8-22-2022 10:00 AM

Towards more complete metagenomic analyses through circularized genomes and conjugative elements

Benjamin R. Joris, *The University of Western Ontario*

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

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Biochemistry

Recommended Citation

Joris, Benjamin R., "Towards more complete metagenomic analyses through circularized genomes and conjugative elements" (2022). *Electronic Thesis and Dissertation Repository*. 8855.
https://ir.lib.uwo.ca/etd/8855

Abstract

Advancements in sequencing technologies have revolutionized biological sciences and led to the emergence of a number of fields of research. One such field of research is metagenomics, which is the study of the genomic content of complex communities of bacteria. The goal of this thesis was to contribute computational methodology that can maximize the data generated in these studies and to apply these protocols to human and environmental metagenomic samples.

Standard metagenomic analyses include a step for binning of assembled contigs, which has previously been shown to exclude mobile genetic elements. I demonstrated that this phenomenon extends to all conjugative elements, which are a subset of mobile genetic elements. I proposed two separate methodologies that could detect contigs that are potential conjugative elements: a curated set of profile hidden Markov models that are very efficient to run, or annotation using the full UniRef90 database, a slower but more sensitive method.

I then applied this framework to a large population-based cohort and to a study examining the association of the maternal human gut microbiota and the development of spina bifida. Broadly, the composition and abundances of conjugative elements were discriminatory between the age and geographic cohorts. In the spina bifida cohort, there was an enrichment of *Campylobacter hominis* and a conjugative element belonging to *Campylobacter hominis*, which was excluded from the metagenomic bins.

Next, I characterized a novel species belonging to the recently discovered manganese-oxidizing genus *Manganitrophus* growing on oil refinery carbon filters. I successfully circularized the genomes of three strains and got quality assemblies for the remaining two samples. Furthermore, I identified a previously uncharacterized conjugative plasmid belonging to the species using my framework developed in chapter 2.

Finally, I developed an assembly pipeline to perform a secondary assembly on binned assemblies using long reads. The secondary assemblies yielded a number of additional circularized sequences that would be useful as scaffolds in future metatranscriptomic, variation analysis, and community dynamic studies.

The methodologies and applications in this thesis provide a framework for more complete metagenomic analyses going forward that will aid in our understanding of microbial ecology.

Keywords: Bioinformatics, computational biology, microbiome, metagenomics

Lay Summary

Over recent years, the technology to determine the DNA sequence of species' genomes has advanced greatly. These technological advances have allowed for the study of bacterial genomes living in complex communities without the need to isolate them individually, a field referred to as metagenomics. As a rapidly expanding field of research, metagenomic analyses required computational tools that can accurately analyze the massive quantities of data being produced. For my thesis, I sought to develop such tools and apply them to the complex bacterial communities that colonize the human intestinal tract and to communities that grow on carbon filters from the wastewater treatment facility of an oil refinery.

Conjugative elements are pieces of DNA that can be exchanged between bacteria. These mobile genetic elements are of clinical interest because they commonly carry cargo genes that can confer antibiotic resistance to the bacteria. I demonstrate that standard metagenomic analyses systematically exclude these elements, and I proposed a methodology to remedy this issue.

I then applied this methodology for identifying conjugative elements to two separate research questions. First, I showed that conjugative systems are different depending on the age and geographical location of the individual, likely due to antibiotic use and diet differences between populations. Additionally, I showed that harmful bacteria carrying a conjugative element in the guts of expectant mothers may play a role in the development of spina bifida.

Growing on the carbon filters of an oil refinery's wastewater treatment plant, I discovered a novel species of bacteria whose closest known relatives on the tree of life are able to use manganese as a source of energy. I also found a previously uncharacterized conjugative element belonging to this species that is likely able to remove heavy metals from its environment.

Finally, I developed a method for assembling additional complete bacterial genomes from the sequencing of complex environments. Additional complete genomes will enable a better ability to understand the full genetic potential of these bacterial communities from the wastewater treatment plant.

Overall, I improved computational methods for analyzing complex bacterial communities and ap-

plied the methods as a proof of principle for their usefulness.

Coauthorship Statement

The research presented in this thesis was conducted by Benjamin Joris with the exceptions noted below. Benjamin Joris conceived, designed, and executed all of the experiments and analyzed the data. With input from Gregory Gloor, Benjamin Joris wrote all manuscripts and thesis chapters.

Chapter 2 - Tyler Browne implemented the conjugative protein pHMM models in Anvi'o and aided in the construction of some of the python scripts used to select contigs.

Chapter 3 - The DNA from the spina bifida microbiome samples were isolated by Hauna Sheyholislami from Dr. Kristin Connor's lab at Carelton University.

Chapter 4 - The DNA isolation and sequencing was performed by Daniel Giguere and Henry Say.

Chapter 5 - The DNA isolation and sequencing was performed by Daniel Giguere, Alec Bahcheli, and Henry Say.

Acknowledgements

First, I would like to thank Dr. Greg Gloor for his unwavering support throughout the course of my thesis. You took me on as a graduate student who had minimal background in bioinformatics and helped sculpt me into the researcher I am today. You helped foster the independence that a doctoral student needs, especially when faced with a global pandemic that cut off in-person research for the bulk of my thesis. You always welcomed newly proposed research ideas and provided your wisdom on how to build them into full research projects. The mentorship you provided transformed me as both a student and an individual.

I would also like to thank my advisors Drs. David Edgell and Robert Hegele for supporting and encouraging me in my proposed research, which was refined into what it became in large part to your feedback. Thank you Dr. Edgell for continually challenging me to better my writing abilities. Without your constructive feedback on drafts of my early manuscripts, I would not have progressed nearly as much as a writer.

To all my fellow students of the biochemistry department at Western University, past and present, I would like to thank you for making graduate school an unforgettable experience. From intramural softball to the annual holiday party, there were so many memories that brought life to graduate school. In particular, I would like to thank all the fellow 'Molecular Biological Laboratory' colleagues. The Gloor lab was always a melting pot for labs in the building to come down and chat about science and life.

A special thanks to all the fellow members of the Gloor lab over the years. Thank you to Dr. Jean Macklaim for getting me started in the lab. Going into my time in the Gloor lab, I had no clue how to use R, Bash, or any other scripting language, and you showed me far more patience than I deserved. To Dr. Daniel Giguere, thank you for paving the road that I walked down. Navigating a PhD program can be pretty opaque, but I could always just ask what you had done one year prior. You also set the stage for so many of my projects and were always willing to give your opinion on any idea I had. Another special thank you to Henry Say who provided the technical expertise of DNA isolation and sequencing that made my final couple of chapters possible in the time frame of my thesis.

I would like to thank my parents for their continual support throughout my many years of post-secondary education. Unfortunately, you will no longer be able to make jokes like ‘Ben is starting grade 21 this fall’, but I am sure you will find new things to tease me about. Finally, I would like to thank Jasmine for being with me every step of the way. Graduate school is full of days where everything goes wrong, but you were always there to make things right.

Contents

List of Figures

List of Tables

List of Appendices

List of Abbreviations, Symbols, and Nomenclature

ASV	amplicon sequence variant
ANI	average nucleotide identity
AT%	percentage of bases that are adenine and thymine
BLAST	basic local alignment search tool
bp	base pair
CAT/BAT	contig annotation tool / bin annotation tool
cDNA	complementary DNA
CE	conjugative element
COG	clusters of orthologous genes
contig	contiguous DNA sequence
CPU	central processing unit
CRISPR	clustered regularly interspaced short palindromic repeats
DNA	deoxyribonucleic acid
dNTP	deoxynucleotide triphosphates
FASTA	file format for genomic and protein sequences
FASTQ	file format for sequencing data
FMT	faecal microbiota transplant
GAC	granular-activated charcoal
GC%	percentage of bases that are guanine and cytosine
GO	gene ontology
GPU	graphical processing unit
HMM	hidden Markov model
IBD	inflammatory bowel disease
ICE	integrative and conjugative element

indel	insertion or deletion
KEGG	Kyoto encyclopedia of genes and genomes
kb	kilo base pair
kmer	substrings of length k contained within a biological sequence
LCA	last common ancestor
MAG	metagenome-assembled genome
Mb	mega base pair
N50 (sequencing)	smallest contig length of which 50% of total bases of the assembly are found
N50 (sequencing)	smallest sequencing read length of which 50% of total bases are found
OLC	overlap-layout-consensus
ORF	open reading frame
PAG	polyaromatic hydrocarbon
PCR	polymerase chain reaction
pHMM	profile hidden Markov model
RNA	ribonucleic acid
rRNA	ribosomal ribonucleic acid
SMS	single-molecule sequencing
T4SS	type IV secretion system
T4CP	type IV coupling protein
tRNA	transfer ribonucleic acid

Chapter 1

General Introduction

Bacterial genomics is a rapidly advancing field centered around the interest in the metabolic potential of bacteria to modulate human health and remediate the environment. Sequencing technology and the data so generated that enables bacterial genomic research has been outpacing Moore's law [1]. As such, the computational methodology to accurately analyze these increasingly large datasets must keep pace. While there has been much progress made in this regard, there is still much data that is overlooked or missed by standard analyses that may be of clinical or environmental interests.

The overarching goal of my thesis was to improve standard metagenomic analyses by employing and adapting novel technologies and then apply these methodological improvements to problems as diverse as human gut microbiome and environmental metagenomic investigations. In Chapter 2, I discovered that conjugative elements from bacteria are systematically excluded from metagenomic bins and developed a pipeline to identify these systems from metagenomic assemblies. I then applied this methodology to gut microbiome datasets to identify conjugative elements in geographic-focused cohorts (cohorts where the primary difference between group is geographic location of sample location) and in mothers who gave birth to infants with spina bifida (Chapter 3). I also applied the identification of conjugative systems to environmental samples to identify a large, circularized conjugative plasmid belonging to a species in the manganese-oxidizing genus

Manganitrophus. In addition, I showed that plasmids and chromosomes could be assembled into circular contigs using a reference-based assembly with long-reads and compared the gene content of these novel strains to the three known species of the genus (Chapter 4). Extending the idea that mapped assemblies using high-quality genomes could yield additional circularized bacterial genomes, I performed secondary assemblies using reads that strongly align to metagenomic bins to better characterize the communities of bacteria growing on the carbon filters found at the wastewater treatment plants of oil refineries (Chapter 5).

1.1 History of sequencing technologies

Ever since the resolution of the three-dimensional double helix structure of DNA by Franklin, Wilkins, Crick, and Watson [2], and the proposition of the central dogma of biology [3], research has been conducted on how to obtain the DNA sequences of organisms. The central dogma of biology led to the belief that the knowledge of the DNA sequence would allow for a comprehensive understanding of phenotypes and diseases. The nature of DNA, with long sequences and very little structural differences between bases, made knowledge of protein sequencing inapplicable [4], so new strategies needed to be developed to rapidly determine nucleic acid sequences. Three generations of DNA sequencing technologies have been developed (Figure 1.1), each generation with its own strengths and weaknesses, which has brought about an ability to complete genomes from complex communities.

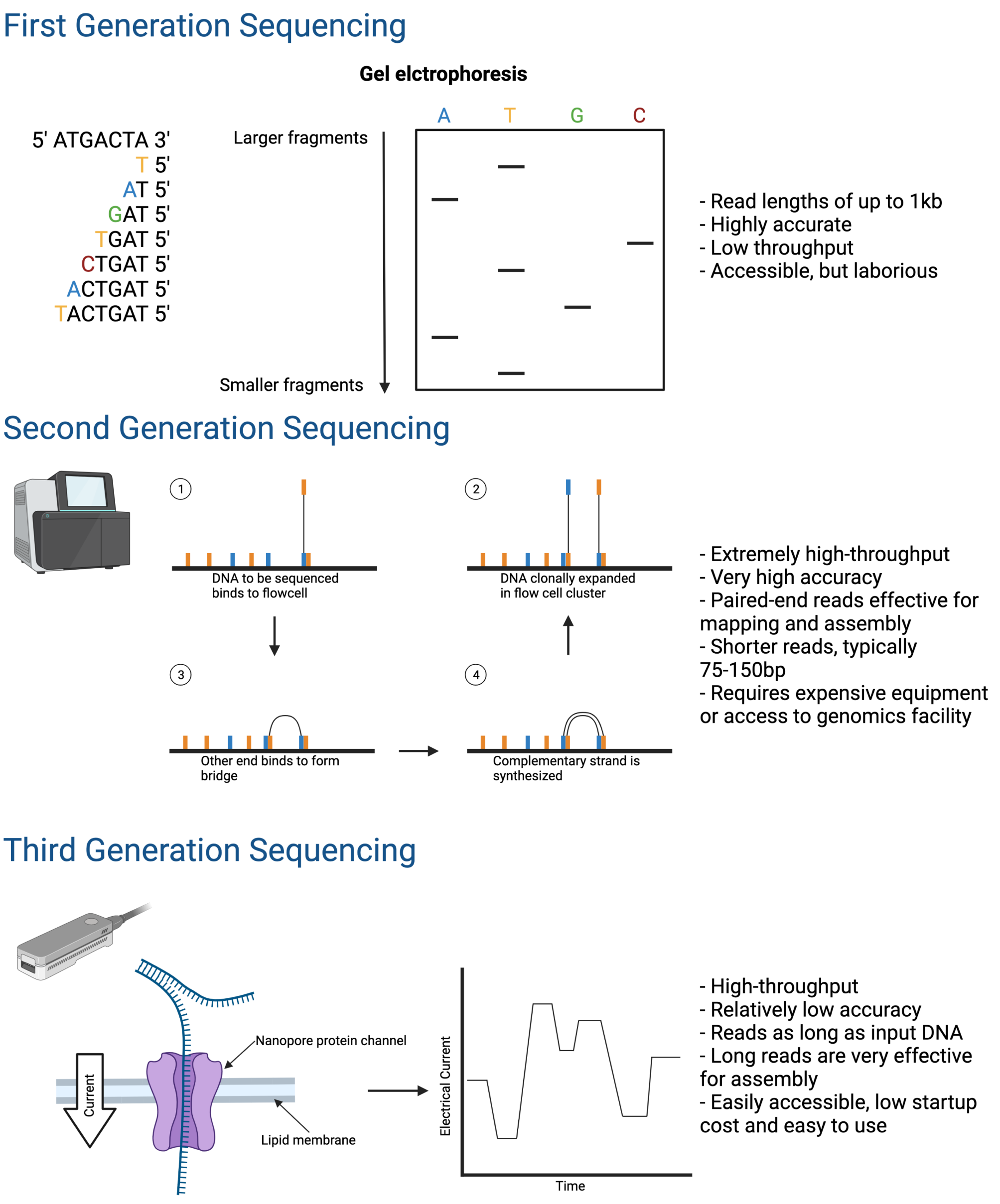

First Generation Sequencing
Gel elctrophoresis
A
T
G
C
5' ATGACTA 3'
T 5'
AT 5'
GAT 5'
TGAT 5'
CTGAT 5'
ACTGAT 5'
TACTGAT 5'
Larger fragments
Smaller fragments
- Read lengths of up to 1kb
- Highly accurate
- Low throughput
- Accessible, but laborious
Second Generation Sequencing
1
DNA to be sequenced binds to flowcell
2
DNA clonally expanded in flow cell cluster
3
Other end binds to form bridge
4
Complementary strand is synthesized
- Extremely high-throughput
- Very high accuracy
- Paired-end reads effective for mapping and assembly
- Shorter reads, typically 75-150bp
- Requires expensive equipment or access to genomics facility
Third Generation Sequencing
Current
Nanopore protein channel
Lipid membrane
Electrical Current
Time
- High-throughput
- Relatively low accuracy
- Reads as long as input DNA
- Long reads are very effective for assembly
- Easily accessible, low startup cost and easy to use

Figure 1.1: The three generations of sequencing technology with key advancements of each generation highlighted. Sanger sequencing in the first generation of sequencing saw the use of chain termination with dideoxynucleotide triphosphates and the use of one-dimensional gel electrophoresis in place of two-dimensional fractionation. In the second generation of sequencing, Solexa/Illumina introduced bridge amplification of DNA spots to create a clonal cluster to greatly increase sequencing accuracy. Oxford Nanopore Technologies introduced the nanopore method of sequencing, which directly sequences the DNA as it transverses the membrane and disrupts the flow of ions. Created with BioRender.

1.1.1 First generation sequencing: artisinal

Sequencing of nucleic acids was first conducted with RNA rather than DNA due to the smaller sequences, RNase enzymes for site-specific cleavage, and established isolation protocols. At first, only the full composition of the sequence, but not the order of the nucleotide could be determined [5]. Eventually, following the realization that RNA used uracil in place of thymine [6], Holley and colleagues were able to establish a method of sequencing capable of sequencing a yeast tRNA [7]. The same year, Sanger developed a two-dimensional fractionation technique of radiolabeled nucleotides that enabled the sequencing of a wider pool of tRNA and rRNA sequences [8, 9, 10]. As steady improvements were made to the two-dimensional fractionation technique over the next decade, the first protein-coding sequence [11] and first complete viral genome were sequenced [12].

Improvements to sequencing again needed to be made to obtain DNA sequences. It was found that the overhanging ends of the lambda phage genome could be filled in using a DNA polymerase supplied with radiolabeled nucleotides. By supplying these nucleotides one at a time, the sequence of the overhanging ends could be determined [13, 14]. This method was further developed and it was discovered that the radiolabeled nucleotides could be inserted anywhere [15, 16, 17], but was still limited to small stretches of DNA.

The next improvements came in the development of two similar techniques that would become the standard practice for the next period of time. Sanger sequencing [18], also referred to as chain-termination sequencing, and Maxam-Gilbert sequencing [19], or chemical cleavage sequencing, both improved on the existing techniques by greatly simplifying the entire process. The change from two-dimensional fractionation to a four lane, size-based gel electrophoresis approach made DNA sequencing much more accessible. The first DNA genome sequenced as a result of this technological advancement was the ΦX174 bacteriophage genome [20], which is still used to this day as a positive control in many sequencing runs. A subsequent improvement to Sanger sequencing was the use of fluorescence in place of phospho-radiolabeling, which allowed for ‘one-pot’ sequencing, which allowed for automation of the process [21, 22, 23, 24, 25, 26, 27].

At the end of the first-generation sequencing era, there were still technological advancements being made that improved the data. The development of polymerase chain reaction [28, 29] allowed for the generation of large quantities of pure DNA for sequencing. These improvements coupled with the incorporation of Sanger sequencing into automated sequencing machines like the Applied Biosystems ABI PRISM allowed for the crowning achievement of the first-generation sequencing era, the first draft human genome [30, 31]. While impressive, sequencing the human genome required a tremendous amount of time and resources to complete as the low throughput of first generation sequencing technology was a major limiting factor.

1.1.2 Second generation sequencing: highly-parallel

Second generation sequencing was spurred by the discovery of the ability to measure the release of pyrophosphate measured by luminescence through the enzyme luciferase [32], also known as pyrosequencing. By washing a nucleotide solution over a bed of template DNA fixed to a solid phase, it was possible to determine which nucleotides are being added to each strand in parallel [33]. This method allowed the synthesis to occur with natural nucleotides and be measured in real-time [34, 35, 36]. A major drawback to the method compared to chain termination was the inability to easily resolve homopolymers, mono-nucleotide stretches in the sequence. Chain termination sequencing is capable of resolving homopolymers, though still imperfectly, because each additional base in the homopolymer increases the length on the gel and can be measured. However, with pyrosequencing, the bases are added in real time and the bases will be added subsequently during the same wash in a homopolymer. Illumination intensity measurement can help to determine up to strings of five consecutive bases, but it cannot be resolved beyond that. These technologies were developed into sequencing machines by 454, which allowed for large libraries and high-throughput. This advancement permitted the sequencing and assembly of a diploid human genome in far less time than the original Human Genome Project [37].

After 454, Solexa (Illumina) brought about the next advancement with their version of sequencing by synthesis, which is, in essence, massively parallel Sanger sequencing. Illumina sequencing includes a step for bridge amplification where solid phase PCR creates a clonal population of DNA sequences, which are then sequenced with a fluorescent dNTP. One of the main benefits of Illumina sequencing is the paired-end reads that it can produce by sequencing the ends of a DNA fragment. Paired-end reads are beneficial when assembling a genome *de novo* or when you are aligning reads to a reference genome, particularly in highly repetitive regions. MGI sequencing has recently developed nanoball sequencing technology, which also produces high quantities of reads at comparable quality to Illumina sequencing [38]. The advancements in the second generation of sequencing technology is primarily where it greatly outpaced Moore's Law [1]. This generation of sequencing saw the draft assembly of the human genome go from a triumphant accomplishment to a relatively mundane exercise.

1.1.3 Third generation sequencing: single molecule

The third generation of sequencing is characterized by single-molecule sequencing (SMS). SMS originated in the early 2000s where it was demonstrated that single DNA molecules could be sequenced using fluorescence microscopy [39]. Early versions of commercialized SMS functioned similarly to Illumina's platforms, but without the bridge amplification to clonally expand the population of DNA. While this early version of SMS did not provide any advantages to the dominant Illumina sequencing platform, it did serve as a springboard for others to enter the SMS space, namely Pacific Biosciences and Oxford Nanopore. PacBio's SMRT sequencing platforms uses fluorescent dNTPs as well, but carries out the process in a nanostructure referred to as a 'zero-mode waveguide', which allows for precise measurement of the fluorophores being added to the DNA in real time. This method of sequencing is fast, relatively accurate, and can produce sequences up to 10kb in length.

Oxford Nanopore Technologies sequencing technology is drastically different than the other methods that have come before it. Rather than reading out the addition of a nucleotide to a chain by fluorescence or luminescence, nanopore sequencing directly sequences a strand of DNA or RNA. It does so by pulling a strand of nucleic acid through a hemolysin ion channel by electrophoresis, which disrupts the flow of current across a lipid bilayer [40]. The disruption of the current in the bilayer is transformed computationally into a DNA sequence through machine learning. Nanopore sequencing has a number of advantages. Most notably is the ability to produce sequences with no theoretical upper bound in length. Because the DNA or RNA is sequenced directly with no synthesis involved, whatever length of nucleic acid chain is input into the sequencer is output to be analyzed. Additionally, platforms such as the MinION or Flongle are extraordinarily portable and can be used to sequence in remote locations, such as tracking an Ebola outbreak in Africa [41]. Nanopore sequencing is, however, not without its drawbacks. Overall sequence quality is comparatively lower than its competitors (though is continually improved with better basecalling algorithms) and it also struggles at sequencing homopolymers.

Sequencing technology has advanced rapidly over the last half century. In tandem, computational tools need to continually evolve and adapt to new avenues of research that continually emerge with these new technological advancements. While it was only possible to sequence a single tRNA sequence fifty years ago, and a single draft human genome twenty years ago, with the advent of second- and third-generation sequencing technologies much more complex sequencing experiments are now possible.

1.2 Metagenomics

One of the fields of research that has been enabled by the high-throughput sequencing technologies developed in the past few decades is metagenomics. Sequencing of isolate genomes of bacteria predates the popularization of high-throughput second generation sequencing with the first bacteria genome of *Haemophilus influenzae* published in 1995 [42]. However, the overwhelming majority of bacteria are not able to be cultured in lab as isolates [43], which complicated their study with low throughput sequencing methods. Metagenomics, the study of complex communities, aims to bypass this problem by directly sequencing the community without the need for culturing of the bacteria. The first instances of metagenomics predates the wide-spread use of second-generation sequencing as well. In 1998, the term metagenomic was coined to describe the study of the collective genomes of the soil microflora [44], where it was proposed that the genomes of uncultivatable bacteria be cloned into cultivable bacteria for study. They highlighted the overwhelming diversity of bacteria in the soil and the potential for beneficial metabolic pathways, such as the synthesis of novel antibiotics, existing in the pool of unknown bacteria.

Similarly to the progression of sequencing technologies, there has been consistent improvements in the methodologies for metagenomics. Due to the limitations in sequencing, the early form of metagenomics came in the form of targeted sequencing of phylogenetically relevant regions such as the 16S rRNA gene. However, as the cost of sequencing has gone down, throughput has increased, and the emergence of third-generation sequencing technology has allowed for long-read sequencing, shotgun metagenomics of all DNA in an environment has gained prominence.

1.2.1 Amplicon Sequencing

Targeted, amplicon sequencing permits the study of the phylogeny of bacteria in a complex community while only sequencing a small portion of the total DNA present. The 16S rRNA gene is large enough to provide sufficient information, conserved throughout the bacterial kingdom, and has a level of sequence variation in line with the evolutionary phylogeny [45]. Study of bacteria using 16S rRNA gene sequences began almost half a century ago with Carl Woese and colleagues used the 16S sequence to classify methanogenic bacteria [46] in 1977 and to classify *Halobacterium volcanii* in 1983 [47]. As sequencing and PCR technology advanced, the 16S rRNA gene sequenced could be amplified and sequenced from complex and uncultivatable communities of bacteria [48, 49, 50].

The general strategy of 16S rRNA gene sequencing is to bind PCR primers to conserved regions that flank the variable regions, PCR amplify the variable region, and then sequence the region (Figure 1.2). Compositional analysis pipelines such as DADA2 are able to take the resultant raw data and pull out the clusters of similar sequences, classify the clusters taxonomically, and then quantify the relative abundances of the clusters [51].

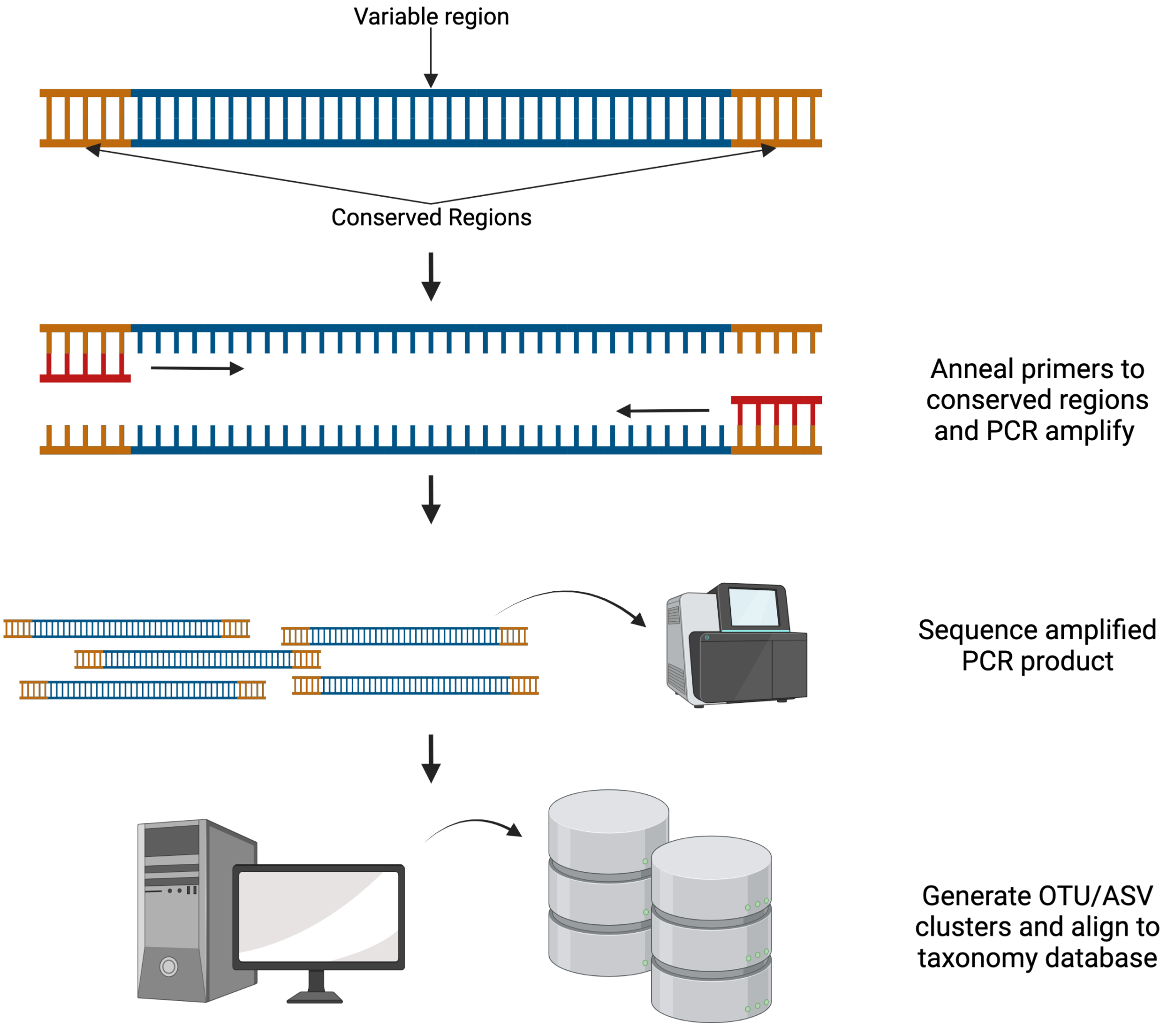

Variable region
Conserved Regions
Anneal primers to conserved regions and PCR amplify
Sequence amplified PCR product
Generate OTU/ASV clusters and align to taxonomy database

Figure 1.2: Example workflow of 16S rRNA gene amplicon sequencing. A variable region within the 16S gene is identified and primers are designed to anneal to regions flanking it that are well conserved. The PCR product is amplified and sequenced. The resultant data is passed to a program such as DADA2 that can generate clusters of sequences (referred to amplicon sequence variants, or ASVs) that can be aligned to a database to assign taxonomy. The relative abundances of the ASVs can be used to assess the taxonomic composition within a sample and differences between groups. Created with BioRender.

16S rRNA gene sequencing, however, has many drawbacks. For one, it struggles broadly to differentiate species and some genera [52, 53]. Some species can display up to 99.9% sequence identity at the 16S rRNA gene level, but have low DNA relatedness in hybridization studies and are clearly distinguishable in biochemical assays [45]. Species identification is also contingent on the quality of the database, which has proven to be problematic in the past with many deposited low-quality or mis-annotated sequences [54]. The nature of 16S rRNA gene sequencing also has some systematic blind spots. For instance, because only the 16S rRNA gene is sequenced, no information on horizontal gene transfer or metabolic capabilities can be confidently derived from the data. Bioinformatic programs such as PICRUSt attempt to predict and reconstruct biochemical pathways on the basis of 16S gene sequencing data by using nearest matches in genome databases to predict the metabolic potential of the bacteria [55]. However, due to the inability of 16S rRNA gene sequencing to resolve species or strains, these predictions are limited and largely speculative.

1.2.2 Shotgun metagenomic sequencing

Shotgun metagenomic sequencing is an extension of shotgun genomic sequencing that has been long-used to assemble some of the first DNA sequences such as the bovine mitochondrion [56]. In contrast to isolating DNA from a single organism or organelle and building a library, shotgun metagenomics builds the sequencing library from all DNA in an environment. This permits the full analysis of all the gene content in the community. In comparison to amplicon sequencing, shotgun metagenomic sequencing also allows for the analysis of strain-level differences and horizontal gene transfer in bacterial communities. The overall pipeline of shotgun metagenomic analysis bears many similarities to 16S rRNA gene sequencing, but also includes steps for genome assembly, binning, and gene prediction that are not possible with amplicon sequencing (Figure 1.3).

