Nitrospirota;c__Nitrospiria;o__S
4282	GAC0	Y	BBL01;f__Manganitrophaceae;g__Manganitroph
			us;s__Manganitrophus morganii
bin.252_9	GAC5	Y	d__Bacteria;p__Pseudomonadota;c__Alphaproteo
bin.867_1	GAC6	Y	bacteria;o__UBA9219;f__UBA9219;g__;s__
bin.405_1	GAC2	Y	
bin.321_1	GAC4	Y	d__Bacteria;p__Pseudomonadota;c__Alphaproteo
bin.596_1	GAC5	Y	bacteria;o__Micropepsales;f__Micropepsaceae;g_
bin.616_1	GAC6	Y	_JACADY01;s__
bin.756_6	GAC7	Y	
bin.666_1	GAC2	Y	d__Bacteria;p__Pseudomonadota;c__Alphaproteo
bin.747_1	GAC3	Y	bacteria;o__UBA11222;f__UBA11222;g__UBA1
			1222;s__

Curriculum Vitae

Name: **Henry Say**

Post-secondary Education and Degrees:

University of Western Ontario
London, Ontario, Canada
2017-2021 BMSc.
Honors Specialization in Biochemistry and Pathology

The University of Western Ontario
London, Ontario, Canada
2021-2023 MSc.
Department of Biochemistry
Supervisor: Dr. Gregory Gloor

Honours and Awards:

Mitacs Accelerate Fellowship
2022-2023

Western Graduate Research Scholarship
2021-2023

Dean's Honor List
2017-2018, 2018-2019, 2020-2021

The Western Scholarship of Excellence
2017

Related Work Experience

Teaching Assistant (BIOCHEM 2280A)
The University of Western Ontario
2021

Teaching Assistant (MBI4850G)
The University of Western Ontario
2022

Preprints:

Say, Henry, Joris, Ben, Giguere, Daniel, Gloor, Gregory B. (2023) Annotating Metagenomically Assembled Bacteriophage from a Unique Ecological System using Protein Structure Prediction and Structure Homology Search. biorXiv, https://doi.org/10.1101/2023.04.19.537516 (April 23, 2023)

Publications:

Slattery, S. S., Giguere, D. J., Stuckless, E. E., Shrestha, A., Briere, L.-A. K., Galbraith, A., Reaume, S., Boyko, X., **Say, H. H.**, Browne, T. S., Frederick, M. I., Lant, J. T., Heinemann, I. U., O'Donoghue, P., Dsouza, L., Martin, S., Howard, P., Jedeszko, C., Ali, K., … Edgell, D. R. (2022). Phosphate-regulated expression of the SARS-CoV-2

receptor-binding domain in the diatom Phaeodactylum tricornutum for pandemic diagnostics. Scientific Reports, 12(1), 7010. https://doi.org/10.1038/s41598-022-11053-7

Cochrane, R. R., Shrestha, A., Severo de Almeida, M. M., Agyare-Tabbi, M., Brumwell, S. L., Hamadache, S., Meaney, J. S., Nucifora, D. P., **Say, H. H**., Sharma, J., Soltysiak, M. P. M., Tong, C., Van Belois, K., Walker, E. J. L., Lachance, M.-A., Gloor, G. B., Edgell, D. R., Shapiro, R. S., & Karas, B. J. (2022). Superior Conjugative Plasmids Delivered by Bacteria to Diverse Fungi. BioDesign Research, 2022. https://doi.org/10.34133/2022/9802168