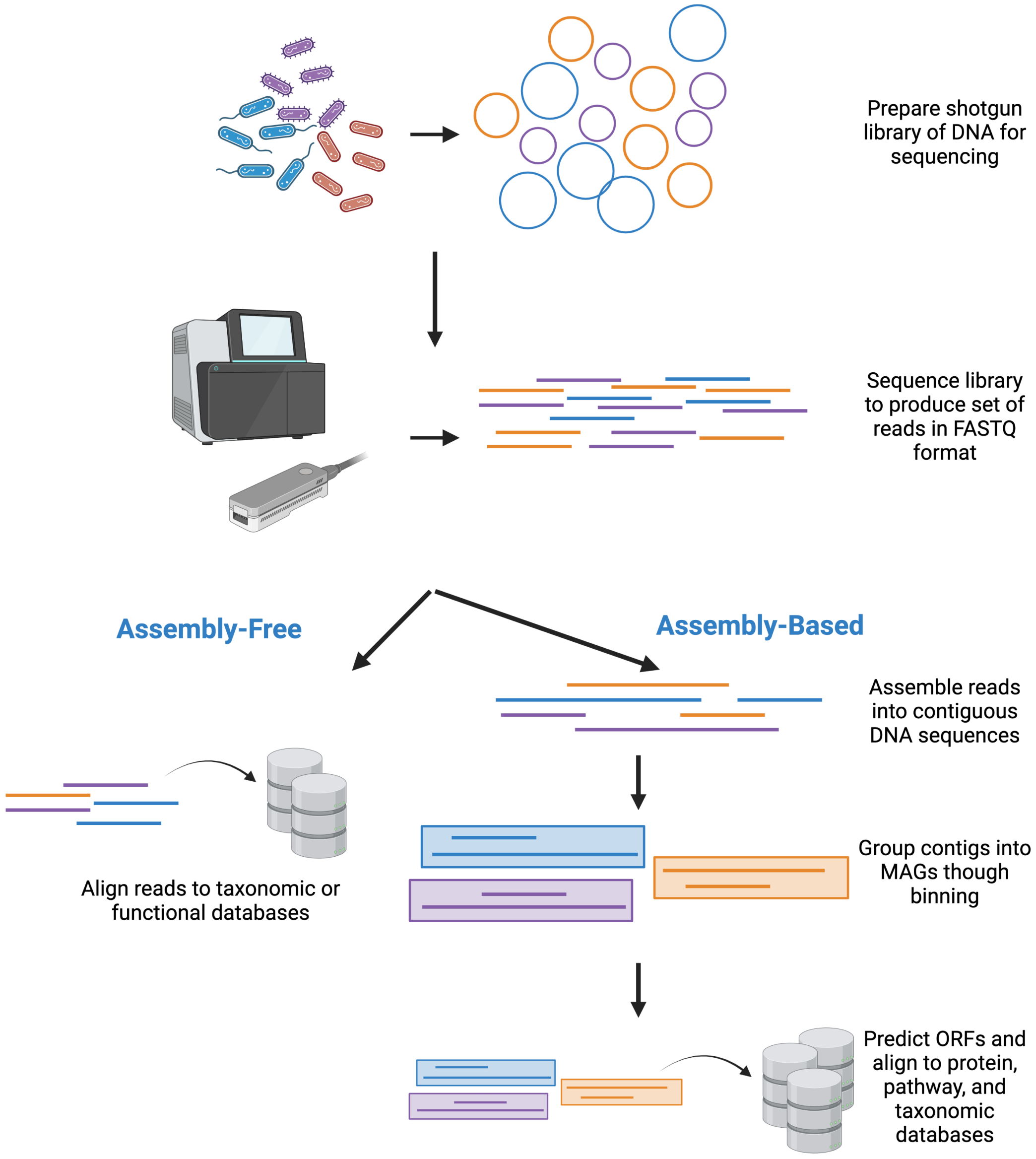

Prepare shotgun library of DNA for sequencing
Sequence library to produce set of reads in FASTQ format
Assembly-Free
Assembly-Based
Assemble reads into contiguous DNA sequences
Align reads to taxonomic or functional databases
Group contigs into MAGs though binning
Predict ORFs and align to protein, pathway, and taxonomic databases

Figure 1.3: Example workflow of shotgun metagenomic sequences. First the DNA from the entire community is isolated and turned into a sequencing library that can be sequenced on second-or third-generation sequencing platforms. The resultant FASTQ sequencing files can be analyzed through two separate, but similar, analysis pipelines. In the assembly-free method of analyzing the shotgun metagenomic data, the sequencing reads are directly analyzed by classifying the taxonomy or gene content of each read. In the assembly-based method of metagenomic analyses, the reads are assembled into larger genomic fragments referred to as contiguous DNA sequences, or contigs. The contigs are then grouped based on features such as mapping coverage and sequence composition into approximations of bacterial genomes called metagenome-assembled genomes (MAGs). The open reading frames of the MAGs are predicted to align to protein and pathway databases and the taxonomy of the full bin is assigned. Created with BioRender.

1.2.3 Metagenomic Assembly

To make the most out of the ability of shotgun metagenomic sequencing to capture full genomic data, the sequencing reads must be stitched together into larger genetic fragments referred to as contigs though metagenomic assembly. Metagenomic assembly of a complex community, such as the human gut microbiome, is an extraordinarily difficult task to accomplish for a multitude of reasons. Bacterial genomes often share a number of highly similar DNA sequences across the genome, which can make separation and resolution of the genomes difficult. Paired-end sequencing with large inserts or long-read third generation sequencing can help to resolve these ambiguities by anchoring the reads outside the region of homology. Additionally, metagenomic assembly of a full metagenome is complicated by the uneven proportions of members of the community. Oftentimes, there are insufficient data for the least abundant members of the community to have a high quality genome produced [57, 58]. One method to overcome this issue would be to simply sequence more deeply with the hope that the sequencing depth threshold for a quality metagenome-assembled genome will be crossed for the lower abundance species in the metagenome. However, this would be expensive and inefficient because most of the additional sequencing will be consumed by the higher abundance species that are already sufficiently sequenced. A recent advancement to overcome this pitfall in metagenomic sequencing and assembly is adaptive sequencing on the Nanopore third-generation sequencers [59]. This method uses a software-controlled enrichment protocol that selectively sequences reads based on their similarity to previously assembled low-abundance MAGs, which prevents the over-sequencing of the high abundance species.

The process of assembly usually follows one of two methods, either overlap-layout-consensus (OLC) or de-Bruijn-graph [60, 61, 62]. OLC is the classical form of assembly that was popularized during the era of Sanger sequencing and assembly of simple, isolate genomes. The basic premise is that overlaps between all reads are formed. Using the overlaps, the reads are laid out in a graph and the consensus sequence is determined. De-Bruijn graph assembly is a newer technique that does not involve overlapping all reads, but rather breaking all reads down into a set of k-mers [63, 64]. The k-mers are used to conduct a Eulerian walk along the sequence of the genome. With the advent of second-generation sequencing, de-Bruijn graph assembly became the preferred method of assembly because of computational and memory limitations of OLC assembly [60].

An early metagenomic assembly algorithm was developed as an extension of the isolate de Bruijn graph assembly algorithm 'Velvet' [65, 66, 67]. MetaVelvet is one of the first dedicated assemblers for metagenomic datasets. Some of the adaptations made to the original algorithm were to use the differential coverage and graph connectivity to create subgraphs within the de Bruijn graphs. Because within a metagenome, each species is likely to have a different sequencing depth, adding in the step for separating the subgraphs based on coverage was quite effective. As a result, this metagenomic-focused assembler was able to generate much higher N50 (the size of contig where equal length and longer contigs compose half the total assembly) than with assemblers that focused on single genomes [67]. However, as the first iteration of a de Bruijn graph short-read assembler, there was still much room for improvement.

Another widely used de Bruijn graph metagenomic assembler for second-generation short-read data was MEGAHIT [68]. Compared to MetaVelvet, MEGAHIT leverages succinct de Bruijn graphs, which is a compressed version of a de Bruijn graph. While difficult, it is much more efficient. Additionally, MEGAHIT also was able to take advantage of the computational power that had begun to be added to graphics cards. Graphics cards have the ability to massively parallelize small computational tasks to greatly enhance the speed of processes. MEGAHIT was able to optimize the run time to 3-5 times quicker than the regular de Bruijn graph assemblers that came before that solely relied on CPU for processing power [68]. MEGAHIT also introduced a step for removing k-mers that only appeared once in the dataset to prevent the use of sequencing errors as part of the assembly. Sequencing errors cause spurious edges and bubbles in the graph. MEGAHIT also implemented multiple k-mer sizes for the building of multiple succinct de Bruijn graphs adapted from IDBA-UD [69], an iterative graph-based assembler. In this approach, the smaller k-mer sizes are useful for filtering edges and gaps from low coverage regions, and large k-mer sizes are useful for repetitive regions in the assembly. Compared to previous metagenomic assemblers, the improvements made to the assembly algorithm yielded better assembly metrics across the board in nearly a tenth of the computational time [68].

The gold-standard of metagenomic assemblers for second-generation sequencing data has emerged to be metaSPAdes [70]. In an independent assessment of the top-of-the-line metagenomic assemblers, metaSPAdes outperformed all other compared assemblers, including MEGAHIT, in assembly metrics across multiple datasets [71]. The design of the metaSPAdes assembly algorithm made improvements on the ability to resolve samples with high levels of microdiversity. The focus of the core algorithm is to construct a consensus backbone while ignoring some levels of the strain-specific features. This principle was previously applied in SPAdes [72] to prune 'bulges' and 'tips' in the assembly graph that represent sequencing errors or exceedingly rare variants. In metaSPAdes, these parameters were weighted to accommodate the high complexity and microdiversity. Part of the SPAdes and metaSPAdes algorithms is the module exSPAnder which aims to utilize paired-end reads to resolve repeats in the assembly graph [73, 74, 75]. Like MetaVelvet, metaSPAdes also uses differential coverage to help separate the subgraphs. Together, all of these advances help metaSPAdes consistently produce highly contiguous and accurate assemblies from short-read data.

Third-generation single-molecule sequencing allows for the sequencing of reads that are far longer than what could be produced with PCR-amplified Illumina libraries. With these longer reads that have obvious benefits and drawbacks that need to be accounted for, a new wave of assembly algorithms needed to be engineered. One such assembler is Canu, which was designed specifically to accommodate noisy and inaccurate reads produced by sequencing platforms such as the Oxford Nanopore Technologies line of sequencers [76]. With longer reads once again being used, the assembly strategy once again returned to a form of OLC that was commonly used for assembly sequencing data produced by Sanger sequencing. To overcome the slow and computationally intensive process of overlapping the reads to build the graph, the authors instead used a MinHash Alignment Process [77]. In essence, the reads use a k-mer hash to seed for candidate overlaps rather than conducting full end to end alignments. This concept was combined with much of the same base algorithm as the Celera Assembler [78] to create an assembler well-adapted to third-generation sequencing. However, while Canu was used for metagenomic assemblies, it was not explicitly designed for that purpose. It was rather designed as a replacement for the Celera Assembler for assemblies of single genomes.

An assembler that was explicitly developed for the assembly of metagenomic datasets is metaFlye [79]. Like metaSPAdes, metaFlye is an extension of a single-genome assembler Flye [80]. Also similarly to metaSPAdes, metaFlye has become the gold standard for metagenomic assembly using long reads. In an independent assessment of the third-generation sequencing metagenomic aseemlbers, metaFlye consistently outperformed all other assemblers [81]. Similarly to Canu, metaFlye builds a set of k-mers to find overlapping reads that can build the assembly graph. metaFlye primarily looks for high-frequency k-mers to find this overlapping set of reads, which can be detrimental when there is uneven coverage of species, because the k-mers that comprise the low abundance species will not be recognized. To address this, metaFlye combines global k-mer counting with local k-mer distributions to find solid k-mers from the lower abundance species. Overall, metaFlye is performant with complex communities of bacteria, even with high levels of microdiversity.

Another method of assembly that arose alongside third-generation sequencing is hybrid assembly which utilizes both long and short reads to construct the assembly graph. One of these assemblers is the metagenomic assembler UniCycler [82]. Hybrid assemblers attempt to take advantage of the strengths of both second- and third-generation sequencing to mask the weaknesses of both. Second generation sequencing is highly accurate, but struggles to assemble through repetitive regions. Long-reads are able to span these regions, but produce assemblies with high levels of error, especially at homopolymers. Unfortunately, UniCycler and other hybrid metagenomic assemblers fail to produce comparable quality of assemblies when compared to Canu or metaFlye [81].

Instead, current best practice is to assemble with long-read assemblers such as metaFlye and then polish the sequences to remove errors. A number of tools exist to fix these errors utilizing both short and long reads. Pilon was developed to polish genomes using Illumina reads by aligning the reads to the genome and assessing areas for consistent, problematic alignments [83]. These principles were extended to polishing genomes with only long-read data in tools such as Homopolish, PEPPER, and Medaka [84, 85, 86], which were able to polish genomes to high-quality levels [87]. Simply assembling metagenomes is not a sufficient analysis. Even with the best sequencing data and the most sophisticated assembly algorithms, metagenomic assemblies yield heavily fragmented assemblies, so the genomic fragments need to be sorted and clustered. Additionally, the genomic sequence alone is not informative. Annotations and taxonomic assignments need to be added on to glean useful information from the assembled genomes.

1.2.4 Metagenomic binning

To overcome the pitfall of fragmented assemblies being output from metagenomic assemblies, the process of metagenomic binning was developed. Broadly, these methods attempt to cluster metagenomically-assembled contigs based on their measurable features. Such features can include sequence composition (as measured by k-mer frequencies or GC%), contig abundance, or assembly graph connectivity. Some forms of binning attempt to preempt the issue of fragmented, mixed assemblies by partitioning the reads instead prior to binning [88, 89, 90, 91, 92, 93]. Typically, these algorithms strictly use k-mer frequencies to cluster the reads. Due to the low information content in short reads and the error frequency of longer reads, these methods are not as effective as the binning of contigs following a metagenomic assembly [94].

One such binning algorithm for metagenomic assemblies is CONCOCT [95]. CONCOCT utilizes the sequence composition and contig mapping coverage across multiple samples to cluster contigs together into bins. The k-mer composition and coverage of each contig are transformed into a combined model projected onto a principal-component plot to reduce the dimensionality of the data while maintaining the vast majority of the information. On this projection, the contigs are clustered using a Gaussian mixture model. Another binning algorithm that utilizes the coverage and sequence composition data to produce metagenomic bins is MetaBAT2 [96]. MetaBAT2 brings additional steps to the algorithm including normalized tetranucleotide frequency scores, an iterative graph partitioning procedure, and further steps to include small contigs into bins. Of the many binning algorithms that utilize coverage and composition to bin the contigs of a metagenomic assembly, MetaBAT2 consistently outperforms the others [97]. Another binning algorithm, which is aimed to compliment the composition and coverage binning algorithms, is GraphBin [98]. GraphBin refines the bins produced by the other algorithms by employing the information that is contained within the assembly graphs. Assembly graphs contain information on the connectivity of contigs that are useful to refine the bin classifications produced by the other binning algorithms. DAS Tool fills a similar niche in that it is a binning tool to refine the outputs of other binning algorithms [99]. Rather than performing a secondary layer of binning to refine the binning results of the other tools, DAS Tool creates an aggregated scoring strategy to create a consensus binning result from the output of the other tools.

There are a handful of issues with metagenomic binning. One common issue is the generation of chimeric bins and assemblies [100]. Chimeric bins occur when contigs with true identity to separate bacteria are included in a single metagenomic bin. In part, DAS Tool's use of a consensus of binning from multiple sources has shown to improve the accuracy [99], however chimerism still remains an issue. To check for the completion level of a metagenomic bin, and whether there is contamination in the bin, a tool such as CheckM can be utilized [101]. CheckM measures the presence of the conserved single-copy core bacterial genes for assessing completion and detects whether any of these genes are deduplicated to measure contamination. Another shortcoming of binning comes in its inability to properly bin plasmids and genomic islands. Maguire *et al.* [102] demonstrated that the binning rate of plasmids and genomic islands were well below 40% whereas the background binning rate of contigs is generally in the 80-90% range.

Metagenomic bins of high-quality, as measured by tools such as CheckM, are referred to as metagenomic-assembled genomes (MAGs). Annotation and quantification of these MAGs are what forms the bulk of metagenomic analyses.

1.2.5 Functional Annotation of Metagenomic-Assembled Genomes

One of the primary benefits of shotgun metagenomic sequencing over amplicon sequencing of the 16S rRNA gene sequencing is the ability to directly assess the metabolic capabilities of a metagenome, not just the inferred capabilities from taxonomy. To annotate MAGs, it is generally a two-step process. First, the open reading frames of the bacteria are predicted to obtain protein sequences. These protein sequences are then aligned to databases to predict the gene, or pathway, that they belong to.

There are a handful of tools available that are able to predict the bacterial open reading frames. Some annotation pipelines choose to bypass this step and align the genomic sequences directly to a nucleotide database, such as with MEGAN4 [103]. This is generally not the accepted strategy as protein sequences are more information dense and easier to align to databases. Instead programs such as Glimmer and Prodigal are used to first predict the open reading frames and translate them into protein sequences [104, 105]. Glimmer uses interpolated Markov Models to predicted the start and stop codons within a bacterial genome [106]. In essence, bacterial genes have regular patterns and features that can be modelled by Markov models, which allows for the accurate prediction of their boundaries in Glimmer's algorithm. Sequencing and assembly errors are a major hurdle for gene prediction algorithms. These errors can commonly cause frameshifts or premature stop codons, which would cause erroneous protein sequences to be predicted. Both Glimmer and Prodigal attempt to fix these errors by modelling the base accuracy and looking for overlapping reading frames indicative of an error at the low quality sites. With Illumina reads, insertions and deletions that cause frameshifts are rare [107], so the impact of errors in assemblies derived from such data are low. However, indels (especially at homopolymers) are common in metagenomic sequencing from long-read sequencing on Nanopore platforms, so these corrections are critical for those assemblies. Prodigal builds on Glimmer's algorithm by adding machine learning models trained on verified bacterial open reading frames to better consider the GC bias of open reading frames and start codon bias [105]. Prodigal calculates scores of every start and stop codon pair in the genome and considers the presence of a Shine-Dalgarno sequence [108] and ribosomal binding sites to determine whether the pair of start and stop codons represents a true open reading frame.

There are two main computational frameworks to align the predicted protein sequences to the available databases, BLAST and hidden Markov models. Basic local alignment search tool, or

BLAST, is a simple and robust method for querying databases with either nucleotide or protein sequences [109]. The core of the algorithm is the splitting of the query sequence into a set of k-mers, searching the database for a sequence that contains a match for the seed k-mer, and then extending from that match to generate an alignment score and find high-scoring segment pairs. The scoring parameters can be adjusted to assign weighting for matches, mismatches, and gaps. Multiple high-scoring segment pairs can be combined on a sequence to form a longer alignment. Once a BLAST 'hit' is found a full, gapped Smith-Waterman alignment will be generated to the matched sequences. Many forms of BLAST can be used depending on what queries and databases are being used. The most basic forms of BLAST, BLASTN and BLASTP, perform nucleotide-to-nucleotide and protein-to-protein alignments, respectively. PSI-BLAST is a form of protein-to-protein BLAST that is more sensitive to distant evolutionary relationships in protein sequences that the standard BLASTP alignment [110]. Another method to map to a protein database is BLASTX which translates a DNA query in all six potential open reading frames and aligns these potential sequences against the database. This is effective to find protein coding sequences in a stretch of genomic DNA or to align cDNA to a protein database. TBLASTX also utilizes all six open reading frames of a nucleotide sequence as a query, but unlike BLASTX, it aligns these sequences to the six open reading frames of a nucleotide database. TBLASTN takes a protein query and aligns it to the translated six open reading frames of a nucleotide database [111]. Each of these variations has its own optimal use case depending on the available data and the questions that need to be answered. Accelerated forms of these BLAST algorithms have been implemented in bioinformatic software that is designed to run on local servers and or cloud computing clusters. For example, the BLASTX module of the DIAMOND alignment tool was benchmarked as being roughly 20,000 times as fast as conventional BLASTX [112].

Hidden Markov models, and more specifically profile hidden Markov models have proven to be effective for searching nucleotide and protein databases for distantly related sequences [113]. Profile hidden Markov model alignments are similar to PSI-BLAST in principle. Certain positions in protein sequences are more conserved in proteins than others and this position-specific information helps to build and score the alignments. The probabilities of the alignment models are built off global alignments of protein families. The alignments help to identify the conserved protein positions, regions of variability, and tolerance of insertions and deletions at a position. The probabilities in a pHMM are commonly converted to additive log-ratios prior to alignment [114]. Dissimilar to BLAST is how each residue is scored. Because the probability of insertions and deletions is calculated within the alignments, the scoring of gaps and insertions is not arbitrary [113]. Profile hidden Markov model alignments have been implemented in HMMER3, which has sped up the alignment to the point where it is now as nearly as fast as BLAST for protein searches [115]. Previously, the sensitivity of pHMM methods were heavily offset by the much slower speeds, but the improved speeds brought by HMMER3 bring the speed to a point where it is appropriate to use on large-scale metagenomic datasets.

For both methods of gene alignment, there are databases available to effectively annotate bacterial MAGs. Perhaps the most comprehensive database to align sequences to is the UniRef and UniProtKB databases [116, 117, 118, 119]. The Uniref (UniProt Reference Clusters) databases are built from a clustered set of sequences from the UniProtKB (UniProt Knowledgebase) database [118]. There are three forms of the UniRef databases: UniRef100, UniRef90, and UniRef50. UniRef100 was constructed by combining all identical sequences and subfragments into single entries in the database. The UniRef90 and UniRef50 databases were constructed by clustering the UniRef100 databases at 90% and 50%, respectively. Each of the databases greatly reduce the size of the databases resulting in reduced memory requirements and increased search speeds. When using the UniRef50 database, cluster homogeneity, measured by Gene Ontology terms, was at over 97% [120], search speed was improved six-fold compared to the UniProtKB database, and was more sensitive to remote similarities [117]. The UniRef databases are formatted as a set of FASTA entries, so they are best searched using a form of BLASTP such as the accelerated BLASTP module of the DIAMOND aligner [112]. A pHMM-based database that is similarly comprehensive to the UniRef databases is the Pfam database [121, 122, 123, 124]. Pfam is a manually curated database where the seeds of clusters are manually annotated and sorted into protein domains and families.

Novel sequences can be added to these clusters by various protein clustering and alignment tools. At the 2021 release of Pfam it represented 77% sequence coverage of the UniProtKB database. For some applications, it is preferable to use a heavily curated database where the pathway and function of each database entry is known rather than drawing from a database where the vast majority of entries have unknown functions. One such database is the Clusters of Orthologous Genes (COG) database [125, 126, 127, 128]. In concept, it is very similar to the Pfam database where it is constructed by clustering protein sequences into families of proteins. In contrast to the Pfam database, which is constructed by comprehensively curating the entirety of the UniProtKB database, the COG database is built from a limited set of bacterial and archaeal genomes. As of the 2021 release of the COG database, there were 1187 bacterial genomes and 122 archaeal genomes used resulting in roughly 5000 clusters of orthologous genes [125]. Recent focus has been invested in improving the diversity of sequences for CRISPR-Cas immunity, sporulation, and photosynthesis [125]. On top of the assignment of the gene/cluster name through annotation, the gene clusters have additional information regarding the biological pathways and protein domains borrowed from databases such as Pfam and InterPro [119, 129]. The structure of the database as clusters of orthologues from bacterial species makes the COG database very effective for the annotation of novel species that are not represented in the database and for conducting comparative and evolutionary analyses. Another focused database that focuses on curation of gene function is the Kyoto Encyclopedia of Genes and Genomes (KEGG) [130, 131, 132]. KEGG was developed to address the need for a method to biologically interpret sequencing data. KEGG connects collections of predicted genes to high-level functions within the cell. The pathway and functional assignments are backed by experimental data with over 75% of the 19000 KEGG entries having PubMed reference links [130]. The genomes used to build KEGG are primary pulled from the NCBI RefSeq database [133], with additional prokaryotic genomes being pulled from NCBI GenBank [134]. Annotation of genomes or metagenomes with KEGG can be conducted using BlastKOALA or GhostKOALA [135]. As the name would suggest, BlastKOALA utilizes BLAST to perform alignment searches for matches. GhostKOALA utilizes GHOSTX, which is much more computationally efficient at the expense of sensitivity [136]. KEGG is also integrated into Anvi'o, which offers modules such as 'anvi-estimate-metabolism' that can estimate the completion of metabolic pathways in a MAG or in the full metagenome [137].

By utilizing multiple sources of annotation, it is possible to make accurate estimations and hypotheses of the metabolic potential of bacteria, even those that have been previously uncharacterized. This ability is undoubtedly the defining strength of shotgun metagenomic sequencing of bacterial communities. As the tools and databases continue to improve, phenotypes of uncultivatable bacteria will be possible to predict without biochemical assays.

1.2.6 Taxonomic Assignment of Metagenomic-Assembled Genomes

Taxonomic assignment of MAGs is similarly important in metagenomic analyses as it is in amplicon sequencing experiments. Taxonomic information of MAGs assembled from a metagenome can place the metabolic pathways predicted to a certain species or assess the taxonomic composition of a microbiome. Because of increased breadth of information there is a greater diversity of methods for assigning taxonomy to a MAG.

As demonstrated with amplicon sequencing, the 16S sequence is an adequate means of assigning taxonomy to a sequence. Within a MAG, the 16S rRNA sequence should be present in at least a single copy if the MAG meets quality control standards. These sequences can be predicted from the genomic sequence with hidden Markov models, which are implemented in programs such as RNAmmer or Barrnap [138, 139]. These predicted sequences can be aligned either locally or through a web server to the SILVA database to assign taxonomy [140]. Using the full sequence of the 16S rRNA gene greatly improves taxonomic assignment compared to the short fragments generated by amplicon sequencing [141]. While considerably more reliable than amplicon sequences, using only the 16S rRNA gene prevent phylogenetic resolution of closely related organisms [142]. Using the 16S rRNA gene sequence to assign taxonomy is an unnecessary limitation when the full genomic sequence could be used.

A form of utilizing the breadth of information present in MAGs has been present in the Genome

Taxonomy Database Toolkit (GTDB-Tk) [143]. Rather than using a single locus to assign taxonomy, it uses 120 bacterial marker genes to assign the taxonomy. For each genome, the ORFs are predicted using Prodigal [105] and aligned to the GTDB marker gene set using HMMER [115]. The reference marker gene set is built from a genome set of over 23,000 bacterial genomes pulled from the NCBI Assembly database [144]. The predicted marker genes from the MAG are concatenated and phylogenetically placed on the domain-specific tree using pplacer [145]. A genome's placement, its relative evolutionary divergence [146], and its average nucleotide identity to reference genomes are the bases of the taxonomic assignment. The tree placement gives the coarse taxonomy and the relative evolutionary distance and average nucleotide identity help to provide a taxonomic assignment at the species level. Another bioinformatic tool that utilizes marker genes is PhyloPhlAn [147]. PhyloPhlAn uses a greater breadth of markers (>400 proteins), but is built from a smaller set of genomes (roughly 87,000). Both marker gene methods outperform taxonomic assignment using the 16S rRNA gene as they are better able to differentiate closely-related organisms and account for events such as horizontal gene transfer.

In an attempt to use as much of the available data as possible, some taxonomic classification algorithms use taxonomic information gathered from all predicted open reading frames. Contig Annotation Tool (CAT) utilizes the taxonomic information that is carried by entries in the NCBI non-redundant (nr) protein database [148]. Within the NCBI-nr database, each protein entry has an associated taxonomy of the genome that the protein sequence originated from. For each open reading frame, the matches within the NCBI database are piped to a last common ancestor (LCA) algorithm that determines the taxonomic level that can be assigned. The taxonomic assignments for each open reading frame are scored based on confidence and aggregated to create a consensus taxonomy for the genome with associated probabilities for the assignment. Another taxonomic assignment algorithm that uses the taxonomic assignments from protein coding sequences is MMseqs2 [149]. One of the major drawbacks of CAT is the computational resources required to run it and the slow speed which is a result of aligning each open reading frame to the full NCBI-nr database with BLAST. Additionally, the use of Prodigal [105] is a time-limiting step due to its inability to be multithreaded. MMseqs2 instead uses the translated six possible open reading frames and has flexible options for the databases used. Overall, MMseqs2 can run up to eighteen times as fast as other taxonomic prediction algorithms with the improvements made [149].

Taxonomic assignment is critical to answer the core metagenomic question of ‘who is there?’ Taxonomic assignment tools can often give divergent answers, so corroborating the assignment with multiple, high-quality sources is a good practice. Using a combination of a marker gene and a full protein alignment will likely result in a high-confidence taxonomic prediction of a MAG. However, many species remain uncharacterized in the databases and lack closely-related common ancestors, so more rigorous means of defining these species is required to understand their place on the tree of life.

1.2.7 Assembly-Free Analysis of Metagenomic Data

Depending on the research questions being asked, sometimes it is not necessary to assemble shotgun metagenomic data. In other instances, assembly-free methods are effective for validating the findings of assembly-based methods. Assembly-free methods are capable of answering both core metagenomic questions of ‘who is there?’ and ‘what are they doing?’.

For long and short reads there are separate tools to determine the taxonomic composition of the metagenome. Specifically designed for accurate short read data from Illumina instruments in the bioinformatic tool MetaPhlAn [150]. Compared to the computationally expensive process of assembling MAGs and annotating them, running MetaPhlAn is very efficient. Metagenomic reads are aligned against a curated set of protein-coding sequences that are strong indicators of bacterial clades. Within the MetaPhlAn database, there are over 1,200 species present with an average of 84 genetic markers per species to assign species-level taxonomy to a read. Additionally, there are over 100,000 markers for high-level taxonomy to assign reads at a higher level if species level assignment is not possible. MEGAN also is a tool that assigns taxonomy to reads, but unlike MetaPhlAn it also has been designed to handle the high error rate of long reads [151, 152]. Alignment in the six possible open reading frames to protein databases with typically is performed with BLASTX, however it is not the optimal alignment algorithm for error-prone long reads. Instead, MEGAN uses LAST for alignment [153, 154]. LAST is a frameshift aware aligner for long reads and is far more suitable than a simple BLASTX. Additionally, MEGAN and MetaPhlAn both offer visualization tools to help interpret the data that they generate. Kraken2 uses k-mer searching to assign taxonomy using an LCA algorithm and attempts to comprehensively assign taxonomy to all reads rather than subsets that contain marker genes such as MetaPhlAn [155].

HUMAnN3 addresses the question of what functions are present in the metagenome and what their composition is [147]. The first step in the HUMAnN3 workflow is to identify the known species present in the community. Once the species are identified, the reads are mapped to pangenomes of each of the identified species, which captures the strain variation in functions of the species. In addition, reads that cannot be assigned a taxonomy present in the HUMAnN3 database are classified using a translated search to assign the function. The output of HUMAnN3 is a table of functions and proteins and the normalized abundance of reads that mapped to them.

The increased computational efficiency of assembly-free methods is a distinct advantage over assembly-based methods and state-of-the-art assembly free methods (MetaPhlAn, MEGAN, and HUMAnN2) are reasonably reliable for most datasets. However, like with methods like taxonomic assignment with only the 16S rRNA gene sequence, many of the benefits of shotgun metagenomic sequencing are unused with these tools. Assembly-free methods can fall short at detecting strain-level differences or when data are being analyzed from poorly studied communities of bacteria.

1.2.8 Differential Analyses of Metagenomic Data

One of the primary objectives of metagenomic research is to examine the differences between groups. The are numerous ways that a table of counts can be generated for samples that can be used as metrics of comparison. For instance, assembly-free methods of taxonomic and functional assignments are designed to create such tables to be analyzed. For an assembly-based approach, reads can be mapped back to the generated MAGs to quantify them with alignment tools such as Minimap2 for long reads [156] or Bowtie2 for short reads [157]. The alignment files can be processed with SAMtools [158] to generate a table of counts of the number of reads mapping to each MAG or contig. Another way of creating quantitative data to assess the metabolism of a metagenome is by performing metatranscriptomic analyses on a community. Metatranscriptomics sequences the actively transcribed mRNA sequence space of a community. Optimally, metatranscriptomic data are aligned to the predicted and annotated coding sequences of an assembled metagenome. Metatranscriptomics allows for conclusions on what genes are being expressed rather than just what genes simply are present in the most abundant bacterial species in a community.

Specialized tools needed to be developed to address the nature of the data and the questions being asked. Metagenomic and metatranscriptomic datasets are extremely high dimensional and are often underpowered, so the statistical tools developed needed to account for these challenges. Two commonly used statistical tools are edgeR and DESeq2 [159, 160]. However, these tools fail to account for the compositional nature of metagenomic and metatranscriptomic datasets [161]. Because the data are limited by the fixed capacity of the instrument, the data are compositional. If the measure abundance of a feature goes up (i.e. a bacteria in a microbiome dataset) because of the increased presence in a sample compared to another, then the measured abundance of something else must go down even if the true abundance of the other feature did not change. This characteristic of composition data necessitates special transformation to apply conventional statistical tests, which are often log-ratio transformations. Indeed, log-ratio transformations have been benchmarked to be much more reliable and provide fewer false positives [162]. Differential abundance analyses feature centred log-ratio transformation have been implemented in ALDEx2 and have proven to provide consistent results in datasets with high dispersion or asymmetry [163]. Compositional data analysis tools can also find associations between clinical metadata and certain bacterial taxa [164]. Differential abundance analyses help to elucidate the bacteria that are of interest in complex communities. However, many of the tools available are prone to false positives. With the sheer number of bacteria in complex communities, false positives are inevitable, but must be minimized through improved algorithms and models. Additionally, differential abundance analyses of high-throughput sequencing is not a complete replacement for benchtop science, but rather a way of generating working hypotheses to be validated *in vitro* or *in vivo*.

1.2.9 Pangenomics of Bacterial Genomes

Beyond differential abundance analysis, shotgun metagenomic data can be used to construct large-scale pangenomes of bacterial clades. A pangenome refers to the total genomic content of a bacterial clade of interest. A pangenome is built by clustering the protein coding sequences of the genomes into orthologues and determining which genes are core and which are dispensable [165]. Through these analyses, it is possible to make inferences about the role of genes in the adaptation of the bacterial clade to environments or the role of bacterial genes in disease pathology.

Many pangenomic analysis tools require a high computational burden to build a pangenome, which is problematic for the high-throughput nature of metagenomic data. For instance, PanOCT [166] and PGAP [167] both utilize an all-versus-all BLAST alignment for all open reading frames in the pangenomes, which even for small-scale pangenomes is very computationally intensive. LS-BSR [168] pre-clusters the coding sequences prior to BLAST alignments to reduce the number of alignments required to build the pangenome greatly. The pangenomic tool Roary [169] also implemented pre-clustering of sequences with CD-HIT [170] and removal of partial sequences, which resulted in a run time of 4.5 hours and a memory usage of 13GB of RAM for a pangenome of 1000 isolates, which is feasible to run on a personal laptop. These step-wise improvements to the pangenomic workflow transformed a computationally intensive process into a workflow that can be run on a low-grade personal computer. Perhaps the best implementation of a pangenomic workflow is built as a module in the Anvi'o toolkit [137]. Anvi'o utilizes the 'minbit' heuristic from the ITEP pangenomic workflow [171] to eliminate weak amino acid matches from the BLAST alignment inputs. Additionally, the MCL algorithm [172] is used to cluster sequences prior to the BLAST alignments with DIAMOND [112] to further reduce the number of alignments required. The true benefit of the Anvi'o pangenomic pipeline is the ability to generate high-quality and interactive visualizations of the pangenome. The interactive interface of Anvi'o allows for the manual curation of the core and accessory (i.e. dispensable) genomes for the clade. Additionally, functional enrichment is implemented in the pangenomic workflow, so comparisons of the strains between groups can be calculated for potential genes involved in disease pathology.

Pangenomics has been applied to metagenomes from widely different environments with different research questions. From a large-scale metagenome survey marine samples the pangenome of the *Prochlorococcus* genome was constructed [173]. The pangenome of *Prochlorococcus* revealed an enrichment of hypervariable gene clusters related to sugar metabolism, suggesting a fitness benefit for harbouring a diverse set of sugar metabolism genes for these marine bacteria. Additionally, in human studies, a pangenome of *Ruminococcus gnavus* was constructed from stool samples of inflammatory bowel disease (IBD) patients (and healthy controls) to identify IBD-specific genes in these bacteria. These IBD-specific genes in *Ruminococcus gnavus* were primarily related to the functions of oxidative stress responses, adhesion, iron-acquisition, and mucus utilization, which could be related to disease pathology [174].

Pangenomic analyses have come a long way in terms of efficiency and power to analyze minute details in bacterial strains. Pangenomics are a powerful tool to complement differential abundance analyses as they can help explain the enrichment of clades in an environment or explain the differences in community phenotype in the absence of differentially abundant clades.

1.3 Applications of Metagenomics

Metagenomic analyses, and the improvements made to them, were developed with the applications to research fields in mind. Perhaps the most prominent field of metagenomic research has been the study of the bacteria that colonize the human body. Bacteria colonize the skin [175], the oral cavity [176, 177, 178], the lungs [179], and most notably the intestinal tract [180, 181, 182]. Many of the bacterial species that colonize the human intestinal tract are uncultivated as yet [183], which makes culture-independent metagenomic and metatranscriptomic approaches the primary methods to analyze them. Human and animals are not alone in being colonized by bacteria as many plant species have close symbiotic relationships with bacteria as well [184, 185]. Other freelyliving environmental bacteria have been discovered to have potential for remediating a wide variety of anthropogenic pollution [186, 187]. As with studies of the human microbiome, the majority of environmental bacteria that are involved with plant growth or bioremediation are difficult to cultivate or uncultivatable and their study has been enabled by the improvements to sequencing technologies and bioinformatic analyses.

1.3.1 Human Health and the Human Microbiome

Many areas of the human body are regularly colonized by bacterial communities. These bacterial communities can form close associations with human tissues and affect human health outcomes. These bacteria can both be protective and causative in disease in instances of dysbiosis, so accurate sequencing and analysis is clinically important.

The skin microbiota, and the dysbioses associated with it, is associated with a number of clinical outcomes. The skin is a challenging environment for bacteria to thrive and is dissimilar to the intestinal environment that gut microbes can thrive in. The surface of the skin is high-salt, acidic, and primarily aerobic [188]. Depending on the location and skin type the normal and healthy compositions of bacterial species can differ greatly. For instance, while species of *Staphylococcus* are found broadly on the body, *Cutibacterium* is concentrated on the face and torso [189] whereas *Corynebacterium* species are concentrated in areas such as the armpits and elbow [190]. Not only the species-level composition is important because different strains of skin microbiota species can have very divergent phenotypes [191]. In a healthy state, the commensal bacteria of the skin play an important role in maintaining immune homeostasis [192]. This includes the regulation of interleukin signaling and the complement cascade [193]. In addition, the skin microbiota also produce a number of antimicrobial peptides [194, 195]. Production of antimicrobial peptides, among other competition mechanisms, help the commensal skin bacteria inhibit the growth of the respiratory tract pathogen *Streptococcus pneumoniae* [196]. *Staphylococcus aureus*, an opportunistic skin pathogen, can also be stimulated to become less pathogenic by *Corynebacterium* commensal strains [197]. Such control of *Staphylococcus aureus* is important in maintaining skin health because an estimated 30% of the population is colonized asymptomatically with *Staphylococcus aureus* [198]. Actively pathogenic *Staphylococcus aureus* can cause abscesses and is associated with severe atopic dermatitis [191]. Another clade of skin colonizers that can become pathogenic is the genus *Mycobacterium.* Species within this genus are known to cause leprosy [199], tuberculosis [200], and Buruli ulcers [201]. Like *Staphylococcus*, *Mycobacterium* are long-term colonizers and opportunistic pathogens that can remain dormant for the entire lifetimes of individuals. It is estimated that up to a quarter of the world's population is colonized by *Mycobacterium tuberculosis*, and despite it being one of the most deadly diseases worldwide, proportionally it only causes disease in a small fraction of the total colonized population [200]. Because of these characteristics, many of the colonizers of the human skin are considered pathobionts, where under regular circumstances they engage with mutualistic behaviours with the human skin, but under circumstances of immune suppression can become pathogenic [202]. On the less severe end of the spectrum of these 'pathobiont' relationships is *Corynebacterium acnes*, which is involved with the development of acne. *Corynebacterium acnes* is in a delicate balance with *Staphylococcus aureus* in the facial microbiota and disruption of the balance can induce skin inflammation and eventually acne. Though many locations of the skin microbiota are aerobic other areas are comparatively anaerobic [188]. Research of communities living in anaerobic follicles, variations in the compositions of the skin microbiota resulting in inflammation and disease, or even community surveillance of antimicrobial-resistant *Staphylococcus aureus* is greatly enabled by cutting-edge metagenomic techniques.

Significantly more complex than the skin microbiota is the microbiota of the oral cavity. Over 700 microbial species colonize the human oral cavity [203], which makes it the second most complex microbiota in the human body behind the gut microbiota [204, 205]. At this level of complexity, isolation of individual strains is not practical, which makes metagenomic approaches to studying this complex environment optimal. Additionally, over half the bacteria of the oral microbiota are uncultivatable [206]. A variety of disease-causing viruses, such as *Herpes simplex* [207], are common to the oral cavity, but the vast majority of viral sequences are bacteriophage in origin [208].

The study of the bacteria that colonize the mouth, and that are infected by the aforementioned bacteriophage, is perhaps the oldest form of microbiota study with Antony van Leeuwenhoek first studying the bacteria of oral plaques in the seventeenth century [209]. Like the skin microbiota, the composition of bacteria is dependent on the site within the cavity [210]. The bacterial phyla of *Bacteroidetes, Proteobacteria, Actinobacteria, Spirochaetes, Fusobacteria*, and *Firmicutes* are the most predominant clades of bacteria that colonize the mouth [211]. There is evidence of a core oral microbiota, with 47% of species-level OTUs being shared between individuals [212]. Similar to the skin microbiota, the colonization of the mouth with commensal bacteria is protective against the infection of pathogens such as *Staphylococcus aureus* [213, 214]. For example, *Streptococcus salivarius* produces a bacteriocin that inhibits the growth of periodontits-causing and halitosis causing bacteria [215, 216]. One of the most common negative health outcomes associated with the oral microbiota is tooth decay, or dental caries, which are in part a result of acid produced by carbohydrate fermentation by oral microbes [204]. Following ingestion of high-levels of carbohydrates, the composition of the oral mircobiota shifts towards one that produces a high amount of acid [217]. Gingivitis has a adult prevalence of over 90% [218] and is caused by the formation of bacterial plaques on the tooth surface. *Actinomyces* species act as the primary colonizers of the surface of the teeth and then the full biofilm forms by coaggregation interactions with other bacterial species [219, 220]. Gingivitis is not necessarily associated with a specific bacterial species, but rather a general overgrowth, and the greater the plaque load, the more severe the disease severity [221]. The oral microbiota is complex, but metagenomic analyses have shown that it is self-regulating with regular dental care and shares a core microbiome between individuals [204, 212].

The most diverse microbiota of humans, which is associated with the a great number of clinical outcomes, is the gut microbiota. Up to 60% of the total dry mass of faeces is composed of bacteria [222]. With the emergence of shotgun metagenomic sequencing, much research has been invested into the composition of these abundant and complex communities. Large-scale studies including the European Metagenomics of the Human Intestinal Tract (MetaHIT) project and the

Human Microbiome Project (HMP) aimed to broadly sequence the gut microbiome to create a similar consensus picture as the Human Genome Project [223, 224, 205]. Recently, using a similar approach, a near-complete and non-redundant set of bacterial genomes was assembled from metagenomic data [183]. This dataset represents over 2,500 bacterial genomes with roughly 2,000 being novel and uncultivatable bacteria identified as part of the study. These genomes were assembled from a comprehensive set of metagenomic samples that were available at the time of publication. Metagenomic studies exploring the association of these microbes with human health outcomes is being actively catalogued on the platform GMrepo [225, 226]. Nearly 72,00 samples are curated with metadata on the platform with a roughly 2:1 split of 16S rRNA gene sequencing and shotgun metagenomic experiments. At present there are 47 phenotype pairs recorded as metadata on GMrepo to associate the gut microbiota with different health outcomes.

Two of the most well-studied clinical disorders associated with the human gut microbiota are irritable bowel syndrome (IBS) and inflammatory bowel disease (IBD). IBS is characterized by recurrent abdominal pain over the course of a long period of time with abnormalities related to defecation, and it afflicts over 10% of the global population [227]. The composition of bacteria that colonize the human intestinal tract have frequent associations with IBS, with broad generalizations such as a relative increase of the phylum *Firmicutes* compared to the phylum *Bacteroidetes* being a common association [228, 229]. Another broad microbiota signature that is associated with IBS is species richness [230]. A lower richness of species found in the intestinal tract is associated with an increased risk of developing IBS. More specific observations have also been made with a greater relative abundance of the genera *Streptococcus* and *Ruminococcus* and lower relative abundances of the genera *Lactobacillus* and *Bifidobacteria* are associated with IBS [231, 232, 233, 234, 235]. IBD is a chronic inflammatory disorder that encompasses ulcerative colitis and Crohn's disease, which has an unclear pathogenesis. Like with the skin microbiota, it is believed that the intestinal microbiota plays a critical role in the development of regular localized immune responses and dysbiosis can play a key role in the development of IDB [236]. Inverse to the microbiota signature seen in IBS, IBD is frequently characterized by a relative increase in abundance of the phylum *Firmi-*

cutes compared to the phylum *Bacteroidetes* [237, 238]. However, similarly to IBS, the microbiota of IBD is characterized by a reduced species richness [239]. *Faecalibacterium prausnitzii* is a gut microbe whose abundance is strongly linked to a protective phenotype for IBD the the abundance is correlated with maintenance of clinical remission [240]. *Faecalibacterium prausnitzii* has been shown to reduce the production of pro-inflammatory cytokines in *in vitro* and *in vivo* models, which may help to explain how it is protective against IBD [241]. In contrast, colonization of the intestinal tract with pro-inflammatory bacteria, such as *Escherichia coli*, is associated with an increased risk of IBD [242, 243, 244]. The metabolic byproducts of certain bacteria can also possess proinflammatory effects. For instance, the intestinal epithelium is damaged by hydrogen-sulfate produced by *Desulfovibrio* and results in mucosal inflammation [245].

Cardiovascular and metabolic disease have also been associated with the composition of the gut microbiota [246]. Serum lipid levels are a major risk factor for cardiovascular disease [247], however, much of the variation in lipid levels cannot be explain by genetics or diet alone [248]. Recent research has revealed that as much as 25% of the variation in high-density lipoprotein levels may be attributable to the gut microbiota [249].

Another class of blood lipid that appears to have its level mediated in part by the microbiota are triglycerides [249]. However, the direct relationship has not been explored directly [246]. For example, as a secondary outcome, triglyceride levels were found to be reduced following bran [250] and inulin [251] ingestion, which are prebiotics that alter the composition of the gut microbiota. A well-studied cardiovascular disorder with a clear mechanism of action in the pathogenesis of the disease is atherosclerosis. The metabolism of L-Carnitine or phosphatidylcholine by the gut microbiota results in the metabolite trimethylamine, which is then converted by the liver into trimethylamine N-oxide (TMAO) [252, 253]. TMAO is directly linked to the pathogenesis of atherosclerosis, with serum levels contributing 11% of the total variation in risk for the disease [254, 255]. TMAO is also a case study in treatment of a disease through modulation of the gut microbiota. By using a choline analogue, Wang *et al.* [256] found that the production of trimethylamine by the microbiota could be reduced without the negative side effects that the blockade of the conversion of trimethylamine to TMAO resulted in [257]. Metabolic disorders such as obesity and type 2 diabetes have also been frequently associated with microbiota composition [258, 259, 260]. Like with other inflammatory-based gut disorders like IBS and IBD, the microbiotas of obesity and type 2 diabetes are characterized by decreased microbial diversity [249, 261, 262]. A major mediator of intestinal inflammation in the gut that is produced by the microbiota are short-chain fatty acids. A decrease in the relative abundance of bacteria that produce the short-chain fatty acid butyrate is associated with the onset of type 2 diabetes [262]. Short-chain fatty acids are potent inducers of regulatory T cells that mediate inflammatory responses in the gut [263]. A major limitation in the study of cardiovascular and metabolic diseases in association with the gut microbiota is the over-reliance on 16S rRNA gene sequencing over shotgun metagenomics [246]. Mechanisms of action are largely speculative with only taxonomic information and hopefully with the lowering cost of sequencing and improved metagenomic analyses available, more studies utilize shotgun metagenomics to better understand the involvement of the gut microbiota in these disorders.

Beyond choline analogues for the treatment of atherosclerosis, there have been a handful of health conditions that have use the gut microbiota as the focus of treatment. Faecal microbiota transplants (FMTs) are one such treatment modality that aims to replace the ‘pathogenic’ stool from an individual with the healthy stool from another [264]. FMTs have proven effective in the treatment of IBD [265] and recurrent *Clostridium difficile* infections [266]. Efforts have also been made to investigate the effectiveness of FMTs in some microbiota-associated metabolic disorders, such as type 2 diabetes, with some success [267]. However, FMTs can be an invasive procedure that can be undesirable to many individuals due to the ’ick‘ factor. Probiotics are a suitable, if less efficacious alternative, due to their packaging into an oral capsule. Perhaps most consistent are the ability of probiotics to modulate stool consistency and prevent outcomes such as antibiotic-associated diarrhea [268]. However, probiotics have also shown effectiveness in reducing the inflammatory markers in IBD patients [269]. With the ever increasing number of clinical outcomes associated with the gut microbiome, research into the effects and safety of FMTs and probiotics is of importance. Given the multifactorial nature of many diseases and disorders, the microbiota is yet another target for pharmaceutical development.

Though there is a distinct over-reliance on amplicon sequencing [246], there is a wealth of research on the associations between the bacteria that colonize the human body and clinical outcomes. As sequencing technologies and analysis methodologies continue to develop, the accuracy and consistency of the results will also continue to improve. For instance, it is now possible to generate lineage-resolved and complete metagenome-assembled genomes from a deeply sequenced gut microbiota [270]. Such advances will help explore the strain-level differences in the microbiota that can be impactful in health and disease.

1.3.2 Bioremediation

Bacteria do not need to colonize eukaryotic organisms to be of interest in metagenomic research. Environmental bacterial communities are able to metabolize or sequester pollutants such as heavy metals [271] and hydrocarbons, including plastics [186]. Anthropogenic pollution as a result of industrialism poses major threats to marine environments [272]. Pollution with oil and hydrocarbons is lethal to many marine organisms and millions of tonnes of plastic are deposited in the ocean annually. The dire nature of these pollutants in the environment necessitates research into how to most effectively remediate sites of contamination. Research has demonstrated that remediation using the naturally occurring biochemical pathways in bacteria (i.e. bioremediation) is more efficient and cost effective than abiotic processes [273].

Marine bacteria are well-adapted to growth in the harsh marine environment, which has fluctuating nutrient availability, pH, temperature, and salinity, and they also harbour biochemical pathways that could be exploited for bioremediation [186]. However, up to 99% of these marine species cannot be isolated in culture and must be studied through culture-free methods such as metagenomics. A great number of genera have been shown to degrade hydrocarbons including *Achromobacter, Alcanivorax, Halomonas,* and *Pseudomonas* [274, 275, 276, 277, 278]. Generally speaking, consortia of the hydrocarbon-degrading bacteria are more efficient than isolates because some only degrade intermediates or different compounds and environmental hydrocarbon pollution is usually a complex mixture [279, 280, 281]. Degradation of aliphatic hydrocarbons by bacteria involves membrane-embedded alkane hydroylase enzymes [282, 283]. For aromatic hydrocarbons, bacteria encode mono- or dioxygenases that cleave the aromatic ring to generate intermediates of the tricarboxylic acid cycle [284, 285]. Additionally, bacteria can produce surfactants, which are more effective than synthetic surfactants at removing hydrocarbons from the environment [276]. To aid the metagenomic identification of hydrocarbon-degrading bacteria, curated databases using the experimentally-validated hydrocarbon degradation genes have been generated. AromaDeg is grounded in the phylogenetic analysis of experimentally validated genes that are involved in the degradation of hydrocarbons [286]. In contrast, CANT-HYD broadly targets genes related to the degradation of hydrocarbons, not just aromatics, and is built as a hidden Markov model database to search for distant sequence similarity [287].

Bacteria are also able to metabolize heavy metals into less toxic products through biotransformation, bioleaching, and biomineralization [288, 289]. In addition, there are a diversity of ways that bacteria can interact with heavy metals to remove them from the environment. Bacteria can secrete extrapolymeric substances, metallothioneins, and siderophores that are able to coordinate and sequester heavy metals in their surrounding environment [271, 290, 291, 292]. Bacteria are also able to sequester heavy metals in the environment by uptaking large quantities into their cytoplasm, a process known as bioaccumulation [293]. Through these processes, bacteria are able to counter environmental pollution of lead, chromium, cadmium, and arsenic [294, 295, 296, 297, 298].

Given the hardiness of marine bacteria and their innate ability for bioremediation of hydrocarbons and heavy metals, they are prime candidates for bioengineering [187]. Some progress has been made in the area by introducing a metallothionein gene into a marine bacteria to induce heavy metal resistance [299]. However, much more work needs to go into understanding the genetics of these marine bacteria. In particular, identification of mobile genetic elements that can act as vectors for bioremediation genes would prove invaluable in efforts to bioengineer marine bacteria.

1.3.3 Conjugative elements

Type IV conjugative elements allow for the exchange of genetic information through cell-to-cell contact mediated by a pilus that bridges the two cells. DNA can also be exchanged through transformation, which is the uptake of DNA by a bacterium from its surrounding environment, and by transduction, which is phage-mediated DNA transfer. Conjugative elements can either be carried on an integrative and conjugative element (ICE) that inserts itself on the chromosomal sequence or on a plasmid, which exists as a separate, often circular, independently-replicating genomic element. Essential to the self-mobilization are three categories of proteins: relaxases, type IV coupling proteins (T4CP), and type IV secretion system (T4SS) proteins [300]. Relaxases catalyze a single stranded nicking reaction at the origin of transfer (*oriT*) sequence and unwind the DNA [300, 301]. Relaxases coordinate a divalent metal ion using a conserved histadine triad, as well as a conserved tyrosine, to catalyze the nicking reaction at the *oriT* [302, 303]. *oriT* sequences are difficult to predict computationally due to poor sequence conservation, so strategies that focus more on the DNA structure (i.e. looking for potential hairpin loops) are more successful [304]. Type IV coupling proteins transit the DNA-relaxase complex to the pore complex for transfer [305]. While T4CP are not required for pilus biogenesis [306], they are critical for efficient transfer of DNA sequences through conjugation [307]. The classical model of cojugation, the *Agrobacterium tumefaciens* pTi plasmid, contains twelve T4SS proteins VirB1 through VirB12 that form the pilus that transfers the relaxase-DNA complex [305]. VirB4 and VirB11 are the ATPases and are involved in the biogenesis of the pilus, though all gram-positive conjugative elements lack the homologue for VirB11 [305, 308]. Other proteins of the pilus, such as VirB8 which is missing a homologue in the *Escherichia coli* F plasmid, also are poorly conserved among conjugative systems [308]. Homologues of the VirB4 ATPase are critical for the active transport of DNA through the pilus and their phylogeny is generally similar to the phylogeny of the bacteria harbouring them [308].

In metagenomic analyses, conjugative elements can be difficult to analyze. As previously mentioned, one of the core elements of a conjugative element, the *oriT* sequence, is difficult to predict computationally due to low sequence conservation [304]. Furthermore, assembly of plasmid sequences, which are a common vector for conjugative elements, are near impossible to assemble from metagenomic data despite their relatively short sequence length [309]. Another factor that has been recently highlighted is the systematic exclusion of mobile genetic element, including plasmids and ICEs, from metagenomic bins, which are the core of many metagenomic analysis workflows [102]. Further refinement of bioinformatic analyses used in metagenomics is needed to properly capture these widely transferred genetic elements.

1.4 Scope and Objective of Thesis

At the start of my thesis research, metagenomic identification of plasmid and conjugative systems was still in early development with computational tools such as PlasFlow [310] only having been just released. I was involved with the sequencing assembly and analysis of a conjugative plasmid that could efficiently and selectively kill bacteria using a CRISPR system [311]. However, when conjugation was attempted with non-lab strains of bacteria, the conjugative plasmid had reduced conjugation efficiency and apparent sequence recombination that deactivated the system. With the eventual goal of using such a system to modulate the composition of the gut microbiota or to selectively kill intestinal pathogens such as *Campylobacter jejuni*, better vectors for CRISPR systems needed to be identified for colonizers of the gut. These early findings in my thesis were the motivation for Chapters 2 and 3 of my thesis. In Chapter 2, I developed a simple, yet effective, method for identifying type IV conjugative elements from metagenomic assemblies. I discovered that these systems were systematically excluded from metagenomic bins, which coincided with a finding published in the same time period stating that plasmids and mobile genetic elements are systematically excluded from metagenomic bins [102]. In Chapter 3, I applied the methodology of identifying these oft-excluded genetic elements to a metagenomic study of the association between the gut microbiota of mothers and spina bifida in their newborns. I revealed that a MAG of *Campylobacter hominis* is highly enriched in the microbiota of mothers who gave birth to infants and that there was a *Campylobacter hominis* conjugative element, not included in the metagenomic bin, that was also enriched. Though it does not appear to carry genes that are involved in disease pathogenesis, such an element could serve as a backbone to effectively deliver a CRISPR-killing system to target this pathogenic bacteria colonizing these mothers.

Also underdeveloped at the outset of my thesis research was third-generation sequencing techniques. I assisted in the completion, and validation, of a number of bacterial genomes from a complex metagenomic sample using third-generation sequencing early in my thesis [312]. However, our methods struggled to assemble complete sequences in more complex communities or when read length was less than optimal. In Chapter 4, I identified a circular and complete sequence of a manganese-oxidizing bacterium from a very complex environmental sample. While present in the other studied samples, the genome could not be resolved into a single, circularized sequence like in the first sample. To address this issue, I used a reference-guided assembly, using the first genome, and circularized an additional two genomes. Applying the methods developed in Chapter 2, I also identified conjugative plasmids from these communities belonging to this bacteria as well. In Chapter 5, I expanded on the reference-based assembly used in Chapter 4 by binning the non-circularized sequences from a metagenomic assembly and using the bins as the basis for a reference-based assembly. A similar approach, Jorg, has been recently published for short-read data [313]. With this methodology, additional circularized genomes and plasmids could be assembled from these complex communities. When paired with the conjugative element identification I developed in Chapter 2, this blueprint for metagenomic analysis maximized the amount of information that could be retrieved from the dataset.

Overall, these chapters present additional improvements to standard metagenomic analyses and apply these improved methodologies to areas of clinical and environmental interest. The studies provide a blueprint for more complete metagenomic analyses. By applying these improved methodologies in the future, it is my hope that additional links between the microbiota and human health can be discovered, novel vectors for microbiota engineering can be identified, and optimal strains for bioremediation can be found.

1.5 References

Bibliography

[1] Erika Check Hayden. Technology: The $1,000 genome. *Nature*, 507(7492):294–295, March 2014.

[2] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, April 1953.

[3] F. H. Crick. On protein synthesis. *Symposia of the Society for Experimental Biology*, 12:138–163, 1958.

[4] Clyde A. Hutchison. DNA sequencing: bench to bedside and beyond. *Nucleic Acids Research*, 35(18):6227–6237, 2007.

[5] Robert W. Holley, Jean Apgar, Susan H. Merrill, and Paul L. Zubkoff. NUCLEOTIDE AND OLIGONUCLEOTIDE COMPOSITIONS OF THE ALANINE-, VALINE-, AND TYROSINE-ACCEPTOR “SOLUBLE” RIBONUCLEIC ACIDS OF YEAST. *Journal of the American Chemical Society*, 83(23):4861–4862, December 1961.

[6] J. T. Madison and R. W. Holley. THE PRESENCE OF 5,6-DIHYDROURIDYLIC ACID IN YEAST ”SOLUBLE” RIBONUCLEIC ACID. *Biochemical and Biophysical Research Communications*, 18:153–157, January 1965.

[7] R. W. Holley, J. Apgar, G. A. Everett, J. T. Madison, M. Marquisee, S. H. Merrill, J. R. Penswick, and A. Zamir. STRUCTURE OF A RIBONUCLEIC ACID. *Science (New York, N.Y.)*, 147(3664):1462–1465, March 1965.

[8] F. Sanger, G. G. Brownlee, and B. G. Barrell. A two-dimensional fractionation procedure for radioactive nucleotides. *Journal of Molecular Biology*, 13(2):373–398, September 1965.

[9] G. G. Brownlee and F. Sanger. Nucleotide sequences from the low molecular weight ribosomal RNA of Escherichia coli. *Journal of Molecular Biology*, 23(3):337–353, February 1967.

[10] S. Cory, K. A. Marcker, S. K. Dube, and B. F. Clark. Primary structure of a methionine

transfer RNA from Escherichia coli. *Nature*, 220(5171):1039–1040, December 1968.

[11] W. Min Jou, G. Haegeman, M. Ysebaert, and W. Fiers. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature*, 237(5350):82–88, May 1972.

[12] W. Fiers, R. Contreras, F. Duerinck, G. Haegeman, D. Iserentant, J. Merregaert, W. Min Jou, F. Molemans, A. Raeymaekers, A. Van den Berghe, G. Volckaert, and M. Ysebaert. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature*, 260(5551):500–507, April 1976.

[13] R. Wu and A. D. Kaiser. Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. *Journal of Molecular Biology*, 35(3):523–537, August 1968.

[14] R. Wu. Nucleotide sequence analysis of DNA. I. Partial sequence of the cohesive ends of bacteriophage lambda and 186 DNA. *Journal of Molecular Biology*, 51(3):501–521, August 1970.

[15] F. Sanger, J. E. Donelson, A. R. Coulson, H. Kössel, and D. Fischer. Use of DNA polymerase I primed by a synthetic oligonucleotide to determine a nucleotide sequence in phage fl DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 70(4):1209–1213, April 1973.

[16] R. Padmanabhan, E. Jay, and R. Wu. Chemical synthesis of a primer and its use in the sequence analysis of the lysozyme gene of bacteriophage T4. *Proceedings of the National Academy of Sciences of the United States of America*, 71(6):2510–2514, June 1974.

[17] R. Padmanabhan and R. Wu. Nucleotide sequence analysis of DNA. IX. Use of oligonucleotides of defined sequence as primers in DNA sequence analysis. *Biochemical and Biophysical Research Communications*, 48(5):1295–1302, September 1972.

[18] F. Sanger and A. R. Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94(3):441–448, May 1975.

[19] A. M. Maxam and W. Gilbert. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2):560–564, February 1977.

[20] F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, C. A. Fiddes, C. A. Hutchison, P. M. Slocombe, and M. Smith. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, 265(5596):687–695, February 1977.

[21] W. Ansorge, B. S. Sproat, J. Stegemann, and C. Schwager. A non-radioactive automated method for DNA sequence determination. *Journal of Biochemical and Biophysical Methods*, 13(6):315–323, December 1986.

[22] W. Ansorge, B. Sproat, J. Stegemann, C. Schwager, and M. Zenke. Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis. *Nucleic Acids Research*, 15(11):4593–4602, June 1987.

[23] J. M. Prober, G. L. Trainor, R. J. Dam, F. W. Hobbs, C. W. Robertson, R. J. Zagursky, A. J. Cocuzza, M. A. Jensen, and K. Baumeister. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science (New York, N.Y.)*, 238(4825):336–341, October 1987.

[24] Hideki Kambara, Tetsuo Nishikawa, Yoshiko Katayama, and Tomoaki Yamaguchi. Optimization of Parameters in a DNA Sequenator Using Fluorescence Detection. *Nature Biotechnology*, 6(7):816–821, July 1988.

[25] H. Swerdlow and R. Gesteland. Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Research*, 18(6):1415–1419, March 1990.

[26] J. A. Luckey, H. Drossman, A. J. Kostichka, D. A. Mead, J. D'Cunha, T. B. Norris, and L. M. Smith. High speed DNA sequencing by capillary electrophoresis. *Nucleic Acids Research*, 18(15):4417–4421, August 1990.

[27] L. M. Smith, S. Fung, M. W. Hunkapiller, T. J. Hunkapiller, and L. E. Hood. The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. *Nucleic Acids Research*, 13(7):2399–2412, April 1985.

[28] R. K. Saiki, S. Scharf, F. Faloona, K. B. Mullis, G. T. Horn, H. A. Erlich, and N. Arnheim. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for

diagnosis of sickle cell anemia. *Science (New York, N.Y.)*, 230(4732):1350–1354, December 1985.

[29] R. K. Saiki, D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuchi, G. T. Horn, K. B. Mullis, and H. A. Erlich. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science (New York, N.Y.)*, 239(4839):487–491, January 1988.

[30] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel,

A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowki, and International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001.

[31] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman,

M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507):1304–1351, February 2001.

[32] P. Nyrén and A. Lundin. Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. *Analytical Biochemistry*, 151(2):504–509, December 1985.

[33] E. D. Hyman. A new method of sequencing DNA. *Analytical Biochemistry*, 174(2):423–436, November 1988.

[34] P. Nyrén. Enzymatic method for continuous monitoring of DNA polymerase activity. *Analytical Biochemistry*, 167(2):235–238, December 1987.

[35] M. Ronaghi, S. Karamohamed, B. Pettersson, M. Uhlén, and P. Nyrén. Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry*, 242(1):84–89, November 1996.

[36] M. Ronaghi, M. Uhlén, and P. Nyrén. A sequencing method based on real-time pyrophosphate. *Science (New York, N.Y.)*, 281(5375):363, 365, July 1998.

[37] Samuel Levy, Granger Sutton, Pauline C. Ng, Lars Feuk, Aaron L. Halpern, Brian P. Walenz, Nelson Axelrod, Jiaqi Huang, Ewen F. Kirkness, Gennady Denisov, Yuan Lin, Jeffrey R. MacDonald, Andy Wing Chun Pang, Mary Shago, Timothy B. Stockwell, Alexia Tsiamouri, Vineet Bafna, Vikas Bansal, Saul A. Kravitz, Dana A. Busam, Karen Y. Beeson, Tina C. McIntosh, Karin A. Remington, Josep F. Abril, John Gill, Jon Borman, Yu-Hui Rogers, Marvin E. Frazier, Stephen W. Scherer, Robert L. Strausberg, and J. Craig Venter. The diploid genome sequence of an individual human. *PLoS biology*, 5(10):e254, September 2007.

[38] Sten Anslan, Vladimir Mikryukov, Kestutis Armolaitis, Jelena Ankuda, Dagnija Lazdina, Kristaps Makovskis, Lars Vesterdal, Inger Kappel Schmidt, and Leho Tedersoo. Highly comparable metabarcoding results from MGI-Tech and Illumina sequencing platforms. *PeerJ*, 9:e12254, 2021.

[39] Ido Braslavsky, Benedict Hebert, Emil Kartalov, and Stephen R. Quake. Sequence information can be obtained from single DNA molecules. *Proceedings of the National Academy of Sciences of the United States of America*, 100(7):3960–3964, April 2003.

[40] J. J. Kasianowicz, E. Brandin, D. Branton, and D. W. Deamer. Characterization of individual

polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences of the United States of America*, 93(24):13770–13773, November 1996.

[41] Joshua Quick, Nicholas J. Loman, Sophie Duraffour, Jared T. Simpson, Ettore Severi, Lauren Cowley, Joseph Akoi Bore, Raymond Koundouno, Gytis Dudas, Amy Mikhail, Nobila Ouédraogo, Babak Afrough, Amadou Bah, Jonathan Hj Baum, Beate Becker-Ziaja, Jan-Peter Boettcher, Mar Cabeza-Cabrerizo, Alvaro Camino-Sanchez, Lisa L. Carter, Juiliane Doerrbecker, Theresa Enkirch, Isabel Graciela García Dorival, Nicole Hetzelt, Julia Hinzmann, Tobias Holm, Liana Eleni Kafetzopoulou, Michel Koropogui, Abigail Kosgey, Eeva Kuisma, Christopher H. Logue, Antonio Mazzarelli, Sarah Meisel, Marc Mertens, Janine Michel, Didier Ngabo, Katja Nitzsche, Elisa Pallash, Livia Victoria Patrono, Jasmine Portmann, Johanna Gabriella Repits, Natasha Yasmin Rickett, Andrea Sachse, Katrin Singethan, Inês Vitoriano, Rahel L. Yemanaberhan, Elsa G. Zekeng, Racine Trina, Alexander Bello, Amadou Alpha Sall, Ousmane Faye, Oumar Faye, N'Faly Magassouba, Cecelia V. Williams, Victoria Amburgey, Linda Winona, Emily Davis, Jon Gerlach, Franck Washington, Vanessa Monteil, Marine Jourdain, Marion Bererd, Alimou Camara, Hermann Somlare, Abdoulaye Camara, Marianne Gerard, Guillaume Bado, Bernard Baillet, Déborah Delaune, Koumpingnin Yacouba Nebie, Abdoulaye Diarra, Yacouba Savane, Raymond Bernard Pallawo, Giovanna Jaramillo Gutierrez, Natacha Milhano, Isabelle Roger, Christopher J. Williams, Facinet Yattara, Kuiama Lewandowski, Jamie Taylor, Philip Rachwal, Daniel Turner, Georgios Pollakis, Julian A. Hiscox, David A. Matthews, Matthew K. O'Shea, Andrew McD Johnston, Duncan Wilson, Emma Hutley, Erasmus Smit, Antonino Di Caro, Roman Woelfel, Kilian Stoecker, Erna Fleischmann, Martin Gabriel, Simon A. Weller, Lamine Koivogui, Boubacar Diallo, Sakoba Keita, Andrew Rambaut, Pierre Formenty, Stephan Gunther, and Miles W. Carroll. Real-time, portable genome sequencing for Ebola surveillance. *Nature*, 530(7589):228–232, February 2016.

[42] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick. Whole-genome random sequenc-

ing and assembly of Haemophilus influenzae Rd. *Science (New York, N.Y.)*, 269(5223):496–512, July 1995.

[43] Michael S. Rappé and Stephen J. Giovannoni. The Uncultured Microbial Majority. *Annual Review of Microbiology*, 57(1):369–394, October 2003.

[44] Jo Handelsman, Michelle R. Rondon, Sean F. Brady, Jon Clardy, and Robert M. Goodman. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology*, 5(10):R245–R249, October 1998.

[45] J. Michael Janda and Sharon L. Abbott. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of Clinical Microbiology*, 45(9):2761–2764, September 2007.

[46] G. E. Fox, L. J. Magrum, W. E. Balch, R. S. Wolfe, and C. R. Woese. Classification of methanogenic bacteria by 16S ribosomal RNA characterization. *Proceedings of the National Academy of Sciences of the United States of America*, 74(10):4537–4541, October 1977.

[47] R. Gupta, J. M. Lanter, and C. R. Woese. Sequence of the 16S Ribosomal RNA from Halobacterium volcanii, an Archaebacterium. *Science (New York, N.Y.)*, 221(4611):656–659, August 1983.

[48] Gillian C. Baker, Shabarni Gaffar, Don A. Cowan, and Adrian R. Suharto. Bacterial community analysis of Indonesian hot springs. *FEMS Microbiology Letters*, 200(1):103–109, June 2001.

[49] Lorraine G. van Waasbergen, David L. Balkwill, Fiona H. Crocker, Bruce N. Bjornstad, and Robert V. Miller. Genetic Diversity among *Arthrobacter* Species Collected across a Heterogeneous Series of Terrestrial Deep-Subsurface Sediments as Determined on the Basis of 16S rRNA and *recA* Gene Sequences. *Applied and Environmental Microbiology*, 66(8):3454–3463, August 2000.

[50] T.R Whitehead and M.A Cotta. Phylogenetic diversity of methanogenic archaea in swine waste storage pits. *FEMS Microbiology Letters*, 179(2):223–226, October 1999.

[51] Benjamin J Callahan, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A

Johnson, and Susan P Holmes. DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, 13(7):581–583, July 2016.

[52] P. P. Bosshard, R. Zbinden, S. Abels, B. Böddinghaus, M. Altwegg, and E. C. Böttger. 16S rRNA gene sequencing versus the API 20 NE system and the VITEK 2 ID-GNB card for identification of nonfermenting Gram-negative bacteria in the clinical laboratory. *Journal of Clinical Microbiology*, 44(4):1359–1366, April 2006.

[53] S. Mignard and J. P. Flandrois. 16S rRNA sequencing in routine bacterial identification: a 30-month experiment. *Journal of Microbiological Methods*, 67(3):574–581, December 2006.

[54] R. A. Clayton, G. Sutton, P. S. Hinkle, C. Bult, and C. Fields. Intraspecific variation in small-subunit rRNA sequences in GenBank: why single sequences may not adequately represent prokaryotic taxa. *International Journal of Systematic Bacteriology*, 45(3):595–599, July 1995.

[55] Gavin M. Douglas, Vincent J. Maffei, Jesse R. Zaneveld, Svetlana N. Yurgel, James R. Brown, Christopher M. Taylor, Curtis Huttenhower, and Morgan G. I. Langille. PICRUSt2 for prediction of metagenome functions. *Nature Biotechnology*, 38(6):685–688, June 2020.

[56] S. Anderson. Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Research*, 9(13):3015–3027, July 1981.

[57] Victor Kunin, Alex Copeland, Alla Lapidus, Konstantinos Mavromatis, and Philip Hugenholtz. A Bioinformatician's Guide to Metagenomics. *Microbiology and Molecular Biology Reviews*, 72(4):557–578, December 2008.

[58] Alexander Sczyrba, Peter Hofmann, Peter Belmann, David Koslicki, Stefan Janssen, Johannes Dröge, Ivan Gregor, Stephan Majda, Jessika Fiedler, Eik Dahms, Andreas Bremges, Adrian Fritz, Ruben Garrido-Oter, Tue Sparholt Jørgensen, Nicole Shapiro, Philip D Blood, Alexey Gurevich, Yang Bai, Dmitrij Turaev, Matthew Z DeMaere, Rayan Chikhi, Niranjan Nagarajan, Christopher Quince, Fernando Meyer, Monika Balvočiūtė, Lars Hestbjerg Hansen, Søren J Sørensen, Burton K H Chia, Bertrand Denis, Jeff L Froula, Zhong

Wang, Robert Egan, Dongwan Don Kang, Jeffrey J Cook, Charles Deltel, Michael Beckstette, Claire Lemaitre, Pierre Peterlongo, Guillaume Rizk, Dominique Lavenier, Yu-Wei Wu, Steven W Singer, Chirag Jain, Marc Strous, Heiner Klingenberg, Peter Meinicke, Michael D Barton, Thomas Lingner, Hsin-Hung Lin, Yu-Chieh Liao, Genivaldo Gueiros Z Silva, Daniel A Cuevas, Robert A Edwards, Surya Saha, Vitor C Piro, Bernhard Y Renard, Mihai Pop, Hans-Peter Klenk, Markus Göker, Nikos C Kyrpides, Tanja Woyke, Julia A Vorholt, Paul Schulze-Lefert, Edward M Rubin, Aaron E Darling, Thomas Rattei, and Alice C McHardy. Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nature Methods*, 14(11):1063–1071, November 2017.

[59] Samuel Martin, Darren Heavens, Yuxuan Lan, Samuel Horsfield, Matthew D. Clark, and Richard M. Leggett. Nanopore adaptive sampling: a tool for enrichment of low abundance species in metagenomic samples. *Genome Biology*, 23(1):11, December 2022.

[60] Z. Li, Y. Chen, D. Mu, J. Yuan, Y. Shi, H. Zhang, J. Gan, N. Li, X. Hu, B. Liu, B. Yang, and W. Fan. Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Briefings in Functional Genomics*, 11(1):25–37, January 2012.

[61] Jason R. Miller, Sergey Koren, and Granger Sutton. Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6):315–327, June 2010.

[62] Paul Flicek and Ewan Birney. Sense from sequence reads: methods for alignment and assembly. *Nature Methods*, 6(11 Suppl):S6–S12, November 2009.

[63] Ramana M. Idury and Michael S. Waterman. A New Algorithm for DNA Sequence Assembly. *Journal of Computational Biology*, 2(2):291–306, January 1995.

[64] P. A. Pevzner, H. Tang, and M. S. Waterman. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America*, 98(17):9748–9753, August 2001.

[65] Daniel R. Zerbino, Gayle K. McEwen, Elliott H. Margulies, and Ewan Birney. Pebble and rock band: heuristic resolution of repeats and scaffolding in the velvet short-read de novo

assembler. *PloS One*, 4(12):e8407, December 2009.

[66] Daniel R. Zerbino and Ewan Birney. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5):821–829, May 2008.

[67] Toshiaki Namiki, Tsuyoshi Hachiya, Hideaki Tanaka, and Yasubumi Sakakibara. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research*, 40(20):e155, November 2012.

[68] Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics (Oxford, England)*, 31(10):1674–1676, May 2015.

[69] Yu Peng, Henry C. M. Leung, S. M. Yiu, and Francis Y. L. Chin. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics (Oxford, England)*, 28(11):1420–1428, June 2012.

[70] Sergey Nurk, Dmitry Meleshko, Anton Korobeynikov, and Pavel A. Pevzner. metaSPAdes: a new versatile metagenomic assembler. *Genome Research*, 27(5):824–834, May 2017.

[71] Ziye Wang, Ying Wang, Jed A. Fuhrman, Fengzhu Sun, and Shanfeng Zhu. Assessment of metagenomic assemblers based on hybrid reads of real and simulated metagenomic sequences. *Briefings in Bioinformatics*, 21(3):777–790, May 2020.

[72] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, Sergey I. Nikolenko, Son Pham, Andrey D. Prjibelski, Alexey V. Pyshkin, Alexander V. Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A. Alekseyev, and Pavel A. Pevzner. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, 19(5):455–477, May 2012.

[73] Andrey D. Prjibelski, Irina Vasilinetc, Anton Bankevich, Alexey Gurevich, Tatiana Krivosheeva, Sergey Nurk, Son Pham, Anton Korobeynikov, Alla Lapidus, and Pavel A. Pevzner. ExSPAnder: a universal repeat resolver for DNA fragment assembly. *Bioinformat-*

ics (Oxford, England), 30(12):i293–301, June 2014.

[74] Irina Vasilinetc, Andrey D. Prjibelski, Alexey Gurevich, Anton Korobeynikov, and Pavel A. Pevzner. Assembling short reads from jumping libraries with large insert sizes. *Bioinformatics (Oxford, England)*, 31(20):3262–3268, October 2015.

[75] Dmitry Antipov, Anton Korobeynikov, Jeffrey S. McLean, and Pavel A. Pevzner. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics (Oxford, England)*, 32(7):1009–1015, April 2016.

[76] Sergey Koren, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, 27(5):722–736, May 2017.

[77] Konstantin Berlin, Sergey Koren, Chen-Shan Chin, James P. Drake, Jane M. Landolin, and Adam M. Phillippy. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature Biotechnology*, 33(6):623–630, June 2015.

[78] E. W. Myers, G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. Reinert, K. A. Remington, E. L. Anson, R. A. Bolanos, H. H. Chou, C. M. Jordan, A. L. Halpern, S. Lonardi, E. M. Beasley, R. C. Brandon, L. Chen, P. J. Dunn, Z. Lai, Y. Liang, D. R. Nusskern, M. Zhan, Q. Zhang, X. Zheng, G. M. Rubin, M. D. Adams, and J. C. Venter. A whole-genome assembly of Drosophila. *Science (New York, N.Y.)*, 287(5461):2196–2204, March 2000.

[79] Mikhail Kolmogorov, Derek M. Bickhart, Bahar Behsaz, Alexey Gurevich, Mikhail Rayko, Sung Bong Shin, Kristen Kuhn, Jeffrey Yuan, Evgeny Polevikov, Timothy P. L. Smith, and Pavel A. Pevzner. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods*, 17(11):1103–1110, November 2020.

[80] Mikhail Kolmogorov, Jeffrey Yuan, Yu Lin, and Pavel A. Pevzner. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, 37(5):540–546, May 2019.

[81] Adriel Latorre-Pérez, Pascual Villalba-Bermell, Javier Pascual, and Cristina Vilanova. Assembly methods for nanopore-based metagenomic sequencing: a comparative study. *Scien-*

tific Reports, 10(1):13588, August 2020.

[82] Ryan R. Wick, Louise M. Judd, Claire L. Gorrie, and Kathryn E. Holt. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS computational biology*, 13(6):e1005595, June 2017.

[83] Zhao Chen, David L. Erickson, and Jianghong Meng. Polishing the Oxford Nanopore long-read assemblies of bacterial pathogens with Illumina short reads to improve genomic analyses. *Genomics*, 113(3):1366–1377, May 2021.

[84] Medaka, April 2022.

[85] Kishwar Shafin, Trevor Pesout, Pi-Chuan Chang, Maria Nattestad, Alexey Kolesnikov, Sidharth Goel, Gunjan Baid, Mikhail Kolmogorov, Jordan M. Eizenga, Karen H. Miga, Paolo Carnevali, Miten Jain, Andrew Carroll, and Benedict Paten. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nature Methods*, 18(11):1322–1332, November 2021.

[86] Yao-Ting Huang, Po-Yu Liu, and Pei-Wen Shih. Homopolish: a method for the removal of systematic errors in nanopore sequencing by homologous polishing. *Genome Biology*, 22(1):95, March 2021.

[87] Jin Young Lee, Minyoung Kong, Jinjoo Oh, JinSoo Lim, Sung Hee Chung, Jung-Min Kim, Jae-Seok Kim, Ki-Hwan Kim, Jae-Chan Yoo, and Woori Kwak. Comparative evaluation of Nanopore polishing tools for microbial genome assembly and polishing strategies for downstream analysis. *Scientific Reports*, 11(1):20740, October 2021.

[88] Brian Cleary, Ilana Lauren Brito, Katherine Huang, Dirk Gevers, Terrance Shea, Sarah Young, and Eric J. Alm. Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nature Biotechnology*, 33(10):1053–1060, October 2015.

[89] Samuele Girotto, Cinzia Pizzi, and Matteo Comin. MetaProb: accurate metagenomic reads binning based on probabilistic sequence signatures. *Bioinformatics (Oxford, England)*, 32(17):i567–i575, September 2016.

[90] Rachid Ounit, Steve Wanamaker, Timothy J. Close, and Stefano Lonardi. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC genomics*, 16:236, March 2015.

[91] L. Schaeffer, H. Pimentel, N. Bray, P. Melsted, and L. Pachter. Pseudoalignment for metagenomic read assignment. *Bioinformatics (Oxford, England)*, 33(14):2082–2088, July 2017.

[92] Kévin Vervier, Pierre Mahé, Maud Tournoud, Jean-Baptiste Veyrieras, and Jean-Philippe Vert. Large-scale machine learning for metagenomics sequence classification. *Bioinformatics (Oxford, England)*, 32(7):1023–1032, April 2016.

[93] Yi Wang, Henry C. M. Leung, S. M. Yiu, and Francis Y. L. Chin. MetaCluster 4.0: a novel binning algorithm for NGS reads and huge number of species. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, 19(2):241–249, February 2012.

[94] Vimac Nolla-Ardèvol, Miriam Peces, Marc Strous, and Halina E. Tegetmeyer. Metagenome from a Spirulina digesting biogas reactor: analysis via binning of contigs and classification of short reads. *BMC microbiology*, 15:277, December 2015.

[95] Johannes Alneberg, Brynjar Smári Bjarnason, Ino de Bruijn, Melanie Schirmer, Joshua Quick, Umer Z Ijaz, Leo Lahti, Nicholas J Loman, Anders F Andersson, and Christopher Quince. Binning metagenomic contigs by coverage and composition. *Nature Methods*, 11(11):1144–1146, November 2014.

[96] Dongwan D. Kang, Feng Li, Edward Kirton, Ashleigh Thomas, Rob Egan, Hong An, and Zhong Wang. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, 7:e7359, 2019.

[97] Yi Yue, Hao Huang, Zhao Qi, Hui-Min Dou, Xin-Yi Liu, Tian-Fei Han, Yue Chen, Xiang-Jun Song, You-Hua Zhang, and Jian Tu. Evaluating metagenomics tools for genome binning with real metagenomic datasets and CAMI datasets. *BMC Bioinformatics*, 21(1):334, December 2020.

[98] Vijini Mallawaarachchi, Anuradha Wickramarachchi, and Yu Lin. GraphBin: refined bin-

ning of metagenomic contigs using assembly graphs. *Bioinformatics*, 36(11):3307–3313, June 2020.

[99] Christian M. K. Sieber, Alexander J. Probst, Allison Sharrar, Brian C. Thomas, Matthias Hess, Susannah G. Tringe, and Jillian F. Banfield. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology*, 3(7):836–843, July 2018.

[100] Naseer Sangwan, Fangfang Xia, and Jack A. Gilbert. Recovering complete and draft population genomes from metagenome datasets. *Microbiome*, 4:8, March 2016.

[101] Donovan H. Parks, Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, and Gene W. Tyson. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 25(7):1043–1055, July 2015.

[102] Finlay Maguire, Baofeng Jia, Kristen L. Gray, Wing Yin Venus Lau, Robert G. Beiko, and Fiona S. L. Brinkman. Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic Islands. *Microbial Genomics*, 6(10), October 2020.

[103] Daniel H. Huson, Suparna Mitra, Hans-Joachim Ruscheweyh, Nico Weber, and Stephan C. Schuster. Integrative analysis of environmental sequences using MEGAN4. *Genome Research*, 21(9):1552–1560, September 2011.

[104] David R. Kelley, Bo Liu, Arthur L. Delcher, Mihai Pop, and Steven L. Salzberg. Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic Acids Research*, 40(1):e9, January 2012.

[105] Doug Hyatt, Gwo-Liang Chen, Philip F. Locascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, 11:119, March 2010.

[106] A. L. Delcher, D. Harmon, S. Kasif, O. White, and S. L. Salzberg. Improved microbial gene identification with GLIMMER. *Nucleic Acids Research*, 27(23):4636–4641, December 1999.

[107] Juliane C. Dohm, Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, 36(16):e105, September 2008.

[108] J. Shine and L. Dalgarno. Terminal-sequence analysis of bacterial ribosomal RNA. Correlation between the 3'-terminal-polypyrimidine sequence of 16-S RNA and translational specificity of the ribosome. *European Journal of Biochemistry*, 57(1):221–230, September 1975.

[109] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, October 1990.

[110] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, September 1997.

[111] E. Michael Gertz, Yi-Kuo Yu, Richa Agarwala, Alejandro A. Schäffer, and Stephen F. Altschul. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC biology*, 4:41, December 2006.

[112] Benjamin Buchfink, Chao Xie, and Daniel H. Huson. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1):59–60, January 2015.

[113] S. R. Eddy. Profile hidden Markov models. *Bioinformatics (Oxford, England)*, 14(9):755–763, 1998.

[114] C. Barrett, R. Hughey, and K. Karplus. Scoring hidden Markov models. *Computer applications in the biosciences: CABIOS*, 13(2):191–199, April 1997.

[115] Sean R. Eddy. Accelerated Profile HMM Searches. *PLoS computational biology*, 7(10):e1002195, October 2011.

[116] Emmanuel Boutet, Damien Lieberherr, Michael Tognolli, Michel Schneider, and Amos Bairoch. UniProtKB/Swiss-Prot. *Methods in Molecular Biology (Clifton, N.J.)*, 406:89–112, 2007.

[117] Baris E. Suzek, Yuqi Wang, Hongzhan Huang, Peter B. McGarvey, Cathy H. Wu, and UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics (Oxford, England)*, 31(6):926–932, March 2015.

[118] Baris E. Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H. Wu. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics (Oxford, England)*, 23(10):1282–1288, May 2007.

[119] Amos Bairoch, Rolf Apweiler, Cathy H. Wu, Winona C. Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J. Martin, Darren A. Natale, Claire O'Donovan, Nicole Redaschi, and Lai-Su L. Yeh. The Universal Protein Resource (UniProt). *Nucleic Acids Research*, 33(Database issue):D154–159, January 2005.

[120] Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Research*, 43(Database issue):D1049–1056, January 2015.

[121] Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A. Salazar, Erik L. L. Sonnhammer, Silvio C. E. Tosatto, Lisanna Paladin, Shriya Raj, Lorna J. Richardson, Robert D. Finn, and Alex Bateman. Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1):D412–D419, January 2021.

[122] Robert D. Finn, Jaina Mistry, John Tate, Penny Coggill, Andreas Heger, Joanne E. Pollington, O. Luke Gavin, Prasad Gunasekaran, Goran Ceric, Kristoffer Forslund, Liisa Holm, Erik L. L. Sonnhammer, Sean R. Eddy, and Alex Bateman. The Pfam protein families database. *Nucleic Acids Research*, 38(Database issue):D211–222, January 2010.

[123] Marco Punta, Penny C. Coggill, Ruth Y. Eberhardt, Jaina Mistry, John Tate, Chris Boursnell, Ningze Pang, Kristoffer Forslund, Goran Ceric, Jody Clements, Andreas Heger, Liisa Holm, Erik L. L. Sonnhammer, Sean R. Eddy, Alex Bateman, and Robert D. Finn. The Pfam protein families database. *Nucleic Acids Research*, 40(Database issue):D290–301, January 2012.

[124] E. L. Sonnhammer, S. R. Eddy, and R. Durbin. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, 28(3):405–420, July 1997.

[125] Michael Y. Galperin, Yuri I. Wolf, Kira S. Makarova, Roberto Vera Alvarez, David Landsman, and Eugene V. Koonin. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Research*, 49(D1):D274–D281, January 2021.

[126] R. L. Tatusov, E. V. Koonin, and D. J. Lipman. A genomic perspective on protein families. *Science (New York, N.Y.)*, 278(5338):631–637, October 1997.

[127] R. L. Tatusov, M. Y. Galperin, D. A. Natale, and E. V. Koonin. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, 28(1):33–36, January 2000.

[128] Michael Y. Galperin, Kira S. Makarova, Yuri I. Wolf, and Eugene V. Koonin. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Research*, 43(Database issue):D261–269, January 2015.

[129] Matthias Blum, Hsin-Yu Chang, Sara Chuguransky, Tiago Grego, Swaathi Kandasaamy, Alex Mitchell, Gift Nuka, Typhaine Paysan-Lafosse, Matloob Qureshi, Shriya Raj, Lorna Richardson, Gustavo A. Salazar, Lowri Williams, Peer Bork, Alan Bridge, Julian Gough, Daniel H. Haft, Ivica Letunic, Aron Marchler-Bauer, Huaiyu Mi, Darren A. Natale, Marco Necci, Christine A. Orengo, Arun P. Pandurangan, Catherine Rivoire, Christian J. A. Sigrist, Ian Sillitoe, Narmada Thanki, Paul D. Thomas, Silvio C. E. Tosatto, Cathy H. Wu, Alex Bateman, and Robert D. Finn. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research*, 49(D1):D344–D354, January 2021.

[130] Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1):D457–462, January 2016.

[131] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Re-*

search, 45(D1):D353–D361, January 2017.

[132] M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, January 2000.

[133] Nuala A. O'Leary, Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, Alexander Astashyn, Azat Badretdin, Yiming Bao, Olga Blinkova, Vyacheslav Brover, Vyacheslav Chetvernin, Jinna Choi, Eric Cox, Olga Ermolaeva, Catherine M. Farrell, Tamara Goldfarb, Tripti Gupta, Daniel Haft, Eneida Hatcher, Wratko Hlavina, Vinita S. Joardar, Vamsi K. Kodali, Wenjun Li, Donna Maglott, Patrick Masterson, Kelly M. McGarvey, Michael R. Murphy, Kathleen O'Neill, Shashikant Pujar, Sanjida H. Rangwala, Daniel Rausch, Lillian D. Riddick, Conrad Schoch, Andrei Shkeda, Susan S. Storz, Hanzhen Sun, Francoise Thibaud-Nissen, Igor Tolstoy, Raymond E. Tully, Anjana R. Vatsan, Craig Wallin, David Webb, Wendy Wu, Melissa J. Landrum, Avi Kimchi, Tatiana Tatusova, Michael DiCuccio, Paul Kitts, Terence D. Murphy, and Kim D. Pruitt. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–745, January 2016.

[134] Dennis A. Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, James Ostell, Kim D. Pruitt, and Eric W. Sayers. GenBank. *Nucleic Acids Research*, 46(D1):D41–D47, January 2018.

[135] Minoru Kanehisa, Yoko Sato, and Kanae Morishima. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *Journal of Molecular Biology*, 428(4):726–731, February 2016.

[136] Shuji Suzuki, Masanori Kakuta, Takashi Ishida, and Yutaka Akiyama. GHOSTX: an improved sequence homology search algorithm using a query suffix array and a database suffix array. *PloS One*, 9(8):e103833, 2014.

[137] A. Murat Eren, Özcan C. Esen, Christopher Quince, Joseph H. Vineis, Hilary G. Morrison, Mitchell L. Sogin, and Tom O. Delmont. Anvi'o: an advanced analysis and visualization

platform for ‘omics data. *PeerJ*, 3:e1319, October 2015.

[138] Karin Lagesen, Peter Hallin, Einar Andreas Rødland, Hans-Henrik Staerfeldt, Torbjørn Rognes, and David W. Ussery. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, 35(9):3100–3108, 2007.

[139] Torsten Seemann. Barrnap, April 2022.

[140] Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41(Database issue):D590–596, January 2013.

[141] Jethro S. Johnson, Daniel J. Spakowicz, Bo-Young Hong, Lauren M. Petersen, Patrick Demkowicz, Lei Chen, Shana R. Leopold, Blake M. Hanson, Hanako O. Agresta, Mark Gerstein, Erica Sodergren, and George M. Weinstock. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications*, 10(1):5029, December 2019.

[142] Nicola Segata and Curtis Huttenhower. Toward an efficient method of identifying core genes for evolutionary and functional microbial phylogenies. *PloS One*, 6(9):e24704, 2011.

[143] Pierre-Alain Chaumeil, Aaron J. Mussig, Philip Hugenholtz, and Donovan H. Parks. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics (Oxford, England)*, page btz848, November 2019.

[144] Paul A. Kitts, Deanna M. Church, Françoise Thibaud-Nissen, Jinna Choi, Vichet Hem, Victor Sapojnikov, Robert G. Smith, Tatiana Tatusova, Charlie Xiang, Andrey Zherikov, Michael DiCuccio, Terence D. Murphy, Kim D. Pruitt, and Avi Kimchi. Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Research*, 44(D1):D73–80, January 2016.

[145] Frederick A. Matsen, Robin B. Kodner, and E. Virginia Armbrust. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC bioinformatics*, 11:538, October 2010.

[146] Donovan H Parks, Maria Chuvochina, David W Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil, and Philip Hugenholtz. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology*, 36(10):996–1004, November 2018.

[147] Francesco Beghini, Lauren J McIver, Aitor Blanco-Míguez, Leonard Dubois, Francesco Asnicar, Sagun Maharjan, Ana Mailyan, Paolo Manghi, Matthias Scholz, Andrew Maltez Thomas, Mireia Valles-Colomer, George Weingart, Yancong Zhang, Moreno Zolfo, Curtis Huttenhower, Eric A Franzosa, and Nicola Segata. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife*, 10:e65088, May 2021.

[148] F. A. Bastiaan von Meijenfeldt, Ksenia Arkhipova, Diego D. Cambuy, Felipe H. Coutinho, and Bas E. Dutilh. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biology*, 20(1):217, December 2019.

[149] M. Mirdita, M. Steinegger, F. Breitwieser, J. Söding, and E. Levy Karin. Fast and sensitive taxonomic assignment to metagenomic contigs. *Bioinformatics (Oxford, England)*, page btab184, March 2021.

[150] Nicola Segata, Levi Waldron, Annalisa Ballarini, Vagheesh Narasimhan, Olivier Jousson, and Curtis Huttenhower. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, 9(8):811–814, June 2012.

[151] Daniel H. Huson, Benjamin Albrecht, Caner Bağcı, Irina Bessarab, Anna Górska, Dino Jolic, and Rohan B. H. Williams. MEGAN-LR: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. *Biology Direct*, 13(1):6, April 2018.

[152] Caner Bağcı, Sina Beier, Anna Górska, and Daniel H. Huson. Introduction to the Analysis of Environmental Sequences: Metagenomics with MEGAN. *Methods in Molecular Biology (Clifton, N.J.)*, 1910:591–604, 2019.

[153] Sergey L. Sheetlin, Yonil Park, Martin C. Frith, and John L. Spouge. Frameshift align-

ment: statistics and post-genomic applications. *Bioinformatics*, 30(24):3575–3582, December 2014.

[154] Szymon M. Kiełbasa, Raymond Wan, Kengo Sato, Paul Horton, and Martin C. Frith. Adaptive seeds tame genomic sequence comparison. *Genome Research*, 21(3):487–493, March 2011.

[155] Derrick E. Wood, Jennifer Lu, and Ben Langmead. Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20(1):257, December 2019.

[156] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics (Oxford, England)*, 34(18):3094–3100, September 2018.

[157] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, April 2012.

[158] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–2079, August 2009.

[159] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1):139–140, January 2010.

[160] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014.

[161] Gregory B. Gloor, Jean M. Macklaim, Vera Pawlowsky-Glahn, and Juan J. Egozcue. Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology*, 8:2224, November 2017.

[162] Thomas P. Quinn, Tamsyn M. Crowley, and Mark F. Richardson. Benchmarking differential expression analysis tools for RNA-Seq: normalization-based vs. log-ratio transformation-based methods. *BMC bioinformatics*, 19(1):274, July 2018.

[163] Andrew D. Fernandes, Jennifer Ns Reid, Jean M. Macklaim, Thomas A. McMurrough,

David R. Edgell, and Gregory B. Gloor. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, 2:15, 2014.

[164] Joseph A. Edwards, Christian M. Santos-Medellín, Zachary S. Liechty, Bao Nguyen, Eugene Lurie, Shane Eason, Gregory Phillips, and Venkatesan Sundaresan. Compositional shifts in root-associated bacterial and archaeal microbiota track the plant life cycle in field-grown rice. *PLoS biology*, 16(2):e2003862, February 2018.

[165] G. S. Vernikos. A Review of Pangenome Tools and Recent Studies. In Hervé Tettelin and Duccio Medini, editors, *The Pangenome: Diversity, Dynamics and Evolution of Genomes*. Springer, Cham (CH), 2020.

[166] Derrick E. Fouts, Lauren Brinkac, Erin Beck, Jason Inman, and Granger Sutton. PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. *Nucleic Acids Research*, 40(22):e172, December 2012.

[167] Yongbing Zhao, Jiayan Wu, Junhui Yang, Shixiang Sun, Jingfa Xiao, and Jun Yu. PGAP: pan-genomes analysis pipeline. *Bioinformatics (Oxford, England)*, 28(3):416–418, February 2012.

[168] Jason W. Sahl, J. Gregory Caporaso, David A. Rasko, and Paul Keim. The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes. *PeerJ*, 2:e332, 2014.

[169] Andrew J. Page, Carla A. Cummins, Martin Hunt, Vanessa K. Wong, Sandra Reuter, Matthew T. G. Holden, Maria Fookes, Daniel Falush, Jacqueline A. Keane, and Julian Parkhill. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics (Oxford, England)*, 31(22):3691–3693, November 2015.

[170] Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics (Oxford, England)*, 28(23):3150–3152, December 2012.

[171] Matthew N Benedict, James R Henriksen, William W Metcalf, Rachel J Whitaker, and Nathan D Price. ITEP: An integrated toolkit for exploration of microbial pan-genomes. *BMC Genomics*, 15(1):8, December 2014.

[172] Stijn van Dongen and Cei Abreu-Goodger. Using MCL to extract clusters from networks. *Methods in Molecular Biology (Clifton, N.J.)*, 804:281–295, 2012.

[173] Tom O. Delmont and A. Murat Eren. Linking pangenomes and metagenomes: the *Prochlorococcus* metapangenome. *PeerJ*, 6:e4320, January 2018.

[174] Andrew Brantley Hall, Moran Yassour, Jenny Sauk, Ashley Garner, Xiaofang Jiang, Timothy Arthur, Georgia K. Lagoudas, Tommi Vatanen, Nadine Fornelos, Robin Wilson, Madeline Bertha, Melissa Cohen, John Garber, Hamed Khalili, Dirk Gevers, Ashwin N. Ananthakrishnan, Subra Kugathasan, Eric S. Lander, Paul Blainey, Hera Vlamakis, Ramnik J. Xavier, and Curtis Huttenhower. A novel Ruminococcus gnavus clade enriched in inflammatory bowel disease patients. *Genome Medicine*, 9(1):103, November 2017.

[175] Allyson L. Byrd, Yasmine Belkaid, and Julia A. Segre. The human skin microbiome. *Nature Reviews. Microbiology*, 16(3):143–155, March 2018.

[176] Yoshihisa Yamashita and Toru Takeshita. The oral microbiome and human health. *Journal of Oral Science*, 59(2):201–206, 2017.

[177] Braden T. Tierney, Zhen Yang, Jacob M. Luber, Marc Beaudin, Marsha C. Wibowo, Christina Baek, Eleanor Mehlenbacher, Chirag J. Patel, and Aleksandar D. Kostic. The Landscape of Genetic Content in the Gut and Oral Human Microbiome. *Cell Host & Microbe*, 26(2):283–295.e8, August 2019.

[178] Digvijay Verma, Pankaj Kumar Garg, and Ashok Kumar Dubey. Insights into the human oral microbiome. *Archives of Microbiology*, 200(4):525–540, May 2018.

[179] Miriam F. Moffatt and William Ocm Cookson. The lung microbiome in health and disease. *Clinical Medicine (London, England)*, 17(6):525–529, December 2017.

[180] Atanu Adak and Mojibur R. Khan. An insight into gut microbiota and its functionalities. *Cellular and molecular life sciences: CMLS*, 76(3):473–493, February 2019.

[181] Eman Zakaria Gomaa. Human gut microbiota/microbiome in health and diseases: a review. *Antonie Van Leeuwenhoek*, 113(12):2019–2040, December 2020.

[182] M. Hasan Mohajeri, Robert J. M. Brummer, Robert A. Rastall, Rinse K. Weersma, Hermie J. M. Harmsen, Marijke Faas, and Manfred Eggersdorfer. The role of the microbiome for human health: from basic science to clinical applications. *European Journal of Nutrition*, 57(Suppl 1):1–14, May 2018.

[183] Alexandre Almeida, Alex L. Mitchell, Miguel Boland, Samuel C. Forster, Gregory B. Gloor, Aleksandra Tarkowska, Trevor D. Lawley, and Robert D. Finn. A new genomic blueprint of the human gut microbiota. *Nature*, 568(7753):499–504, April 2019.

[184] Gustavo Santoyo, Gabriel Moreno-Hagelsieb, Ma del Carmen Orozco-Mosqueda, and Bernard R. Glick. Plant growth-promoting bacterial endophytes. *Microbiological Research*, 183:92–99, February 2016.

[185] Imran Afzal, Zabta Khan Shinwari, Shomaila Sikandar, and Shaheen Shahzad. Plant beneficial endophytic bacteria: Mechanisms, diversity, host range and genetic determinants. *Microbiological Research*, 221:36–49, April 2019.

[186] Luis Felipe Muriel-Millán, Sofía Millán-López, and Liliana Pardo-López. Biotechnological applications of marine bacteria in bioremediation of environments polluted with hydrocarbons and plastics. *Applied Microbiology and Biotechnology*, 105(19):7171–7185, October 2021.

[187] Hirak R. Dash, Neelam Mangwani, Jaya Chakraborty, Supriya Kumari, and Surajit Das. Marine bacteria: potential candidates for enhanced bioremediation. *Applied Microbiology and Biotechnology*, 97(2):561–571, January 2013.

[188] B. Matard, T. Meylheuc, R. Briandet, I. Casin, P. Assouly, B. Cavelier-balloy, and P. Reygagne. First evidence of bacterial biofilms in the anaerobe part of scalp hair follicles: a pilot comparative study in folliculitis decalvans. *Journal of the European Academy of Dermatology and Venereology: JEADV*, 27(7):853–860, July 2013.

[189] Christian F. P. Scholz and Mogens Kilian. The natural history of cutaneous propionibacteria,

and reclassification of selected species within the genus Propionibacterium to the proposed novel genera Acidipropionibacterium gen. nov., Cutibacterium gen. nov. and Pseudopropionibacterium gen. nov. *International Journal of Systematic and Evolutionary Microbiology*, 66(11):4422–4432, November 2016.

[190] Elizabeth A. Grice, Heidi H. Kong, Sean Conlan, Clayton B. Deming, Joie Davis, Alice C. Young, NISC Comparative Sequencing Program, Gerard G. Bouffard, Robert W. Blakesley, Patrick R. Murray, Eric D. Green, Maria L. Turner, and Julia A. Segre. Topographical and temporal diversity of the human skin microbiome. *Science (New York, N.Y.)*, 324(5931):1190–1192, May 2009.

[191] Allyson L. Byrd, Clay Deming, Sara K. B. Cassidy, Oliver J. Harrison, Weng-Ian Ng, Sean Conlan, NISC Comparative Sequencing Program, Yasmine Belkaid, Julia A. Segre, and Heidi H. Kong. Staphylococcus aureus and Staphylococcus epidermidis strain diversity underlying pediatric atopic dermatitis. *Science Translational Medicine*, 9(397):eaal4651, July 2017.

[192] Shruti Naik, Nicolas Bouladoux, Christoph Wilhelm, Michael J. Molloy, Rosalba Salcedo, Wolfgang Kastenmuller, Clayton Deming, Mariam Quinones, Lily Koo, Sean Conlan, Sean Spencer, Jason A. Hall, Amiran Dzutsev, Heidi Kong, Daniel J. Campbell, Giorgio Trinchieri, Julia A. Segre, and Yasmine Belkaid. Compartmentalized control of skin immunity by resident commensals. *Science (New York, N.Y.)*, 337(6098):1115–1119, August 2012.

[193] Christel Chehoud, Stavros Rafail, Amanda S. Tyldsley, John T. Seykora, John D. Lambris, and Elizabeth A. Grice. Complement modulates the cutaneous microbiome and inflammatory milieu. *Proceedings of the National Academy of Sciences of the United States of America*, 110(37):15061–15066, September 2013.

[194] Gitte J. M. Christensen, Christian F. P. Scholz, Jan Enghild, Holger Rohde, Mogens Kilian, Andrea Thürmer, Elzbieta Brzuszkiewicz, Hans B. Lomholt, and Holger Brüggemann. Antagonism between Staphylococcus epidermidis and Propionibacterium acnes and its ge-

nomic basis. *BMC genomics*, 17:152, February 2016.

[195] Anna L. Cogen, Kenshi Yamasaki, Katheryn M. Sanchez, Robert A. Dorschner, Yuping Lai, Daniel T. MacLeod, Justin W. Torpey, Michael Otto, Victor Nizet, Judy E. Kim, and Richard L. Gallo. Selective antimicrobial action is provided by phenol-soluble modulins derived from Staphylococcus epidermidis, a normal resident of the skin. *The Journal of Investigative Dermatology*, 130(1):192–200, January 2010.

[196] Lindsey Bomar, Silvio D. Brugger, Brian H. Yost, Sean S. Davies, and Katherine P. Lemon. Corynebacterium accolens Releases Antipneumococcal Free Fatty Acids from Human Nostril and Skin Surface Triacylglycerols. *mBio*, 7(1):e01725–01715, January 2016.

[197] Matthew M. Ramsey, Marcelo O. Freire, Rebecca A. Gabrilska, Kendra P. Rumbaugh, and Katherine P. Lemon. Staphylococcus aureus Shifts toward Commensalism in Response to Corynebacterium Species. *Frontiers in Microbiology*, 7:1230, 2016.

[198] J. E. E. Totté, W. T. van der Feltz, M. Hennekam, A. van Belkum, E. J. van Zuuren, and S. G. M. A. Pasmans. Prevalence and odds of Staphylococcus aureus carriage in atopic dermatitis: a systematic review and meta-analysis. *The British Journal of Dermatology*, 175(4):687–695, October 2016.

[199] V. N. Sehgal. Leprosy. *Dermatologic Clinics*, 12(4):629–644, October 1994.

[200] Rein M. G. J. Houben and Peter J. Dodd. The Global Burden of Latent Tuberculosis Infection: A Re-estimation Using Mathematical Modelling. *PLoS medicine*, 13(10):e1002152, October 2016.

[201] Mark Wansbrough-Jones and Richard Phillips. Buruli ulcer: emerging from obscurity. *Lancet (London, England)*, 367(9525):1849–1858, June 2006.

[202] Heather Lehman. Skin manifestations of primary immune deficiency. *Clinical Reviews in Allergy & Immunology*, 46(2):112–119, April 2014.

[203] Robert J. Palmer Jr. Composition and development of oral bacterial communities: Oral bacterial communities. *Periodontology 2000*, 64(1):20–39, February 2014.

[204] William G. Wade. The oral microbiome in health and disease. *Pharmacological Research*,

69(1):137–143, March 2013.

[205] Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, June 2012.

[206] William G. Wade. *Advances in Applied Microbiology*. Elsevier, August 2004.

[207] D. A. Scott, W. A. Coulter, and P.-J. Lamey. Oral shedding of herpes simplex virus type 1: a review. *Journal of Oral Pathology and Medicine*, 26(10):441–447, November 1997.

[208] David T Pride, Julia Salzman, Matthew Haynes, Forest Rohwer, Clara Davis-Long, Richard A White, Peter Loomer, Gary C Armitage, and David A Relman. Evidence of a robust resident bacteriophage population revealed through analysis of the human salivary virome. *The ISME Journal*, 6(5):915–926, May 2012.

[209] J R Porter. Antony van Leeuwenhoek: tercentenary of his discovery of bacteria. *Bacteriological Reviews*, 40(2):260–269, June 1976.

[210] Nicola Segata, Susan Haake, Peter Mannon, Katherine P Lemon, Levi Waldron, Dirk Gevers, Curtis Huttenhower, and Jacques Izard. Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome Biology*, 13(6):R42, 2012.

[211] Floyd E. Dewhirst, Tuste Chen, Jacques Izard, Bruce J. Paster, Anne C. R. Tanner, Wen-Han Yu, Abirami Lakshmanan, and William G. Wade. The Human Oral Microbiome. *Journal of Bacteriology*, 192(19):5002–5017, October 2010.

[212] Egija Zaura, Bart JF Keijser, Susan M Huse, and Wim Crielaard. Defining the healthy ”core microbiome” of oral microbial communities. *BMC Microbiology*, 9(1):259, 2009.

[213] E J Vollaard and H A Clasener. Colonization resistance. *Antimicrobial Agents and Chemotherapy*, 38(3):409–414, March 1994.

[214] Åsa Sullivan, Charlotta Edlund, and Carl Erik Nord. Effect of antimicrobial agents on the ecological balance of human microflora. *The Lancet Infectious Diseases*, 1(2):101–114, September 2001.

[215] Philip A Wescombe, Nicholas CK Heng, Jeremy P Burton, Chris N Chilcott, and John R

Tagg. Streptococcal bacteriocins and the case for *Streptococcus* salivarius as model oral probiotics. *Future Microbiology*, 4(7):819–835, September 2009.

[216] J.P. Burton, C.N. Chilcott, C.J. Moore, G. Speiser, and J.R. Tagg. A preliminary study of the effect of probiotic Streptococcus salivarius K12 on oral malodour parameters. *Journal of Applied Microbiology*, 100(4):754–764, April 2006.

[217] N. Takahashi and B. Nyvad. The Role of Bacteria in the Caries Process: Ecological Perspectives. *Journal of Dental Research*, 90(3):294–303, March 2011.

[218] J. Coventry, G. Griffiths, C. Scully, and M. Tonetti. ABC of oral health: Periodontal disease. *BMJ*, 321(7252):36–39, July 2000.

[219] Paul E. Kolenbrander, Robert J. Palmer Jr, Alexander H. Rickard, Nicholas S. Jakubovics, Natalia I. Chalmers, and Patricia I. Diaz. Bacterial interactions and successions during plaque development. *Periodontol*, pages 42–47, 2006.

[220] Bente Nyvad and Mogens Kilian. Microbiology of the early colonization of human enamel and root surfaces in vivo. *European Journal of Oral Sciences*, 95(5):369–380, October 1987.

[221] S. S. Socransky. Microbiology of Periodontal Disease—Present Status and Future Considerations. *Journal of Periodontology*, 48(9):497–504, September 1977.

[222] A. M. Stephen and J. H. Cummings. The microbial contribution to human faecal mass. *Journal of Medical Microbiology*, 13(1):45–56, February 1980.

[223] Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, Nicolas Pons, Florence Levenez, Takuji Yamada, Daniel R. Mende, Junhua Li, Junming Xu, Shaochuan Li, Dongfang Li, Jianjun Cao, Bo Wang, Huiqing Liang, Huisong Zheng, Yinlong Xie, Julien Tap, Patricia Lepage, Marcelo Bertalan, Jean-Michel Batto, Torben Hansen, Denis Le Paslier, Allan Linneberg, H. Bjørn Nielsen, Eric Pelletier, Pierre Renault, Thomas Sicheritz-Ponten, Keith Turner, Hongmei Zhu, Chang Yu, Shengting Li, Min Jian, Yan Zhou, Yingrui Li, Xiuqing Zhang, Songgang Li, Nan Qin, Huanming Yang, Jian Wang, Søren Brunak, Joel Doré, Francisco

Guarner, Karsten Kristiansen, Oluf Pedersen, Julian Parkhill, Jean Weissenbach, MetaHIT Consortium, Peer Bork, S. Dusko Ehrlich, and Jun Wang. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65, March 2010.

[224] Human Microbiome Project Consortium. A framework for human microbiome research. *Nature*, 486(7402):215–221, June 2012.

[225] Die Dai, Jiaying Zhu, Chuqing Sun, Min Li, Jinxin Liu, Sicheng Wu, Kang Ning, Li-jie He, Xing-Ming Zhao, and Wei-Hua Chen. GMrepo v2: a curated human gut microbiome database with special focus on disease markers and cross-dataset comparison. *Nucleic Acids Research*, 50(D1):D777–D784, January 2022.

[226] Sicheng Wu, Chuqing Sun, Yanze Li, Teng Wang, Longhao Jia, Senying Lai, Yaling Yang, Pengyu Luo, Die Dai, Yong-Qing Yang, Qibin Luo, Na L Gao, Kang Ning, Li-jie He, Xing-Ming Zhao, and Wei-Hua Chen. GMrepo: a database of curated and consistently annotated human gut metagenomes. *Nucleic Acids Research*, 48(D1):D545–D553, January 2020.

[227] Caroline Canavan, Joe West, and Timothy Card. The epidemiology of irritable bowel syndrome. *Clinical Epidemiology*, 6:71–80, 2014.

[228] Ian B. Jeffery, Paul W. O'Toole, Lena Öhman, Marcus J. Claesson, Jennifer Deane, Eamonn M. M. Quigley, and Magnus Simrén. An irritable bowel syndrome subtype defined by species-specific alterations in faecal microbiota. *Gut*, 61(7):997–1006, July 2012.

[229] Mirjana Rajilić-Stojanović, Elena Biagi, Hans G. H. J. Heilig, Kajsa Kajander, Riina A. Kekkonen, Sebastian Tims, and Willem M. de Vos. Global and deep molecular analysis of microbiota signatures in fecal samples from patients with irritable bowel syndrome. *Gastroenterology*, 141(5):1792–1801, November 2011.

[230] Julien Tap, Muriel Derrien, Hans Törnblom, Rémi Brazeilles, Stéphanie Cools-Portier, Joël Doré, Stine Störsrud, Boris Le Nevé, Lena Öhman, and Magnus Simrén. Identification of an Intestinal Microbiota Signature Associated With Severity of Irritable Bowel Syndrome. *Gastroenterology*, 152(1):111–123.e8, January 2017.

[231] Angèle P. M. Kerckhoffs, Melvin Samsom, Michel E. van der Rest, Joris de Vogel, Jan

Knol, Kaouther Ben-Amor, and Louis M. A. Akkermans. Lower Bifidobacteria counts in both duodenal mucosa-associated and fecal microbiota in irritable bowel syndrome patients. *World Journal of Gastroenterology*, 15(23):2887–2892, June 2009.

[232] Erja Malinen, Teemu Rinttilä, Kajsa Kajander, Jaana Mättö, Anna Kassinen, Lotta Krogius, Maria Saarela, Riitta Korpela, and Airi Palva. Analysis of the fecal microbiota of irritable bowel syndrome patients and healthy controls with real-time PCR. *The American Journal of Gastroenterology*, 100(2):373–382, February 2005.

[233] Sung Noh Hong and Poong-Lyul Rhee. Unraveling the ties between irritable bowel syndrome and intestinal microbiota. *World Journal of Gastroenterology*, 20(10):2470–2481, March 2014.

[234] Anna Kassinen, Lotta Krogius-Kurikka, Harri Mäkivuokko, Teemu Rinttilä, Lars Paulin, Jukka Corander, Erja Malinen, Juha Apajalahti, and Airi Palva. The fecal microbiota of irritable bowel syndrome patients differs significantly from that of healthy subjects. *Gastroenterology*, 133(1):24–33, July 2007.

[235] Delphine M. Saulnier, Kevin Riehle, Toni-Ann Mistretta, Maria-Alejandra Diaz, Debasmita Mandal, Sabeen Raza, Erica M. Weidler, Xiang Qin, Cristian Coarfa, Aleksandar Milosavljevic, Joseph F. Petrosino, Sarah Highlander, Richard Gibbs, Susan V. Lynch, Robert J. Shulman, and James Versalovic. Gastrointestinal microbiome signatures of pediatric patients with irritable bowel syndrome. *Gastroenterology*, 141(5):1782–1791, November 2011.

[236] Atsushi Nishida, Ryo Inoue, Osamu Inatomi, Shigeki Bamba, Yuji Naito, and Akira Andoh. Gut microbiota in the pathogenesis of inflammatory bowel disease. *Clinical Journal of Gastroenterology*, 11(1):1–10, February 2018.

[237] Alan W. Walker, Jeremy D. Sanderson, Carol Churcher, Gareth C. Parkes, Barry N. Hudspith, Neil Rayment, Jonathan Brostoff, Julian Parkhill, Gordon Dougan, and Liljana Petrovska. High-throughput clone library analysis of the mucosa-associated microbiota reveals dysbiosis and differences between inflamed and non-inflamed regions of the intestine in inflammatory bowel disease. *BMC microbiology*, 11:7, January 2011.

[238] Daniel N. Frank, Allison L. St Amand, Robert A. Feldman, Edgar C. Boedeker, Noam Harpaz, and Norman R. Pace. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proceedings of the National Academy of Sciences of the United States of America*, 104(34):13780–13785, August 2007.

[239] C. Manichanh, L. Rigottier-Gois, E. Bonnaud, K. Gloux, E. Pelletier, L. Frangeul, R. Nalin, C. Jarrin, P. Chardon, P. Marteau, J. Roca, and J. Dore. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut*, 55(2):205–211, February 2006.

[240] E. Varela, C. Manichanh, M. Gallart, A. Torrejón, N. Borruel, F. Casellas, F. Guarner, and M. Antolin. Colonisation by Faecalibacterium prausnitzii and maintenance of clinical remission in patients with ulcerative colitis. *Alimentary Pharmacology & Therapeutics*, 38(2):151–161, July 2013.

[241] Harry Sokol, Bénédicte Pigneur, Laurie Watterlot, Omar Lakhdari, Luis G. Bermúdez-Humarán, Jean-Jacques Gratadoux, Sébastien Blugeon, Chantal Bridonneau, Jean-Pierre Furet, Gérard Corthier, Corinne Grangette, Nadia Vasquez, Philippe Pochart, Germain Trugnan, Ginette Thomas, Hervé M. Blottière, Joël Doré, Philippe Marteau, Philippe Seksik, and Philippe Langella. Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proceedings of the National Academy of Sciences of the United States of America*, 105(43):16731–16736, October 2008.

[242] Kyohei Nishino, Atsushi Nishida, Ryo Inoue, Yuki Kawada, Masashi Ohno, Shigeki Sakai, Osamu Inatomi, Shigeki Bamba, Mitsushige Sugimoto, Masahiro Kawahara, Yuji Naito, and Akira Andoh. Analysis of endoscopic brush samples identified mucosa-associated dysbiosis in inflammatory bowel disease. *Journal of Gastroenterology*, 53(1):95–106, January 2018.

[243] Martin Baumgart, Belgin Dogan, Mark Rishniw, Gil Weitzman, Brian Bosworth, Rhonda Yantiss, Renato H. Orsi, Martin Wiedmann, Patrick McDonough, Sung Guk Kim, Douglas

Berg, Ynte Schukken, Ellen Scherl, and Kenneth W. Simpson. Culture independent analysis of ileal mucosa reveals a selective increase in invasive Escherichia coli of novel phylogeny relative to depletion of Clostridiales in Crohn's disease involving the ileum. *The ISME journal*, 1(5):403–418, September 2007.

[244] Helen M. Martin, Barry J. Campbell, C. Anthony Hart, Chiedzo Mpofu, Manu Nayar, Ravinder Singh, Hans Englyst, Helen F. Williams, and Jonathan M. Rhodes. Enhanced Escherichia coli adherence and invasion in Crohn's disease and colon cancer. *Gastroenterology*, 127(1):80–93, July 2004.

[245] Julien Loubinoux, Jean-Pierre Bronowicki, Ines A. C. Pereira, Jean-Louis Mougenel, and Alain E. Faou. Sulfate-reducing bacteria in human feces and their association with inflammatory bowel diseases. *FEMS microbiology ecology*, 40(2):107–112, May 2002.

[246] Benjamin R. Joris and Gregory B. Gloor. Unaccounted risk of cardiovascular disease: the role of the microbiome in lipid metabolism. *Current Opinion in Lipidology*, 30(2):125–133, April 2019.

[247] A. C. I. Boullart, J. de Graaf, and A. F. Stalenhoef. Serum triglycerides and risk of cardiovascular disease. *Biochimica Et Biophysica Acta*, 1821(5):867–875, May 2012.

[248] Jacqueline S. Dron, Jian Wang, Cécile Low-Kam, Sumeet A. Khetarpal, John F. Robinson, Adam D. McIntyre, Matthew R. Ban, Henian Cao, David Rhainds, Marie-Pierre Dubé, Daniel J. Rader, Guillaume Lettre, Jean-Claude Tardif, and Robert A. Hegele. Polygenic determinants in extremes of high-density lipoprotein cholesterol. *Journal of Lipid Research*, 58(11):2162–2170, November 2017.

[249] Jingyuan Fu, Marc Jan Bonder, María Carmen Cenit, Ettje F. Tigchelaar, Astrid Maatman, Jackie A. M. Dekens, Eelke Brandsma, Joanna Marczynska, Floris Imhann, Rinse K. Weersma, Lude Franke, Tiffany W. Poon, Ramnik J. Xavier, Dirk Gevers, Marten H. Hofker, Cisca Wijmenga, and Alexandra Zhernakova. The Gut Microbiome Contributes to a Substantial Proportion of the Variation in Blood Lipids. *Circulation Research*, 117(9):817–824, October 2015.

[250] Dorothy A. Kieffer, Brian D. Piccolo, Maria L. Marco, Eun Bae Kim, Michael L. Goodson, Michael J. Keenan, Tamara N. Dunn, Knud Erik Bach Knudsen, Sean H. Adams, and Roy J. Martin. Obese Mice Fed a Diet Supplemented with Enzyme-Treated Wheat Bran Display Marked Shifts in the Liver Metabolome Concurrent with Altered Gut Bacteria. *The Journal of Nutrition*, 146(12):2445–2460, December 2016.

[251] Alissa C. Nicolucci, Megan P. Hume, Inés Martínez, Shyamchand Mayengbam, Jens Walter, and Raylene A. Reimer. Prebiotics Reduce Body Fat and Alter Intestinal Microbiota in Children Who Are Overweight or With Obesity. *Gastroenterology*, 153(3):711–722, September 2017.

[252] Yijun Zhu, Eleanor Jameson, Marialuisa Crosatti, Hendrik Schäfer, Kumar Rajakumar, Timothy D. H. Bugg, and Yin Chen. Carnitine metabolism to trimethylamine by an unusual Rieske-type oxygenase from human microbiota. *Proceedings of the National Academy of Sciences of the United States of America*, 111(11):4268–4273, March 2014.

[253] Robert A. Koeth, Bruce S. Levison, Miranda K. Culley, Jennifer A. Buffa, Zeneng Wang, Jill C. Gregory, Elin Org, Yuping Wu, Lin Li, Jonathan D. Smith, W. H. Wilson Tang, Joseph A. DiDonato, Aldons J. Lusis, and Stanley L. Hazen. -Butyrobetaine is a proatherogenic intermediate in gut microbial metabolism of L-carnitine to TMAO. *Cell Metabolism*, 20(5):799–812, November 2014.

[254] Zeneng Wang, Elizabeth Klipfell, Brian J. Bennett, Robert Koeth, Bruce S. Levison, Brandon Dugar, Ariel E. Feldstein, Earl B. Britt, Xiaoming Fu, Yoon-Mi Chung, Yuping Wu, Phil Schauer, Jonathan D. Smith, Hooman Allayee, W. H. Wilson Tang, Joseph A. DiDonato, Aldons J. Lusis, and Stanley L. Hazen. Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature*, 472(7341):57–63, April 2011.

[255] Brian J. Bennett, Thomas Q. de Aguiar Vallim, Zeneng Wang, Diana M. Shih, Yonghong Meng, Jill Gregory, Hooman Allayee, Richard Lee, Mark Graham, Rosanne Crooke, Peter A. Edwards, Stanley L. Hazen, and Aldons J. Lusis. Trimethylamine-N-oxide, a metabolite associated with atherosclerosis, exhibits complex genetic and dietary regulation. *Cell*

Metabolism, 17(1):49–60, January 2013.

[256] Zeneng Wang, Adam B. Roberts, Jennifer A. Buffa, Bruce S. Levison, Weifei Zhu, Elin Org, Xiaodong Gu, Ying Huang, Maryam Zamanian-Daryoush, Miranda K. Culley, Anthony J. DiDonato, Xiaoming Fu, Jennie E. Hazen, Daniel Krajcik, Joseph A. DiDonato, Aldons J. Lusis, and Stanley L. Hazen. Non-lethal Inhibition of Gut Microbial Trimethylamine Production for the Treatment of Atherosclerosis. *Cell*, 163(7):1585–1595, December 2015.

[257] Manya Warrier, Diana M. Shih, Amy C. Burrows, Daniel Ferguson, Anthony D. Gromovsky, Amanda L. Brown, Stephanie Marshall, Allison McDaniel, Rebecca C. Schugar, Zeneng Wang, Jessica Sacks, Xin Rong, Thomas de Aguiar Vallim, Jeff Chou, Pavlina T. Ivanova, David S. Myers, H. Alex Brown, Richard G. Lee, Rosanne M. Crooke, Mark J. Graham, Xiuli Liu, Paolo Parini, Peter Tontonoz, Aldon J. Lusis, Stanley L. Hazen, Ryan E. Temel, and J. Mark Brown. The TMAO-Generating Enzyme Flavin Monooxygenase 3 Is a Central Regulator of Cholesterol Balance. *Cell Reports*, 10(3):326–338, January 2015.

[258] Thomas Greiner and Fredrik Bäckhed. Effects of the gut microbiota on obesity and glucose homeostasis. *Trends in endocrinology and metabolism: TEM*, 22(4):117–123, April 2011.

[259] Peter J. Turnbaugh, Ruth E. Ley, Michael A. Mahowald, Vincent Magrini, Elaine R. Mardis, and Jeffrey I. Gordon. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444(7122):1027–1031, December 2006.

[260] Vanessa K. Ridaura, Jeremiah J. Faith, Federico E. Rey, Jiye Cheng, Alexis E. Duncan, Andrew L. Kau, Nicholas W. Griffin, Vincent Lombard, Bernard Henrissat, James R. Bain, Michael J. Muehlbauer, Olga Ilkayeva, Clay F. Semenkovich, Katsuhiko Funai, David K. Hayashi, Barbara J. Lyle, Margaret C. Martini, Luke K. Ursell, Jose C. Clemente, William Van Treuren, William A. Walters, Rob Knight, Christopher B. Newgard, Andrew C. Heath, and Jeffrey I. Gordon. Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science (New York, N.Y.)*, 341(6150):1241214, September 2013.

[261] Emmanuelle Le Chatelier, Trine Nielsen, Junjie Qin, Edi Prifti, Falk Hildebrand, Gwen Falony, Mathieu Almeida, Manimozhiyan Arumugam, Jean-Michel Batto, Sean Kennedy,

Pierre Leonard, Junhua Li, Kristoffer Burgdorf, Niels Grarup, Torben Jørgensen, Ivan Brandslund, Henrik Bjørn Nielsen, Agnieszka S. Juncker, Marcelo Bertalan, Florence Levenez, Nicolas Pons, Simon Rasmussen, Shinichi Sunagawa, Julien Tap, Sebastian Tims, Erwin G. Zoetendal, Søren Brunak, Karine Clément, Joël Doré, Michiel Kleerebezem, Karsten Kristiansen, Pierre Renault, Thomas Sicheritz-Ponten, Willem M. de Vos, Jean-Daniel Zucker, Jeroen Raes, Torben Hansen, MetaHIT consortium, Peer Bork, Jun Wang, S. Dusko Ehrlich, and Oluf Pedersen. Richness of human gut microbiome correlates with metabolic markers. *Nature*, 500(7464):541–546, August 2013.

[262] Junjie Qin, Yingrui Li, Zhiming Cai, Shenghui Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, Wenwei Zhang, Yuanlin Guan, Dongqian Shen, Yangqing Peng, Dongya Zhang, Zhuye Jie, Wenxian Wu, Youwen Qin, Wenbin Xue, Junhua Li, Lingchuan Han, Donghui Lu, Peixian Wu, Yali Dai, Xiaojuan Sun, Zesong Li, Aifa Tang, Shilong Zhong, Xiaoping Li, Weineng Chen, Ran Xu, Mingbang Wang, Qiang Feng, Meihua Gong, Jing Yu, Yanyan Zhang, Ming Zhang, Torben Hansen, Gaston Sanchez, Jeroen Raes, Gwen Falony, Shujiro Okuda, Mathieu Almeida, Emmanuelle LeChatelier, Pierre Renault, Nicolas Pons, Jean-Michel Batto, Zhaoxi Zhang, Hua Chen, Ruifu Yang, Weimou Zheng, Songgang Li, Huanming Yang, Jian Wang, S. Dusko Ehrlich, Rasmus Nielsen, Oluf Pedersen, Karsten Kristiansen, and Jun Wang. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55–60, October 2012.

[263] Nicholas Arpaia, Clarissa Campbell, Xiying Fan, Stanislav Dikiy, Joris van der Veeken, Paul deRoos, Hui Liu, Justin R. Cross, Klaus Pfeffer, Paul J. Coffer, and Alexander Y. Rudensky. Metabolites produced by commensal bacteria promote peripheral regulatory T-cell generation. *Nature*, 504(7480):451–455, December 2013.

[264] Jiunn-Wei Wang, Chao-Hung Kuo, Fu-Chen Kuo, Yao-Kuang Wang, Wen-Hung Hsu, Fang-Jung Yu, Huang-Ming Hu, Ping-I. Hsu, Jaw-Yuan Wang, and Deng-Chyang Wu. Fecal microbiota transplantation: Review and update. *Journal of the Formosan Medical Association*, 118:S23–S31, March 2019.

[265] T.J. Borody, L. George, P. Andrews, S. Brandl, S. Noonan, P. Cole, L. Hyland, A. Morgan, J. Maysey, and D. Moore-Jones. Bowel-flora alteration: a potential cure for inflammatory bowel disease and irritable bowel syndrome? *Medical Journal of Australia*, 150(10):604–604, May 1989.

[266] Els van Nood, Anne Vrieze, Max Nieuwdorp, Susana Fuentes, Erwin G. Zoetendal, Willem M. de Vos, Caroline E. Visser, Ed J. Kuijper, Joep F.W.M. Bartelsman, Jan G.P. Tijssen, Peter Speelman, Marcel G.W. Dijkgraaf, and Josbert J. Keller. Duodenal Infusion of Donor Feces for Recurrent Clostridium difficile. *New England Journal of Medicine*, 368(5):407–415, January 2013.

[267] Anne Vrieze, Els Van Nood, Frits Holleman, Jarkko Salojärvi, Ruud S. Kootte, Joep F.W.M. Bartelsman, Geesje M. Dallinga–Thie, Mariette T. Ackermans, Mireille J. Serlie, Raish Oozeer, Muriel Derrien, Anne Druesne, Johan E.T. Van Hylckama Vlieg, Vincent W. Bloks, Albert K. Groen, Hans G.H.J. Heilig, Erwin G. Zoetendal, Erik S. Stroes, Willem M. de Vos, Joost B.L. Hoekstra, and Max Nieuwdorp. Transfer of Intestinal Microbiota From Lean Donors Increases Insulin Sensitivity in Individuals With Metabolic Syndrome. *Gastroenterology*, 143(4):913–916.e7, October 2012.

[268] F. Cremonini, S. Di Caro, E. C. Nista, F. Bartolozzi, G. Capelli, G. Gasbarrini, and A. Gasbarrini. Meta-analysis: the effect of probiotic administration on antibiotic-associated diarrhoea. *Alimentary Pharmacology & Therapeutics*, 16(8):1461–1467, August 2002.

[269] M. Lorea Baroja, P. V. Kirjavainen, S. Hekmat, and G. Reid. Anti-inflammatory effects of probiotic yogurt in inflammatory bowel disease patients. *Clinical and Experimental Immunology*, 149(3):470–479, September 2007.

[270] Derek M. Bickhart, Mikhail Kolmogorov, Elizabeth Tseng, Daniel M. Portik, Anton Korobeynikov, Ivan Tolstoganov, Gherman Uritskiy, Ivan Liachko, Shawn T. Sullivan, Sung Bong Shin, Alvah Zorea, Victòria Pascal Andreu, Kevin Panke-Buisse, Marnix H. Medema, Itzhak Mizrahi, Pavel A. Pevzner, and Timothy P. L. Smith. Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities.

Nature Biotechnology, January 2022.

[271] Pratima Gupta and Batul Diwan. Bacterial Exopolysaccharide mediated heavy metal removal: A Review on biosynthesis, mechanism and remediation strategies. *Biotechnology Reports (Amsterdam, Netherlands)*, 13:58–71, March 2017.

[272] Donat-P. Häder, Anastazia T. Banaszak, Virginia E. Villafañe, Maite A. Narvarte, Raúl A. González, and E. Walter Helbling. Anthropogenic pollution of aquatic ecosystems: Emerging problems with global implications. *Science of The Total Environment*, 713:136586, April 2020.

[273] Raymond A. Wuana and Felix E. Okieimen. Heavy Metals in Contaminated Soils: A Review of Sources, Chemistry, Risks and Best Available Strategies for Remediation. *ISRN Ecology*, 2011:1–20, October 2011.

[274] Mao-Cheng Deng, Jing Li, Fu-Rui Liang, Meisheng Yi, Xiao-Ming Xu, Jian-Ping Yuan, Juan Peng, Chou-Fei Wu, and Jiang-Hai Wang. Isolation and characterization of a novel hydrocarbon-degrading bacterium Achromobacter sp. HZ01 from the crude oil-contaminated seawater at the Daya Bay, southern China. *Marine Pollution Bulletin*, 83(1):79–86, June 2014.

[275] Yi-Chen Liu, Ling-Zhi Li, Ying Wu, Wei Tian, Li-Ping Zhang, Lian Xu, Qi-Rong Shen, and Biao Shen. Isolation of an alkane-degrading Alcanivorax sp. strain 2B5 and cloning of the alkB gene. *Bioresource Technology*, 101(1):310–316, January 2010.

[276] Meriam Cheffi, Dorra Hentati, Alif Chebbi, Najla Mhiri, Sami Sayadi, Ana Maria Marqués, and Mohamed Chamkha. Isolation and characterization of a newly naphthalene-degrading Halomonas pacifica, strain Cnaph3: biodegradation and biosurfactant production studies. *3 Biotech*, 10(3):89, March 2020.

[277] Luis Felipe Muriel-Millán, José Luis Rodríguez-Mejía, Elizabeth Ernestina Godoy-Lozano, Nancy Rivera-Gómez, Rosa-María Gutierrez-Rios, Daniel Morales-Guzmán, María R. Trejo-Hernández, Alejandro Estradas-Romero, and Liliana Pardo-López. Functional and Genomic Characterization of a Pseudomonas aeruginosa Strain Isolated From the South-

western Gulf of Mexico Reveals an Enhanced Adaptation for Long-Chain Alkane Degradation. *Frontiers in Marine Science*, 6:572, September 2019.

[278] Josefien Van Landuyt, Lorenzo Cimmino, Charles Dumolin, Ioanna Chatzigiannidou, Felix Taveirne, Valérie Mattelin, Yu Zhang, Peter Vandamme, Alberto Scoma, Adam Williamson, and Nico Boon. Microbial enrichment, functional characterization and isolation from a cold seep yield piezotolerant obligate hydrocarbon degraders. *FEMS Microbiology Ecology*, 96(9):fiaa097, September 2020.

[279] Qingguo Chen, Jingjing Li, Mei Liu, Huiling Sun, and Mutai Bao. Study on the biodegradation of crude oil by free and immobilized bacterial consortium in marine environment. *PLOS ONE*, 12(3):e0174445, March 2017.

[280] S. Santisi, M. Catalfamo, M. Bonsignore, G. Gentile, E. Di Salvo, M. Genovese, M. Mahjoubi, A. Cherif, G. Mancini, M. Hassanshahian, G. Pioggia, and S. Cappello. Biodegradation ability of two selected microbial autochthonous consortia from a chronically polluted marine coastal area (Priolo Gargallo, Italy). *Journal of Applied Microbiology*, 127(3):618–629, September 2019.

[281] Baojiang Wang, Qiliang Lai, Zhisong Cui, Tianfeng Tan, and Zongze Shao. A pyrene-degrading consortium from deep-sea sediment of the West Pacific and its key member *Cycloclasticus* sp. P1. *Environmental Microbiology*, 10(8):1948–1963, August 2008.

[282] Fernando Rojo. Degradation of alkanes by bacteria. *Environmental Microbiology*, 11(10):2477–2490, October 2009.

[283] Yong Nie, Chang-Qiao Chi, Hui Fang, Jie-Liang Liang, She-Lian Lu, Guo-Li Lai, Yue-Qin Tang, and Xiao-Lei Wu. Diverse alkane hydroxylase genes in microorganisms and environments. *Scientific Reports*, 4(1):4968, May 2015.

[284] Kevin W. George and Anthony G. Hay. Bacterial Strategies for Growth on Aromatic Compounds. In *Advances in Applied Microbiology*, volume 74, pages 1–33. Elsevier, 2011.

[285] Junwei Cao, Qiliang Lai, Jun Yuan, and Zongze Shao. Genomic and metabolic analysis of fluoranthene degradation pathway in Celeribacter indicus P73T. *Scientific Reports*,

5(1):7741, July 2015.

[286] Márcia Duarte, Ruy Jauregui, Ramiro Vilchez-Vargas, Howard Junca, and Dietmar H. Pieper. AromaDeg, a novel database for phylogenomics of aerobic bacterial degradation of aromatics. *Database: The Journal of Biological Databases and Curation*, 2014:bau118, 2014.

[287] Varada Khot, Jackie Zorz, Daniel A. Gittins, Anirban Chakraborty, Emma Bell, María A. Bautista, Alexandre J. Paquette, Alyse K. Hawley, Breda Novotnik, Casey R. J. Hubert, Marc Strous, and Srijak Bhatnagar. CANT-HYD: A Curated Database of Phylogeny-Derived Hidden Markov Models for Annotation of Marker Genes Involved in Hydrocarbon Degradation. *Frontiers in Microbiology*, 12:764058, 2021.

[288] Varenyam Achal, Xiangliang Pan, Qinglong Fu, and Daoyong Zhang. Biomineralization based remediation of As(III) contaminated soil by Sporosarcina ginsengisoli. *Journal of Hazardous Materials*, 201-202:178–184, January 2012.

[289] Jinghong Zhang, Xu Zhang, Yongqing Ni, Xiaojuan Yang, and Hongyu Li. Bioleaching of arsenic from medicinal realgar by pure and mixed cultures. *Process Biochemistry*, 9(42):1265–1271, 2007.

[290] W. C. Leung, M.-F. Wong, H. Chua, W. Lo, P. H. F. Yu, and C. K. Leung. Removal and recovery of heavy metals by bacteria isolated from activated sludge treating industrial effluents and municipal wastewater. *Water Science and Technology*, 41(12):233–240, June 2000.

[291] H. A. Elliott, M. R. Liberati, and C. P. Huang. Competitive Adsorption of Heavy Metals by Soils. *Journal of Environmental Quality*, 15(3):214–219, July 1986.

[292] Milind Mohan Naik and Santosh Kumar Dubey. Lead-enhanced siderophore production and alteration in cell morphology in a Pb-resistant Pseudomonas aeruginosa strain 4EA. *Current Microbiology*, 62(2):409–414, February 2011.

[293] Lina Velásquez and Jenny Dussan. Biosorption and bioaccumulation of heavy metals on dead and living biomass of Bacillus sphaericus. *Journal of Hazardous Materials*, 167(1-

3):713–716, August 2009.

[294] Sonia M. Tiquia-Arashiro. Lead absorption mechanisms in bacteria as strategies for lead bioremediation. *Applied Microbiology and Biotechnology*, 102(13):5437–5444, July 2018.

[295] Shahid Sher and Abdul Rehman. Use of heavy metals resistant bacteria—a strategy for arsenic bioremediation. *Applied Microbiology and Biotechnology*, 103(15):6007–6021, August 2019.

[296] Bhupendra Pushkar, Pooja Sevak, Sejal Parab, and Nikita Nilkanth. Chromium pollution and its bioremediation mechanisms in bacteria: A review. *Journal of Environmental Management*, 287:112279, June 2021.

[297] Padma Seragadam, Abhilasha Rai, Kartik Chandra Ghanta, Badri Srinivas, Sandip Kumar Lahiri, and Susmita Dutta. Bioremediation of hexavalent chromium from wastewater using bacteria-a green technology. *Biodegradation*, 32(4):449–466, August 2021.

[298] Abd Elnaby Hanan, M Abou Elela Gehan, and A El Sersy Nermeen. Cadmium resisting bacteria in Alexandria Eastern Harbor (Egypt) and optimization of cadmium bioaccumulation by Vibrio harveyi. *African Journal of Biotechnology*, 10(17):3412–3423, April 2011.

[299] null Sode, null Yamamoto, and null Hatano. Construction of a marine cyanobacterial strain with increased heavy metal ion tolerance by introducing exogenic metallothionein gene. *Journal of Marine Biotechnology*, 6(3):174–177, August 1998.

[300] Chris Smillie, M Pilar Garcillán-Barcia, M Victoria Francia, Eduardo P C Rocha, and Fernando de la Cruz. Mobility of plasmids. *Microbiol Mol Biol Rev*, 74(3):434–52, September 2010.

[301] M Victoria Francia, Athanasia Varsaki, M Pilar Garcillán-Barcia, Amparo Latorre, Constantin Drainas, and Fernando de la Cruz. A classification scheme for mobilization regions of bacterial plasmids. *FEMS Microbiol Rev*, 28(1):79–100, February 2004.

[302] E C Becker and R J Meyer. Recognition of oriT for DNA processing at termination of a round of conjugal transfer. *J Mol Biol*, 300(5):1067–77, July 2000.

[303] Rebekah Potts Nash, Sohrab Habibi, Yuan Cheng, Scott A Lujan, and Matthew R Red-

inbo. The mechanism and control of DNA transfer by the conjugative relaxase of resistance plasmid pCU1. *Nucleic Acids Res*, 38(17):5929–43, September 2010.

[304] Jan Zrimec and Aleš Lapanje. DNA structure at the plasmid origin-of-transfer indicates its potential transfer range. *Sci Rep*, 8(1):1820, January 2018.

[305] Rémi Fronzes, Peter J Christie, and Gabriel Waksman. The structural biology of type IV secretion systems. *Nat Rev Microbiol*, 7(10):703–14, October 2009.

[306] Elena Cabezón, Jorge Ripoll-Rozada, Alejandro Peña, Fernando de la Cruz, and Ignacio Arechaga. Towards an integrated model of bacterial conjugation. *FEMS Microbiol Rev*, 39(1):81–95, January 2015.

[307] Anabel Alperi, Delfina Larrea, Esther Fernández-González, Christoph Dehio, Ellen L Zechner, and Matxalen Llosa. A translocation motif in relaxase TrwC specifically affects recruitment by its conjugative type IV secretion system. *J Bacteriol*, 195(22):4999–5006, November 2013.

[308] Minny Bhatty, Jenny A Laverde Gomez, and Peter J Christie. The expanding bacterial type IV secretion lexicon. *Res Microbiol*, 164(6):620–39, August 2013.

[309] Sergio Arredondo-Alonso, Rob J Willems, Willem van Schaik, and Anita C Schürch. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microb Genom*, 3(10):e000128, October 2017.

[310] Pawel S. Krawczyk, Leszek Lipinski, and Andrzej Dziembowski. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Research*, 46(6):e35, April 2018.

[311] Thomas A. Hamilton, Gregory M. Pellegrino, Jasmine A. Therrien, Dalton T. Ham, Peter C. Bartlett, Bogumil J. Karas, Gregory B. Gloor, and David R. Edgell. Efficient inter-species conjugative transfer of a CRISPR nuclease for targeted bacterial killing. *Nature Communications*, 10(1):4544, December 2019.

[312] Daniel J Giguere, Alexander T Bahcheli, Benjamin R Joris, Julie M Paulssen, Lisa M Gieg, Martin W Flatley, and Gregory B Gloor. Complete and validated genomes from a

metagenome. preprint, Bioinformatics, April 2020.

[313] Lauren M. Lui, Torben N. Nielsen, and Adam P. Arkin. A method for achieving complete microbial genomes and improving bins from metagenomics data. *PLOS Computational Biology*, 17(5):e1008972, May 2021.

Chapter 2

Identification of type IV conjugative systems that are systematically excluded from metagenomic bins

2.1 Introduction

Bacteria can acquire exogenous DNA through horizontal gene transfer. Conjugation is a common mechanism of horizontal gene transfer that relies on direct cell-cell contact to unidirectionally transfer DNA from a bacterial donor to a recipient cell. In bacteria, integrative conjugative elements (ICEs) and conjugative plasmids are mobilizable through the actions of type IV secretion systems (T4SS). ICEs are integrated on the chromosomal sequence whereas plasmids are circular and separate elements from the chromosome. Approximately half of the known plasmids are mobilizable in *trans* where the conjugative machinery is on a different genetic element than the transferred element, and the remainder are mobilizable in *cis* because the conjugative machinery is present on the same genetic element [1]. ICEs encode their own T4SS, and can mobilize other elements [2]. Conjugative elements (CEs) often contain antibiotic resistance genes, but also can harbour useful biosynthetic and biodegradation genes [3]. Furthermore, conjugative systems can serve as vectors to introduce clustered regularly interspaced short palindromic repeats (CRISPR) systems, metabolic pathways or novel functions into the gut microbiota [4, 5, 6, 7, 8, 9]. Therefore, characterizing the full complement of conjugative systems in the human gut could expand the number of useable vectors for these applications. Precise identification of conjugative systems from metagenomic samples could also provide insights to their distribution in populations and their correlation with antibiotic exposure, age, and health status.

For a DNA sequence to be considered mobilizable by conjugation, it must encode an origin of transfer (*oriT*) sequence that is recognized and nicked by a relaxase protein [1, 10]. Relaxase proteins contain a conserved histidine triad that coordinates a divalent metal ion, as well as tyrosine residues that catalyze the nicking reaction at the *oriT* DNA sequence [11, 12]. In addition to a relaxase gene and an *oriT* sequence, a full complement of type IV secretion system and coupling proteins are required for a sequence to be conjugative. In the well-studied *Agrobacterium tumefaciens* conjugative system, there are 12 proteins involved in the transfer of the DNA-relaxase complex from one bacterial cell to another [13, 14]. Homologs of the VirB4 ATPase that are essential for assembly of the conjugative system and DNA transfer are generally similar to the phylogeny of the bacteria harbouring them [15] and thus are useful for classifying conjugative systems [16]. The synteny of conjugative transfer genes is also highly conserved among conjugative systems [14]. Both the synteny and presence of highly-conserved genes involved in conjugation facilitates the classification of genetic elements as potentially conjugative if the sequences are annotated as belonging to the components of the T4SS [17] (Figure 2.1).

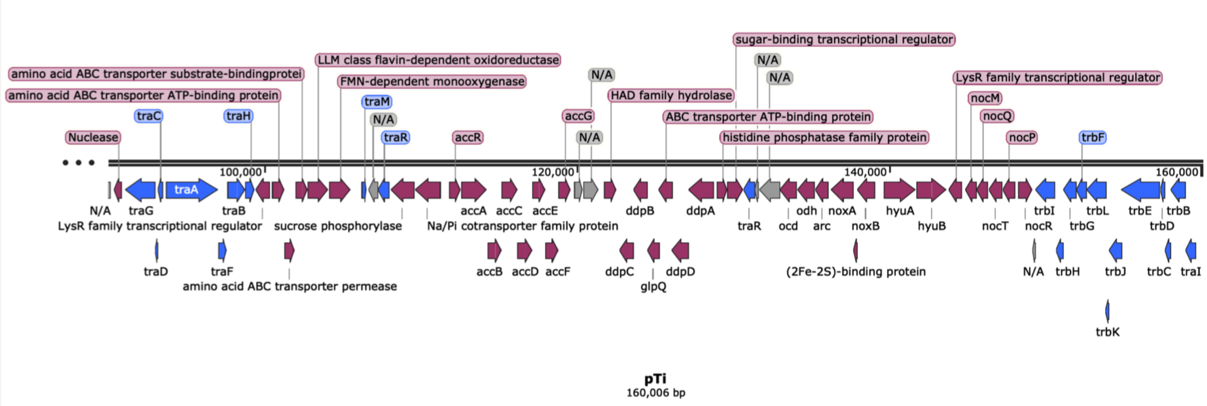

sugar-binding transcriptional regulator
LLM class flavin-dependent oxidoreductase
amino acid ABC transporter substrate-bindingprotei
amino acid ABC transporter ATP-binding protein
FMN-dependent monooxygenase
Nuclease
traC
traH
traM
N/A
traR
accR
accG
N/A
HAD family hydrolase
ABC transporter ATP-binding protein
histidine phosphatase family protein
N/A
N/A
LysR family transcriptional regulator
nocM
nocQ
nocP
trbF
100,000
120,000
140,000
160,000
traA
N/A
traG
traB
LysR family transcriptional regulator
sucrose phosphorylase
accA
accC
accE
Na/Pi cotransporter family protein
ddpB
ddpA
traR
ocd
odh
arc
noxA
noxB
hyuA
hyuB
nocT
trbI
nocR
trbL
trbG
trbE
trbB
trbD
traD
traF
amino acid ABC transporter permease
accB
accD
accF
ddpC
glpQ
ddpD
(2Fe-2S)-binding protein
N/A
trbH
trbJ
trbC
traI
trbK
pTi
160,006 bp

Figure 2.1: Example schematic of the gene organization of a bacterial conjugation system on the *Agrobacterium tumefaciens* pTi plasmid.

Previous work has identified novel CEs in the human and animal gut microbiomes, but the focus was mainly on ICEs and not on conjugative plasmids [3, 18, 19]. Identifying conjugative plasmids from a short-read metagenomic assembly is difficult for several reasons. The initial barrier is the difficulty in assembling circularized plasmids from short-read sequencing data [20]. A second barrier is that the contiguous DNA sequences (contigs) that compose metagenome-assembled genomes (MAGs) are binned together based on sequence composition and coverage. Binning of a plasmid with its cognate genome will not happen unless the contigs that compose the plasmid are maintained in the same copy-number and have the same sequence composition as the chromosome. These criteria are generally not met because conjugative systems are usually more AT rich than the cognate chromosome [1] and often do not have a unit copy number. Since nearly 80% of the non-redundant set of genomes from the human-gut microbiome are from difficult-to-culture species that are known only from MAGs [21], alternate methods must be employed to assemble and identify conjugative plasmids from the metagenomic sequencing data. Computational tools have recently been developed to identify plasmids from metagenomic assemblies [22], but would be less than optimal if applied to already binned data that systematically excludes plasmids [23]. Methods that identify CEs prior to binning should be able to capture the full spectrum of ICEs and conjugative plasmids.

Here, we show that T4SS conjugative systems can be identified using two distinct methods (Figure 2.2). First, we used profile HMMs (pHMMs) to identify conjugative systems directly from metagenomic assemblies of North American inflammatory bowel disease (IBD) and North American pre-term infant samples. Second, we searched predicted protein sequences from those same assemblies versus UniRef90 [24] for proteins involved in conjugation to identify conjugative systems. Both methods were able to find conjugative systems in raw metagenomic assemblies with pHMMs being computationally more efficient but less sensitive. Finally, we demonstrate that the majority of conjugative systems produced by a metagenomic assembly are not included in high-quality bins that are used as proxies for bacterial genomes in metagenomic analysis pipelines. Our findings provide a roadmap to integrate the analysis of conjugative systems alongside the chromosomal content of bacteria.

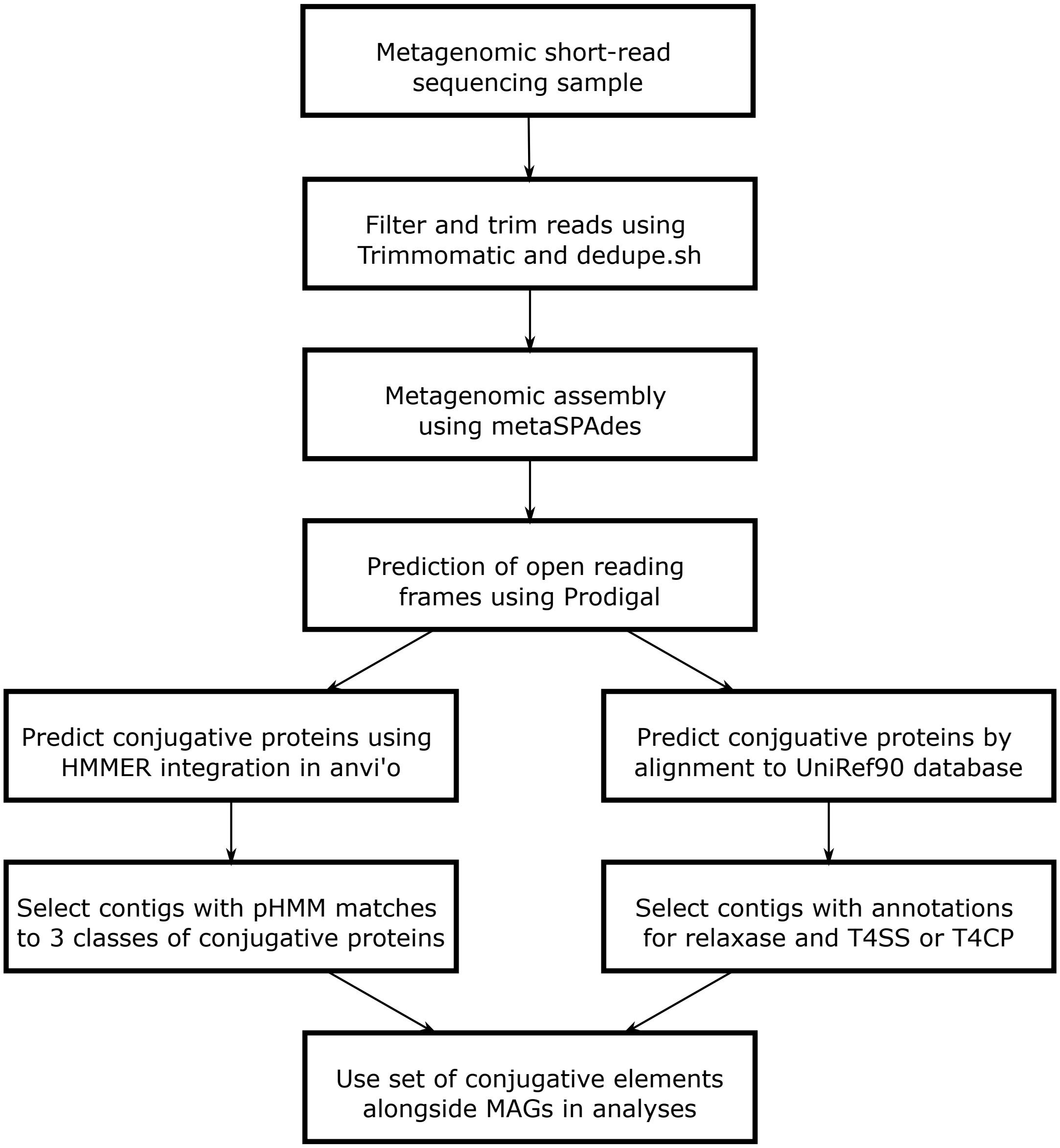

Metagenomic short-read sequencing sample
Filter and trim reads using Trimmomatic and dedupe.sh
Metagenomic assembly using metaSPAdes
Prediction of open reading frames using Prodigal
Predict conjugative proteins using HMMER integration in anvi'o
Predict conjguative proteins by alignment to UniRef90 database
Select contigs with pHMM matches to 3 classes of conjugative proteins
Select contigs with annotations for relaxase and T4SS or T4CP
Use set of conjugative elements alongside MAGs in analyses

Figure 2.2: Overview of methods employed in this study. In the left panel is the workflow used to identify conjugative systems from previously assembled human gut bacterial genomes. The right panel outlines the workflow for the assembly of select North American samples and the use of pHMMs to identify the conjugative systems.

2.2 Methods

2.2.1 Assembly and identification of conjugative systems in North American short-read data

Samples belonging to a North American IBD (n=50) [25] and a North American pre-term infant cohort (n=51) [26] were assembled *de novo* as follows. Reads from these samples were downloaded from the Sequence Read Archive using the SRA toolkit version 2.9.2, deduplicated with 'dedupe.sh' [27], and trimmed with Trimmomatic version 0.36 [28] with options 'LEADING:10 TRAILING:10'. Processed reads were assembled sample-by-sample using SPAdes version 3.14.0, option '–meta' [29].

2.2.2 Identification of conjugative systems using Profile hidden Markov models

The resultant assemblies were imported into Anvi'o version 6.0 [30] where the presence of T4SS, T4CP, and relaxase proteins were predicted using the 'anvi-run-hmms' module, which integrates HMMER3 functionality [31]. Instructions for installation of type IV conjugation pHMMs into Anvi'o can be found in the online code repository
(https://github.com/bjoris33/humanGutConj_Microbiome). In short, these pHMMs are the Pfam models for relaxases, type IV coupling proteins, and for the type IV secretion pilus proteins [32]. Contigs that contained pHMM matches for all three classes of conjugative proteins were extracted and annotated by aligning open reading frames (ORFs) predicted with Prodigal version 2.6.3 [33] to the UniRef90 database [24]. Taxonomic prediction of the contigs was conducted with Kaiju version 1.7.2 utilizing the RefSeq non-redundant protein database [34].

2.2.3 Identification of conjugative systems using protein alignments to the UniRef90 database

The contigs of the raw metagenomic assemblies had their ORFs predicted using Prodigal version 2.6.3 [33]. The predicted ORFs were then aligned to the UniRef90 database [24] using the 'blastp' module of DIAMOND version 0.9.14 [35]. By using keywords such as 'conjugal' or 'mobilization', the protein alignments were searched for contigs that contained annotations for both a relaxase and either a type IV secretion system protein or a type IV coupling protein. Through manual curation, type IV secretion system proteins and coupling proteins often shared identical or very similar annotation entries in the UniRef90 database, so the decision was made not to distinguish between the two.

2.2.4 Binning of Assemblies

For each assembly, all 101 samples were mapped to the contigs using Bowtie2 [36]. The mapping files were sorted and indexed with SAMtools [37] and then the assemblies were binned using MetaBAT2 version 2.12.1 [38]. CheckM version 1.1.2 was used to assess the quality of the resultant bins [39]. High-quality bins were defined using the same cutoffs (>90% completion and <5% redundancy) as Almeida *et al* [21] defined. Bins not passing that threshold were classified as 'low-quality'. The previously identified contigs with conjugative systems were classified based on their presence in bins, and the types of bins they were present in. Results of this classification were visualized using SankeyMATIC (http://sankeymatic.com/).

2.3 Results

2.3.1 Profile hidden Markov models and database alignment successfully identify conjugative elements from metagenomic assemblies

Fifty-one samples from a pre-term infant cohort and 50 from a North American IBD cohort were assembled sample-by-sample using metaSPAdes [29] to identify T4SS conjugative systems from a full pool of assembled contigs (i.e. not binned). Two separate methods we employed to search for contigs containing type IV conjugative proteins. For the pHMM method of identifying conjugative systems, contigs with conjugative systems were defined by pHMM matches for a relaxase, a type IV coupling protein, and a type IV secretion system, which offers a fast and precise method to annotate a limited number of protein families. From the assembly of the pre-term infant cohort 96 of 470500 contigs met the criteria, whereas 268 of 15100646 contigs from the IBD cohort did.

The second method of identifying conjugative systems utilizes the UniRef90 database by aligning the predicted ORFs to it using DIAMOND [35]. The alignment results are searched using a keyword strategy for contigs that contain an alignment for a relaxase or mobilization protein and an alignment for a type IV secretion or type IV coupling protein. From the pre-term infant cohort assemblies 242 of 470500 contigs met the described criteria, and 4244 of the 15100646 contigs from the IBD cohort met the same criteria.

The two outlined methods represent potentially complimentary methods of tackling the same problem–identifying conjugative systems from a pool of metagenomic-assembled contigs. There is a large-degree of overlap between the two methods, however alignment to the UniRef90 database appears to be much more sensitive with only 280 of the 4486 identified conjugative systems also identified using the pHMM method (Figure 2.3). While it may be less sensitive, the pHMM method of identifying conjugative systems has a much smaller computational footprint as it does not rely on aligning to a large protein database, but rather to a small and specific set of profile hidden Markov models.

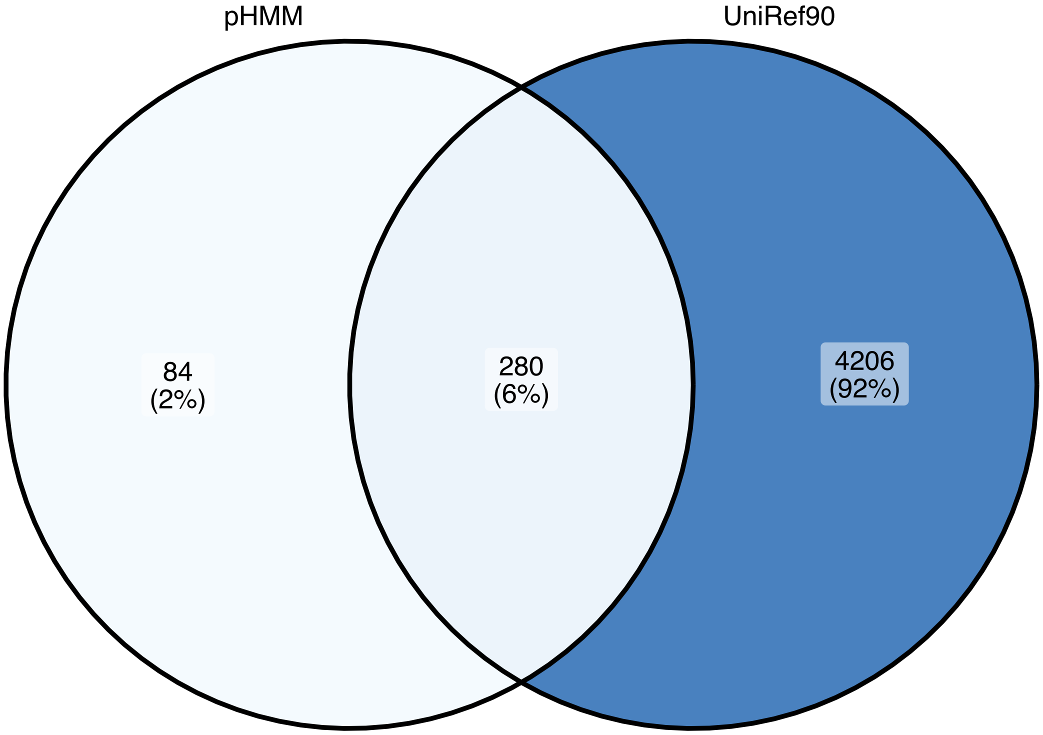


Figure 2.3: Venn diagram illustrating the overlap of the two methods of identifying type IV conjuagtive systems from the 101 assembled metagenomic samples. Each number represents the quantity of contigs that met the classification criteria for the labelled identification method. Intensity of blue shading indicated the proportion of the total contigs that were identified by a method or methods.

2.3.2 The majority of conjugative systems identified are omitted from metagenomic bins

Metagenomic assemblies from two distinct cohorts were binned using MetaBAT2 [38] to explore how conjugative systems are distributed within common metagenomic analyses. Of the 364 assembled contigs containing pHMM matches to all three protein categories, 270 were not included in any metagenomic bins (Figure 2.4). For the 94 contigs included in metagenomic bins, 65 of those were found in high-quality bins (>90% completion and <5% redundancy). This is in stark contrast to the background binning rate of contigs; For contigs above 5kb in size the binning rate with MetaBAT2 [38] was 70.4% (116112/164843 contigs) and for contigs above 10kb the binning rate was 79.1% (57214/72300 contigs). Among the 29 contigs included in bins that do not meet the aforementioned threshold, 8 are within bins that are less than or equal to 1 megabase in size, potentially suggesting that fragments of a conjugative plasmid may have binned together.

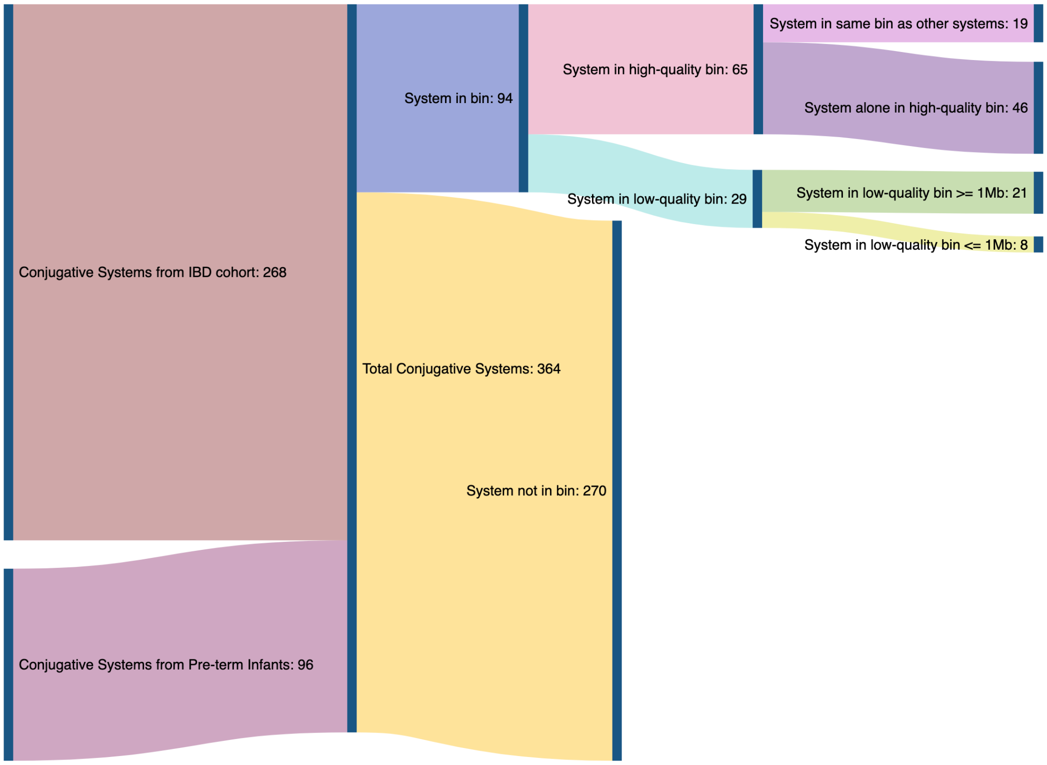


Figure 2.4: Sankey diagram representing the flow of 364 contigs containing conjugative systems identified using pHMMs into bins generated by MetaBAT2 from assembled data. Bin quality determined by CheckM.

For the conjugative systems established using alignments to the UniRef90 database, there is an even lower rate of binning (Figure 2.5). Of the 4486 conjugative systems, only 287 of them were binned–a rate of 6.4%. Again, a number of the bins that do form are low quality bins below 1mb in size that may be the collection of contigs that form a conjugative plasmid.

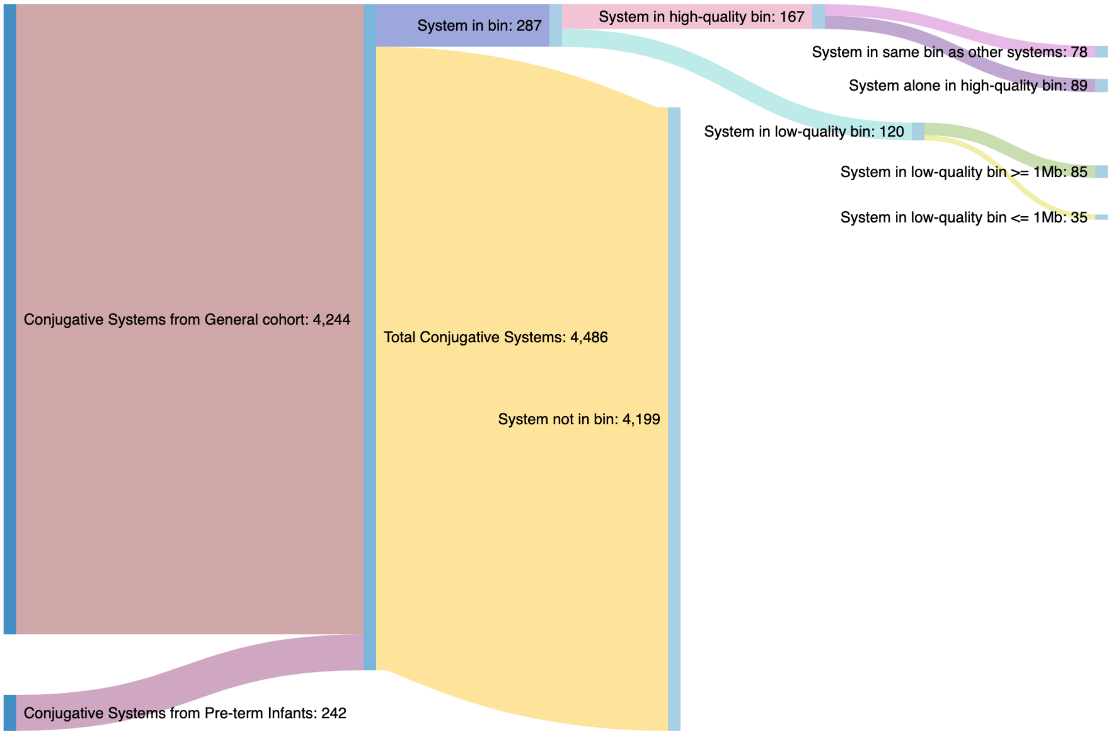


Figure 2.5: Sankey diagram representing the flow of 4486 contigs containing conjugative systems identified using predicted protein alignments to the UniRef90 database into bins generated by MetaBAT2 from assembled data. Bin quality determined by CheckM.

2.4 Discussion

To produce MAGs, contigs generated by metagenomic assembly are typically binned using a program such as MetaBAT2 [38]. Conjugative systems are often more AT rich than the parent genomes [1], which would result in the conjugative system and cognate genome not occurring in the same metagenomic bin because binning algorithms use GC content as a parameter for clustering. Additionally, plasmids are not necessarily maintained in a unit copy number within the cell, causing differential sequence coverage in comparison to the parent genome, which is another factor that leads to plasmids being excluded from MAGs. Therefore to capture a more complete image of the conjugative systems present in an environment, identification of the systems must take place before binning.

We have outlined two methods for identifying contigs carrying potentially functional type IV conjugative systems from raw metagenomic assemblies. Using a curated set of pHMMs of the three main classes of type IV conjugative proteins (relaxases, secretion proteins, and coupling proteins), we were able to classify 364 total contigs as being potentially conjugative. In comparison, the method that utilized predicted protein alignments to the UniRef90 database found 4486 contigs that met the criteria, which indicates that it may be the more sensitive method. However, aligning all predicted open reading frames found in a metagenomic assembly to the full UniRef90 database is a computationally expensive task and the reduced criteria for classification may lead to more false positives. Considering that many of the conjugative systems identified by pHMMs are also found by the protein alignment method (280 of 364 contigs), using pHMMs may be appropriate in as a first pass method or in situations where computational resources are scarce.

The assembled contigs were binned with MetaBAT2 as a way of quantifying the effect of binning, which revealed that the vast majority of the assembled conjugative systems were not included in metagenomic bins and therefore would not be included in a MAG database, which confirms recent findings [23]. The binning rate of contigs carrying type IV conjugative systems identified by pHMMs and alignment to UniRef90 was considerably lower than the background binning rate of equivalently sized contigs (25.8% and 6.4% compared to 70.4%, respectively). Many of the binned conjugative systems were not within a bin that would pass the quality cutoff to be included in a curated MAG genome set as well [21]. Interestingly, eight of the conjugative systems were binned into low-quality bins that were smaller than 1MB in size, which may suggest that the fragments of a conjugative plasmid could be binned together, which would increase the completeness of the conjugative system.

2.5 Conclusions

Conjugative systems could differ between cohorts and require special consideration to ensure that they are included in metagenomic analyses. ICEs and plasmids can carry harmful systems, such as antimicrobial resistance, but also can act as vectors for bile salt metabolism and for detoxification modules [3]. These cargo genes are relevant for research relating to the gut microbiome's role in pathogenicity as well as metabolism, digestion, and host effector molecules. Comprehensive identification and quantification of conjugative systems could allow for association of conjugative systems with different health outcomes. Because assembled conjugative systems are rarely included in metagenomic bins [23] (Figure 2.4 and Figure 2.5), they need to be identified and analyzed outside of standard binning pipelines. At present, it is not possible to assemble complete chromosomes or plasmids from short-read metagenomic data [20], so it may helpful to identify smaller bins containing conjugative systems in an attempt to cluster the fragments of plasmids present in an assembly together. Identifying type IV conjugative systems using pHMMs or UniRef90 annotations and using tools such as PlasFlow [22] to identify plasmids out of a full assembly in parallel with standard binning analyses will enhance research of the associations between the human gut microbiome and human health.

In the future, expanding the curated set of pHMMs by building exhaustive protein alignments for each of the known conjugative system proteins (T4SS, T4CP, and relaxases) could increase the sensitivity of the method to detect conjugative systems. Additionally, creating a curated set of conjugation proteins from the UniRef90, instead of exhaustively annotating with the full database, should improve the computational efficiency of the UniRef method. Additionally, improvements in assembly and binning algorithms will continue to improve the recovery of low relative abundance conjugative elements and improve the completeness and accuracy of the assembled fragments. For instance, long-read assembly permits the circularization of genomes and plasmids [40, 41] and the binning of plasmids to their cognate genomes using methylation data [42], which will reduce the ambiguity of the origins of conjugative systems (i.e. whether they are an ICE or independently circularized plasmid) and provide a more complete picture of the cargo they carry and the differences between cohorts.

2.6 References

Bibliography

[1] Chris Smillie, M Pilar Garcillán-Barcia, M Victoria Francia, Eduardo P C Rocha, and Fernando de la Cruz. Mobility of plasmids. *Microbiol Mol Biol Rev*, 74(3):434–52, September 2010.

[2] Aurélie Daccord, Daniela Ceccarelli, and Vincent Burrus. Integrating conjugative elements of the SXT/R391 family trigger the excision and drive the mobilization of a new class of Vibrio genomic islands. *Mol Microbiol*, 78(3):576–88, November 2010.

[3] Xiaofang Jiang, Andrew Brantley Hall, Ramnik J Xavier, and Eric J Alm. Comprehensive analysis of chromosomal mobile genetic elements in the gut microbiome reveals phylum-level niche-adaptive gene pools. *PLoS One*, 14(12):e0223680, 2019.

[4] Kevin Neil, Nancy Allard, Frédéric Grenier, Vincent Burrus, and Sébastien Rodrigue. Highly efficient gene transfer in the mouse gut microbiota is enabled by the Incl2 conjugative plasmid TP114. *Commun Biol*, 3(1):523, September 2020.

[5] Thomas A Hamilton, Gregory M Pellegrino, Jasmine A Therrien, Dalton T Ham, Peter C Bartlett, Bogumil J Karas, Gregory B Gloor, and David R Edgell. Efficient inter-species conjugative transfer of a CRISPR nuclease for targeted bacterial killing. *Nat Commun*, 10(1):4544, October 2019.

[6] Jason M Peters, Byoung-Mo Koo, Ramiro Patino, Gary E Heussler, Cameron C Hearne, Jiuxin Qu, Yuki F Inclan, John S Hawkins, Candy H S Lu, Melanie R Silvis, M Michael Harden, Hendrik Osadnik, Joseph E Peters, Joanne N Engel, Rachel J Dutton, Alan D Grossman, Carol A Gross, and Oren S Rosenberg. Enabling genetic analysis of diverse bacteria with Mobile-CRISPRi. *Nat Microbiol*, 4(2):244–250, February 2019.

[7] Robert J Citorik, Mark Mimee, and Timothy K Lu. Sequence-specific antimicrobials using efficiently delivered RNA-guided nucleases. *Nat Biotechnol*, 32(11):1141–5, November 2014.

[8] Ahmed A Gomaa, Heidi E Klumpe, Michelle L Luo, Kurt Selle, Rodolphe Barrangou, and Chase L Beisel. Programmable removal of bacterial strains by use of genome-targeting CRISPR-Cas systems. *mBio*, 5(1):e00928–13, January 2014.

[9] David Bikard, Chad W Euler, Wenyan Jiang, Philip M Nussenzweig, Gregory W Goldberg, Xavier Duportet, Vincent A Fischetti, and Luciano A Marraffini. Exploiting CRISPR-Cas nucleases to produce sequence-specific antimicrobials. *Nat Biotechnol*, 32(11):1146–50, November 2014.

[10] M Victoria Francia, Athanasia Varsaki, M Pilar Garcillán-Barcia, Amparo Latorre, Constantin Drainas, and Fernando de la Cruz. A classification scheme for mobilization regions of bacterial plasmids. *FEMS Microbiol Rev*, 28(1):79–100, February 2004.

[11] Rebekah Potts Nash, Sohrab Habibi, Yuan Cheng, Scott A Lujan, and Matthew R Redinbo. The mechanism and control of DNA transfer by the conjugative relaxase of resistance plasmid pCU1. *Nucleic Acids Res*, 38(17):5929–43, September 2010.

[12] E C Becker and R J Meyer. Recognition of oriT for DNA processing at termination of a round of conjugal transfer. *J Mol Biol*, 300(5):1067–77, July 2000.

[13] Rémi Fronzes, Peter J Christie, and Gabriel Waksman. The structural biology of type IV secretion systems. *Nat Rev Microbiol*, 7(10):703–14, October 2009.

[14] Elena Cabezón, Jorge Ripoll-Rozada, Alejandro Peña, Fernando de la Cruz, and Ignacio Arechaga. Towards an integrated model of bacterial conjugation. *FEMS Microbiol Rev*, 39(1):81–95, January 2015.

[15] Minny Bhatty, Jenny A Laverde Gomez, and Peter J Christie. The expanding bacterial type IV secretion lexicon. *Res Microbiol*, 164(6):620–39, August 2013.

[16] Julien Guglielmini, Fernando de la Cruz, and Eduardo P C Rocha. Evolution of conjugation and type IV secretion systems. *Mol Biol Evol*, 30(2):315–31, February 2013.

[17] Julien Guglielmini, Leonor Quintais, Maria Pilar Garcillán-Barcia, Fernando de la Cruz, and Eduardo P C Rocha. The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. *PLoS Genet*, 7(8):e1002222, August 2011.

[18] N Shterzer and I Mizrahi. The animal gut as a melting pot for horizontal gene transfer. *Can J Microbiol*, 61(9):603–5, September 2015.

[19] James H Kaufman, Ignacio Terrizzano, Gowri Nayar, Ed Seabolt, Akshay Agarwal, Ilya B Slizovskiy, and Noelle Noyes. Integrative and Conjugative Elements (ICE) and Associated Cargo Genes within and across Hundreds of Bacterial Genera. *bioRxiv*, 2020. Publisher: Cold Spring Harbor Laboratory _eprint: https://www.biorxiv.org/content/early/2020/08/03/2020.04.07.030320.full.pdf.

[20] Sergio Arredondo-Alonso, Rob J Willems, Willem van Schaik, and Anita C Schürch. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microb Genom*, 3(10):e000128, October 2017.

[21] Alexandre Almeida, Alex L Mitchell, Miguel Boland, Samuel C Forster, Gregory B Gloor, Aleksandra Tarkowska, Trevor D Lawley, and Robert D Finn. A new genomic blueprint of the human gut microbiota. *Nature*, 568(7753):499–504, April 2019.

[22] Pawel S Krawczyk, Leszek Lipinski, and Andrzej Dziembowski. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res*, 46(6):e35, April 2018.

[23] Finlay Maguire, Baofeng Jia, Kristen L Gray, Wing Yin Venus Lau, Robert G Beiko, and Fiona S L Brinkman. Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic Islands. *Microb Genom*, 6(10), October 2020.

[24] Baris E Suzek, Yuqi Wang, Hongzhan Huang, Peter B McGarvey, Cathy H Wu, and UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–32, March 2015.

[25] Andrew Brantley Hall, Moran Yassour, Jenny Sauk, Ashley Garner, Xiaofang Jiang, Timothy Arthur, Georgia K Lagoudas, Tommi Vatanen, Nadine Fornelos, Robin Wilson, Madeline Bertha, Melissa Cohen, John Garber, Hamed Khalili, Dirk Gevers, Ashwin N Ananthakrishnan, Subra Kugathasan, Eric S Lander, Paul Blainey, Hera Vlamakis, Ramnik J Xavier, and

Curtis Huttenhower. A novel Ruminococcus gnavus clade enriched in inflammatory bowel disease patients. *Genome Med*, 9(1):103, November 2017.

[26] Molly K Gibson, Bin Wang, Sara Ahmadi, Carey-Ann D Burnham, Phillip I Tarr, Barbara B Warner, and Gautam Dantas. Developmental dynamics of the preterm infant gut microbiota and antibiotic resistome. *Nat Microbiol*, 1:16024, March 2016.

[27] Brian Bushnell, Jonathan Rood, and Esther Singer. BBMerge - Accurate paired shotgun read merging via overlap. *PLoS One*, 12(10):e0185056, 2017.

[28] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–20, August 2014.

[29] Sergey Nurk, Dmitry Meleshko, Anton Korobeynikov, and Pavel A Pevzner. metaSPAdes: a new versatile metagenomic assembler. *Genome Res*, 27(5):824–834, May 2017.

[30] A Murat Eren, Özcan C Esen, Christopher Quince, Joseph H Vineis, Hilary G Morrison, Mitchell L Sogin, and Tom O Delmont. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*, 3:e1319, 2015.

[31] Sean R Eddy. Accelerated Profile HMM Searches. *PLoS Comput Biol*, 7(10):e1002195, October 2011.

[32] Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik L L Sonnhammer, Silvio C E Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, Robert D Finn, and Alex Bateman. Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1):D412–D419, January 2021.

[33] Doug Hyatt, Gwo-Liang Chen, Philip F Locascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11:119, March 2010.

[34] Peter Menzel, Kim Lee Ng, and Anders Krogh. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat Commun*, 7:11257, April 2016.

[35] Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*, 12(1):59–60, January 2015.

[36] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 9(4):357–9, March 2012.

[37] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–9, August 2009.

[38] Dongwan D Kang, Feng Li, Edward Kirton, Ashleigh Thomas, Rob Egan, Hong An, and Zhong Wang. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, 7:e7359, 2019.

[39] Donovan H Parks, Michael Imelfort, Connor T Skennerton, Philip Hugenholtz, and Gene W Tyson. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*, 25(7):1043–55, July 2015.

[40] Eli L. Moss, Dylan G. Maghini, and Ami S. Bhatt. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nature Biotechnology*, 2020. ISBN: 1546-1696.

[41] Daniel J Giguere, Alexander T Bahcheli, Benjamin R Joris, Julie M Paulssen, Lisa M Gieg, Martin W Flatley, and Gregory B Gloor. Complete and validated genomes from a metagenome. *bioRxiv*, 2020. Publisher: Cold Spring Harbor Laboratory _eprint: https://www.biorxiv.org/content/early/2020/04/09/2020.04.08.032540.full.pdf.

[42] John Beaulaurier, Shijia Zhu, Gintaras Deikus, Ilaria Mogno, Xue-Song Zhang, Austin Davis-Richardson, Ronald Canepa, Eric W Triplett, Jeremiah J Faith, Robert Sebra, Eric E Schadt, and Gang Fang. Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nat Biotechnol*, 36(1):61–69, January 2018.

Chapter 3

Separation of cohorts on the basis of bacterial type IV conjugation systems

3.1 Introduction

The human gut microbiome has been a recent target of much research on the connection between its composition and metabolism with human health. Broadly, the relative abundances of bacterial species in the human gut microbiome is differential between geographically-focused cohorts [1]. For instance, the microbiota compositions of hunter-gatherer tribes from West Africa are considerably different than the compositions of individuals from Western societies who are exposed to oral antibiotics and processed foods [1], which has helped spur investigations on the different phenotypes that can be linked to the compositions of the gut microbiota. These differences in composition can also be a causal or contributing factor to a number of aspects of human health.

A disrupted gut microbiota balance, and the resultant proinflammatory state, has been linked to a number of gastrointestinal, metabolic, immunological, and neurological diseases [2]. For example, inflammatory disorders that have associations with the species composition and metabolic output of the gut microbiota include irritable bowel syndrome, inflammatory bowel disease, and non-alcoholic fatty liver disease [3, 4, 5, 6]. Additionally, microbiota composition has been linked with metabolic diseases such as obesity and type 2 diabetes [7, 8], with certain bacterial genera such as *Campylobacter*, *Porphyromonas*, *Staphylococcus*, and *Ruminococcus* being enriched in the microbiota of obese individuals [9]. Beyond metabolic and gastrointestinal disorders, which have relatively clear mechanisms of pathogenesis from the microbiome, recent research has revealed associations with neurological diseases, such as the association between the inflammation caused by lipopolysaccharides and the amyloids formed in Alzheimer disease [10].

Spina bifida (also referred to as myelomeningocele) lies at the interface of metabolic and neurological disorders. Like the disorders previously mentioned, there is a sizeable proportion of risk of spina bifida that can be attributed to genetics (60%), however much of the attributable risk is unknown [11]. Some of the risk of development of spina bifida is related to the behaviours of the mother during pregnancy. For instance, smoking and alcohol intake are risk factors for congenital neural tube defects [12] and are also known to modulate the composition of the gut microbiota in a way that is proinflammatory and resembles the microbiotas of inflammatory bowel disease and obesity [13, 14]. Consumption of high glycemic index foods are also correlated with the risk of developing spina bifida [15]; this is a biomarker that can be regulated by gut microbiome [8]. Intake and plasma levels of vitamins and nutrients such as folate [16], vitamin B12 [16, 17], methionine [18], choline [19], vitamin C [20], and zinc [21] are all positively associated with the development of spina bifida. Many of these vitamins and nutrients are synthesized or metabolized by the human gut microbiome [22]. Folate, whose deficiency is potentially the most strongly associated nutrient with spina bifida risk is produced by a wide variety of bacterial taxa, with an estimated 13% of the species colonizing the human gut having the necessary genes for folate synthesis [23]; folate produced by the gut microbiota is bioavailable and a major source of serum folate [24]. Methionine deficiency is another risk factor for spina bifida that could be mediated through the gut microbiome. Metabolism of methionine by the gut microbiome results in the production of the short-chain fatty acids propionate and butyrate [25], which are key in the regulation of bowel permeability and inflammation [26]. Bowel inflammation and a ‘leaky gut’ are associated with a poorer uptake of nutrients [27], which could lower the serum and amniotic fluid levels of important vitamins such as folate and cobalamin and contribute to spina bifida risk. While there are a number of potential connections between spina bifida and the gut microbiome, there is an absence of research on the direct relationship.

In this chapter, the separation of cohorts on the basis of conjugative systems will be explored in cohorts that differ primarily on geographic location as a proof of principle. This is to establish that conjugative systems are differentially abundant between groups, like metagenomic-assembled genomes [1]. In the study of the association between spina bifida and the composition of the gut microbiota of mothers, we processed metageonomic reads following our approach that allows for the analysis of conjugative elements. As shown in Chapter 2, conjugative systems are systematically excluded from metagenomic bins, so they must be separately analyzed for a complete understanding of the differences in microbiota composition between cohorts. Assembly-based and assembly-free methods were utilized to examine the differences between the microbiotas of mothers who gave birth to infants with spina bifida compared to the mothers who gave birth to healthy infants. Conjugative systems were identified using the methods outlined in Chapter 2 and compared independently between groups from the genomic bins. All methods show robust differences between groups that reveal an association between human gut microbiota composition in mothers and spina bifida.

3.2 Methods

3.2.1 Reference human gut metagenome set

A near-complete and non-redundant set of human gut microbiome genomes were downloaded from the European Bioinformatics Institute FTP site [28]. These genomes were assembled from 13,133 metagenomic samples, a comprehensive set of samples available at the time, using SPAdes [29] and binned using MetaBAT2 [30]. The quality of binned genomes were assessed using CheckM [31]. High-quality genomes were defined as greater than 90% completeness and less than 5% contamination and medium-quality genomes were defined as greater than 50% completeness and less than 10% contamination, and these genomes were used to create the non-redundant set of genomes. The program dRep was used to cluster the genomes at 99% sequence identity [32] thereby dereplicating the genome bins, creating a set of 2505 genomes [28].

3.2.2 Identifying and quantifying conjugative systems in reference human gut metagenome set

ORFs were predicted in the genome by Prodigal version 2.6.3 [33]. The predicted protein sequences were then aligned to the UniRef90 database [34] using the Diamond protein aligner version 0.9.14 [35]. Contigs were extracted from the genomes if they contained annotations for a relaxase/mobilization protein and a type IV secretion/type IV coupling protein using a word-search strategy. Short-read data from 785 samples (Supplemental Table 2) [36, 37, 38, 39, 40, 41, 42, 43] were downloaded from the Sequence Read Archive using the SRA toolkit version 2.9.2. The downloaded reads were deduplicated with 'dedupe.sh' [44], and trimmed with Trimmomatic version 0.36 [45] with options 'LEADING:10 TRAILING:10'. Subregions of the contigs where annotations for conjugative proteins were present, with no more than 20 ORFs between successive UniRef90 annotations for conjugative proteins, were extracted. The processed read data were mapped to the extracted conjugative systems using Bowtie2 version 2.3.5 [46] with the settings '–no-unal –no-mixed –no-discordant'. Extraction of the sub-regions was to avoid an artificially high proportion of reads mapping in samples where the bacterium is present, but the ICE has not integrated in its chromosome (Figure 3.1). The proportion of reads mapping to the conjugative systems was extracted from the Bowtie2 output, and the mapping data was visualized using Anvi'o [47]. Raw counts of reads mapping to the extracted conjugative systems were transformed using a centered log-ratio. The principal component coordinates of the first three components were used for clustering by hdbscan as those components contained the majority of the variance explained [48].

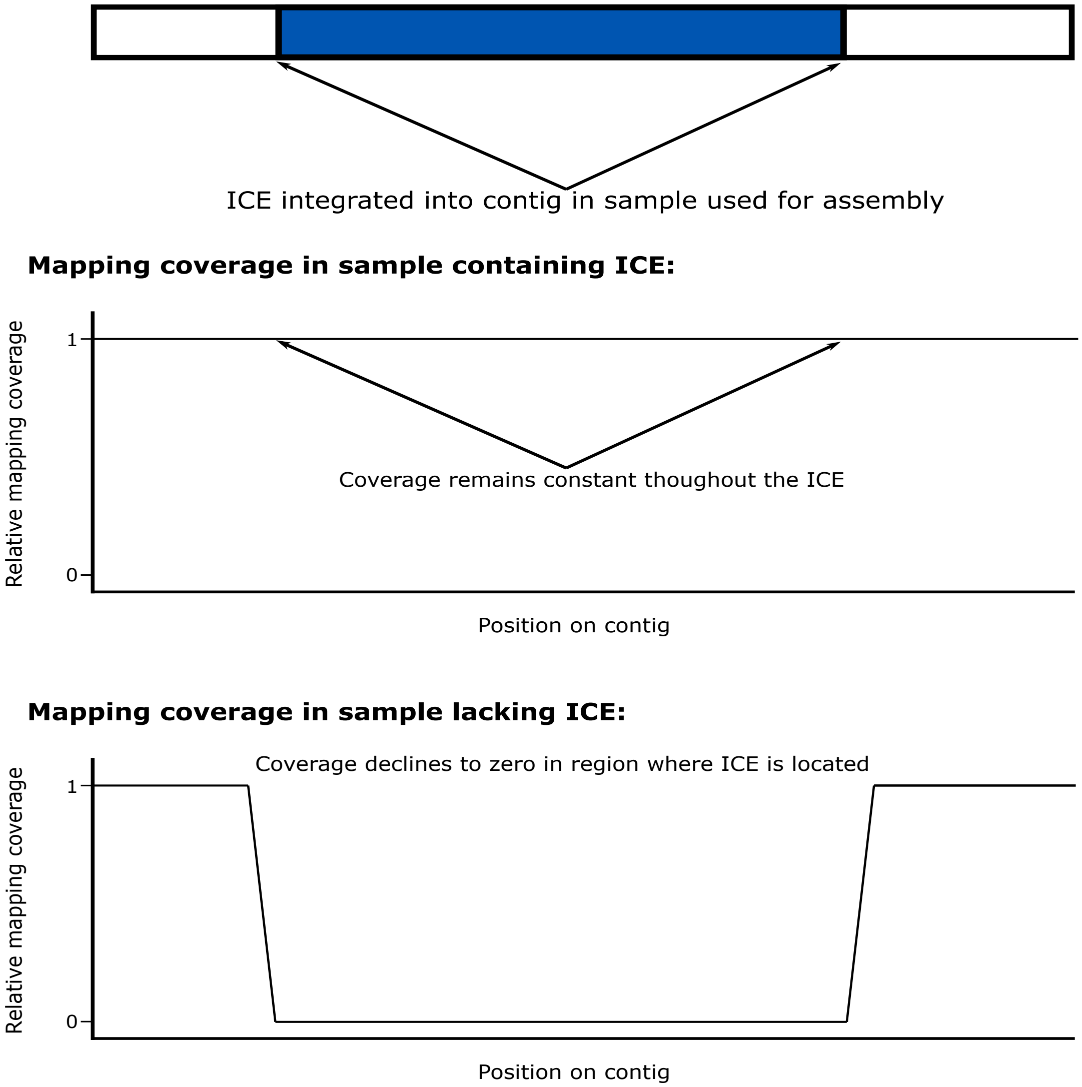

ICE integrated into contig in sample used for assembly
Mapping coverage in sample containing ICE:
Relative mapping coverage
1
0
Coverage remains constant thoughout the ICE
Position on contig
Mapping coverage in sample lacking ICE:
Relative mapping coverage
1
0
Coverage declines to zero in region where ICE is located
Position on contig

Figure 3.1: Conceptual diagram of the mapping coverage of an assembled integrative and conjugative element. The mapping coverage in the first plot shows an even mapping coverage across the contig because the ICE is present in the sample and the average mapping coverage of the contig would be an accurate metric. In the second plot, the ICE is missing in the sample and the mapping coverage falls to zero where the ICE is located on the contig. As a way to quantify the presence of the ICE, the average mapping coverage for the entire contig would be artificially high. Limiting the mapping to only the region containing the conjugative proteins solves this issue.

3.2.3 Metagenomic assembly of spina bifida microbiome

Sequencing reads obtained from the stool microbiomes of 15 mothers who had given birth to children with spina bifida and 18 samples from mothers who gave birth to healthy children were processed prior to assembly. The reads were first aligned to the Genome Reference Consortium Human Build 38 genome with Bowtie2 [46] to remove any human reads. As previously described, the reads were processed with 'dedupe.sh' [44] and Trimmomatic version 0.36 [45] to remove duplicated and low quality reads from the dataset. Using the processed reads, the samples were assembled sample-by-sample using metaSPAdes [29]. To bin the assemblies into approximations of bacterial genomes, the reads from all samples were mapped to all assembled metagenomes using Bowtie2 [46], sorted and indexed using SAMtools [49], and binned using MetaBAT2 version 2.12.1 [30]. All MetaBAT2 bins were pooled and dereplicated at a 99% sequence identity using dRep [32], a threshold that was used for the non-redundant gut microbiome genome set [28]. Bins with greater than 50% completion and less than 10% redundancy were used for quantitative analyses.

3.2.4 Identification of conjugative elements in spina bifida microbiome

Conjugative elements were identified from the unbinned metagenomic assembly by leveraging the profile hidden Markov model method of identifying conjugative systems developed in Chapter 2. For the assembly of each sample, the contigs were imported into Anvi'o [47] where the HMMER [50, 51] tool was used to search the open reading frames of the assembled contigs for conjugative systems. If a contig contained pHMM alignments for all three classes of type IV conjugative proteins (relaxase, type IV secretion system, and type IV coupling), then the contig would be deemed as conjugative.

3.2.5 Annotation and quantification of bins and conjugative elements in the spina bifida microbiome

Dereplicated bins and conjugative contigs were taxonomically assigned using two methods. Firstly, the bins were classified taxonomically using PhyloPhlAn with the database version SGB.Jul20 [52]. The primary assignment of the bin was used as the putative PhyloPhlAn taxonomic assignment. Secondly, the bins and conjugative contigs were taxonomically assigned with the bin assignment tool of the CAT software package [53]. Processed sequencing reads were aligned to the bins using Bowtie2 version 2.3.5 [46] with the settings '–no-unal –no-mixed –no-discordant' to quantify the abundances in each sample; read counts were obtained from the resultant SAM alignment files. The dereplicated bins were annotated with KEGG to estimate the metabolic capabilities of each bin [54]. KEGG analyses were conducted using the implementation in Anvi'o version 7.0 [47] with the functions 'anvi-run-kegg-kofams' and 'anvi-estimate-metabolism'. The open reading frames (ORFs) of the dereplicated bins were also predicted separately using Prodigal version 2.6.3 [33]. The predicted ORFs were annotated using InterProScan version 5.48-83.0 [55] using the Pfam [56] and GO databases [57, 58]. Average coverage of the predicted ORFs was computed using 'anvi-export-gene-coverage-and-detection' module of Anvi'o, which was then converted into separate count tables for both Pfam and GO database entries. Differences in relative abundances of bins and conjugative elements between groups was assessed using the ALDEx2 R package [59].

3.3 Results

3.3.1 Mapping human gut microbiome data from cohorts to conjugative systems reveals distinct geographic-based patterns

As a proof of principle of the utility of the conjugative system identification frameworks established in chapter 2, we explored relative abundances of conjugative systems in a larger number of cohorts, without having to conduct computationally-expensive metagenomic assemblies. For this analysis, conjugative systems were identified from a set of 2505 bacterial genomes, which represent a non-redundant and near-complete picture of the human gut microbiome [28]. A total of 1598 contigs from 787 genomes that contain UniRef90 annotations for relaxase/mobilization and T4SS/T4CP proteins were identified. From these contigs, 3216 subregions where conjugative protein annotations were concentrated on the contig were extracted, with 2413 being >1kb in size and used for visualization. Short-read human gut microbiome sequencing data from 785 samples, spread across 8 cohorts were aligned to the extracted subregions.

With the conjugative systems identified from the human gut metagenome set, there are distinct patterns that arise that are distinct to each cohort (Figure 3.2). Only a very small number of the reads from the North American and European Infant cohorts mapped to conjugative systems. The only notable signal is in the *Proteobacteria* phylum for the North American pre-term infants, a finding consistent with what was found by *de novo* assembly of these samples in chapter 2.

The West African and South American cohorts also share similar characteristics as both have an overall lower apparent relative abundance of conjugative systems compared to the other non-infant cohorts, particularly in the *Bacteroidetes* phylum. The other four cohorts appear similar with regards to the presence and absence of the conjugative systems. The cohorts separated into three distinct clusters (Figure 3.3 and 3.4), when the principal components of the centered log-ratio transformed data were clustered using hdbscan [48]. In this analysis infant cohorts were excluded because of their extreme sparsity. The majority of the West African and South American samples clustered together consistent with Figure 3.2. Not readily apparent from the cladogram was the East Asian cohort that clustered primarily on its own. The North American Indigenous, North American IBD, and Western European general samples largely clustered together. The distinct clustering of samples into geographic locations on the basis of the abundances of conjugative elements is similar to that of the broader gut microbiome [1].

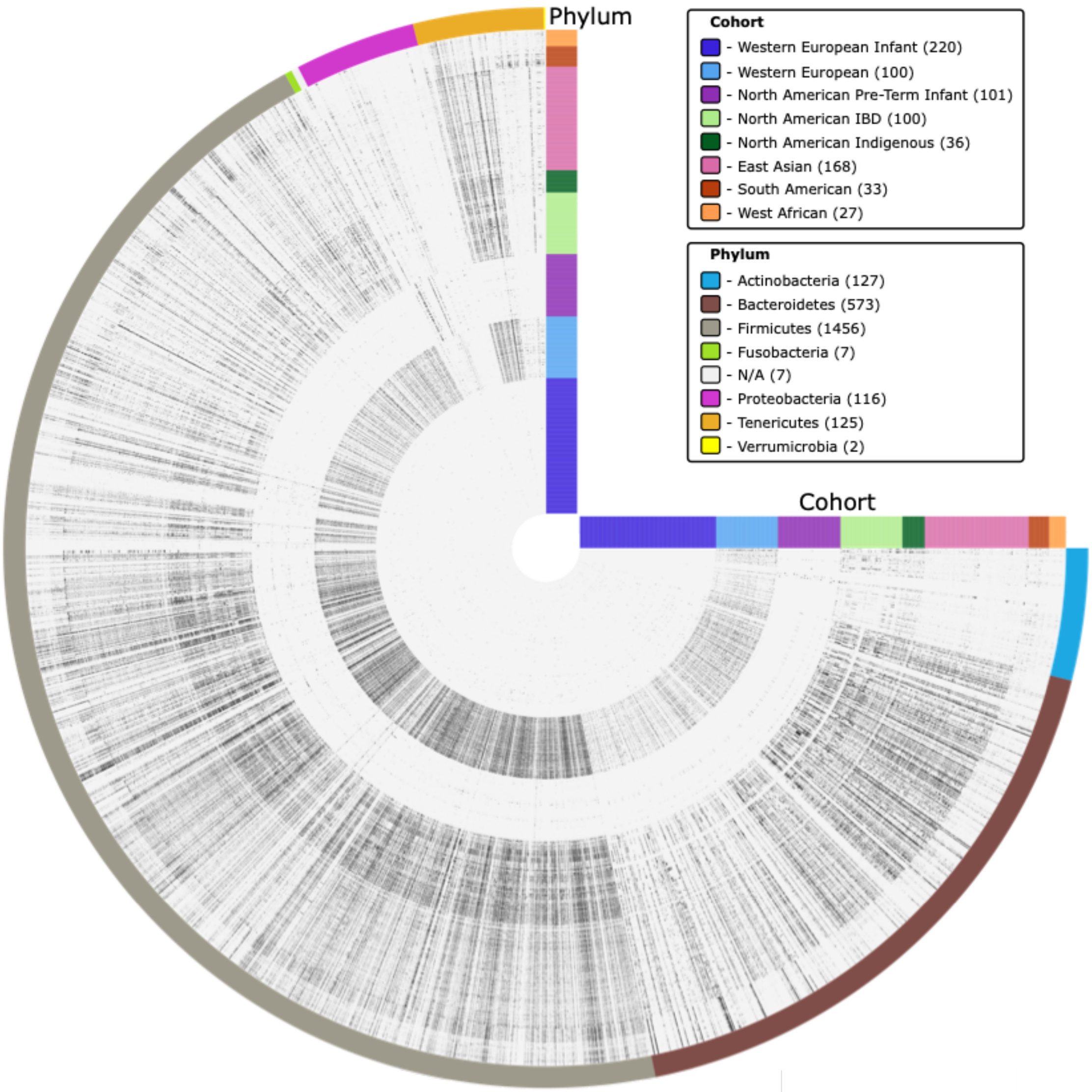


Figure 3.2: Anvi'o cladogram of potentially conjugative systems originating from 785 samples across 8 cohorts. Inner rings of the phylogram represent individual samples and the outermost ring being the phylum of conjugative system. Each slice of the circle phylogram are individual conjugative regions. For each point on the inner plot, the intensity of the black colouring corresponds to the mean coverage of the system for a given sample proportional to the other conjugative systems.

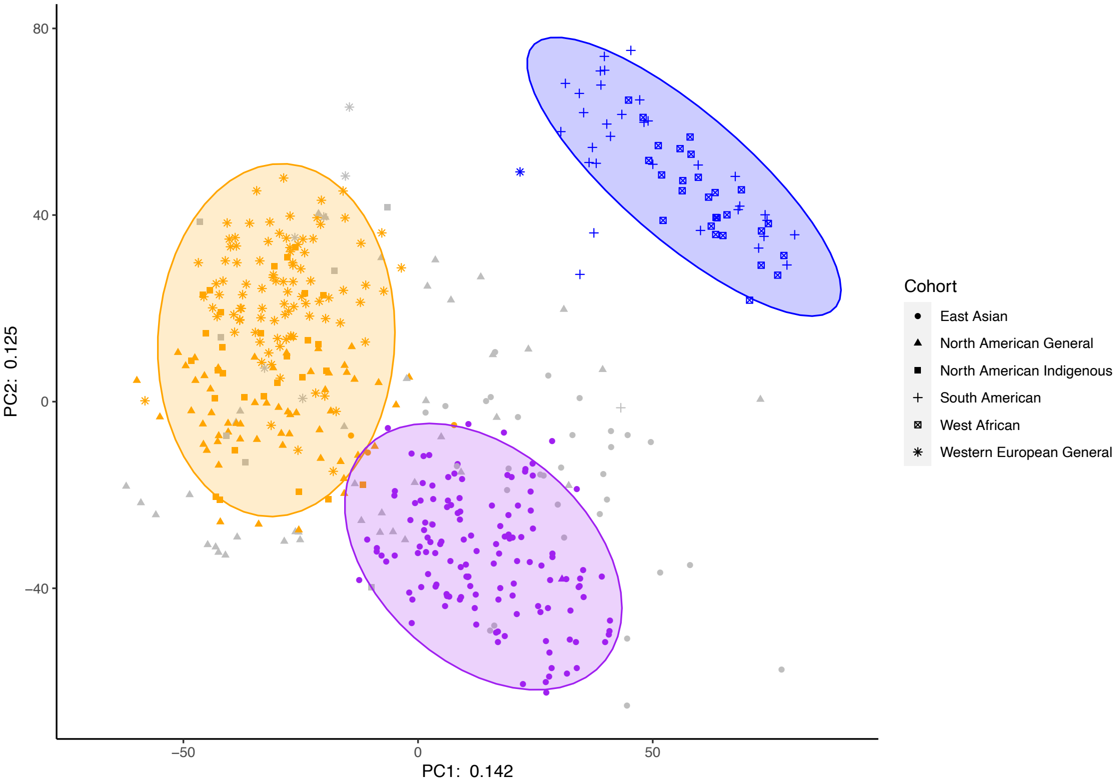


Figure 3.3: Clustering of the principal component coordinates of the CLR transformed relative abundances of the extracted conjugative regions from the genome database. Coloured points represent membership to clusters with grey points not belonging to a cluster. Ellipses represent a 95 percent confidence interval using a multivariate t-distribution about the cluster.

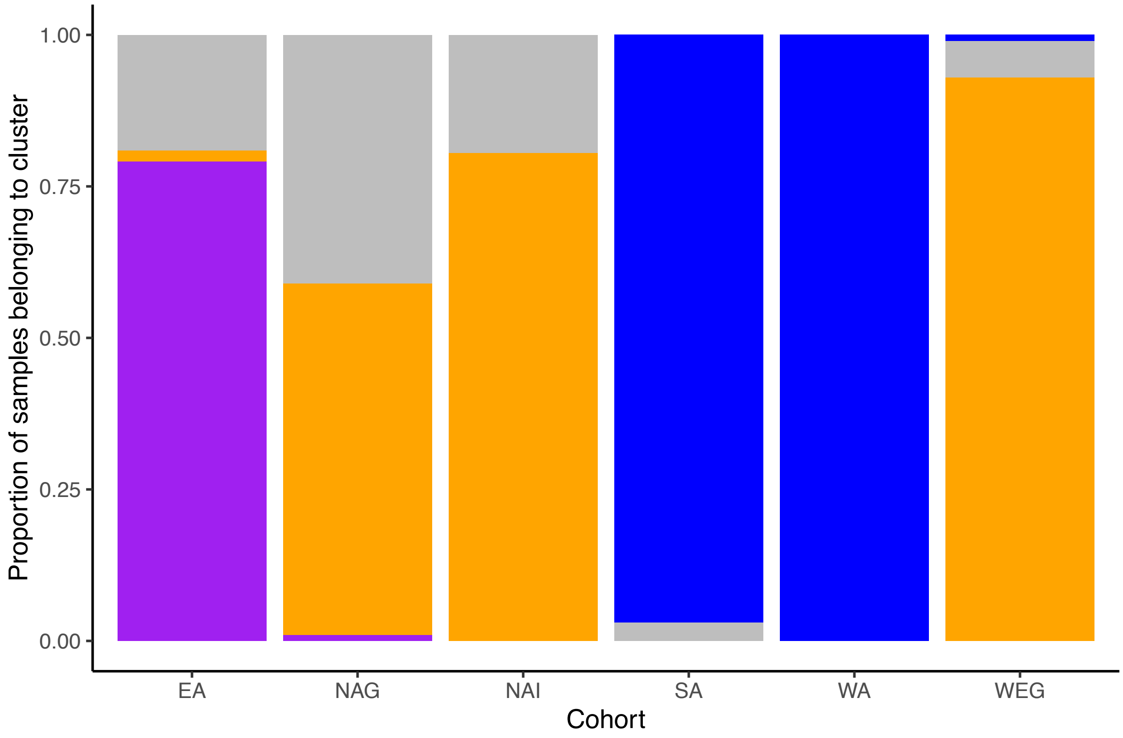


Figure 3.4: Stacked bar plot of the proportions of samples belonging to the hdbscan clusters from each cohort.

3.3.2 Differential abundances of taxa and functions in metagenomic bins in spina bifida microbiome

Having established that conjugative elements could discriminate between populations, we applied the protocol to human health-focused study. Metagenomic assembly and binning of 33 metagenomic samples from 15 mothers who gave birth to children with spina bifida and 18 who gave birth to healthy children yielded a total of 406 medium quality bins; Medium quality bins are defined as having greater than 50% completion and less than 10% redundancy as measured by CheckM using the single-copy bacterial genes [31]. dRep [32] was utilized to remove genomic bins that were duplicated in the assembly process in multiple samples by clustering the bins at a 99% sequence identity. Reads from each sample were mapped to the deduplicated bins to examine the differential abundances of taxa between the spina bifida and control cohorts. A principal component analysis of the abundances of the bins show some degree of separation between the spina bifida and control samples on components 1 and 2 (Figure 3.5). Differential abundance analysis via ALDEx2 [59] on count tables shows that the most differentially abundant bin belongs to the bacteria *Campylobacter hominis*, which had an effect size of 0.96 and was enriched in the mothers who gave birth to infants with spina bifida (Figure 3.6). Secondarily, there was a genomic bin belonging to the genus *Peptoniphilus* that also was enriched in the cases, with an effect size of 0.73. To confirm that *Campylobacter hominis* and *Peptoniphilus* are differentially abundant between the groups, an orthogonal method, MetaPhlAn [60], was used to classify the taxonomy of each read to create a count table of the taxa abundances in each sample. By this method, the two most differentially abundant species are *Campylobacter hominis* and *Peptoniphilus* with effect sizes of 0.86 and 0.79, respectively, which confirms the enrichment in mothers who gave birth to infants with spina bifida.

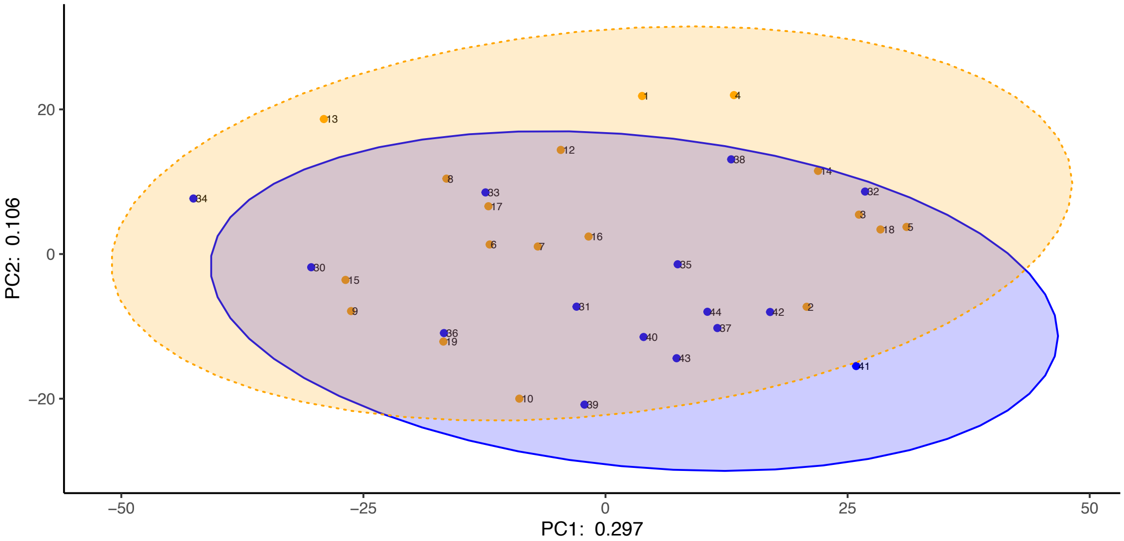


Figure 3.5: Principal component analysis diagram of the 15 spina bifida and 18 control samples for the mapping to the dereplicated bins. Ellipses represent a 95% confidence interval using a multivariate t-distribution about the samples belonging to each group. Orange points are samples from the control group and blue points are for the samples of mothers that gave birth to infants with spina bifida.

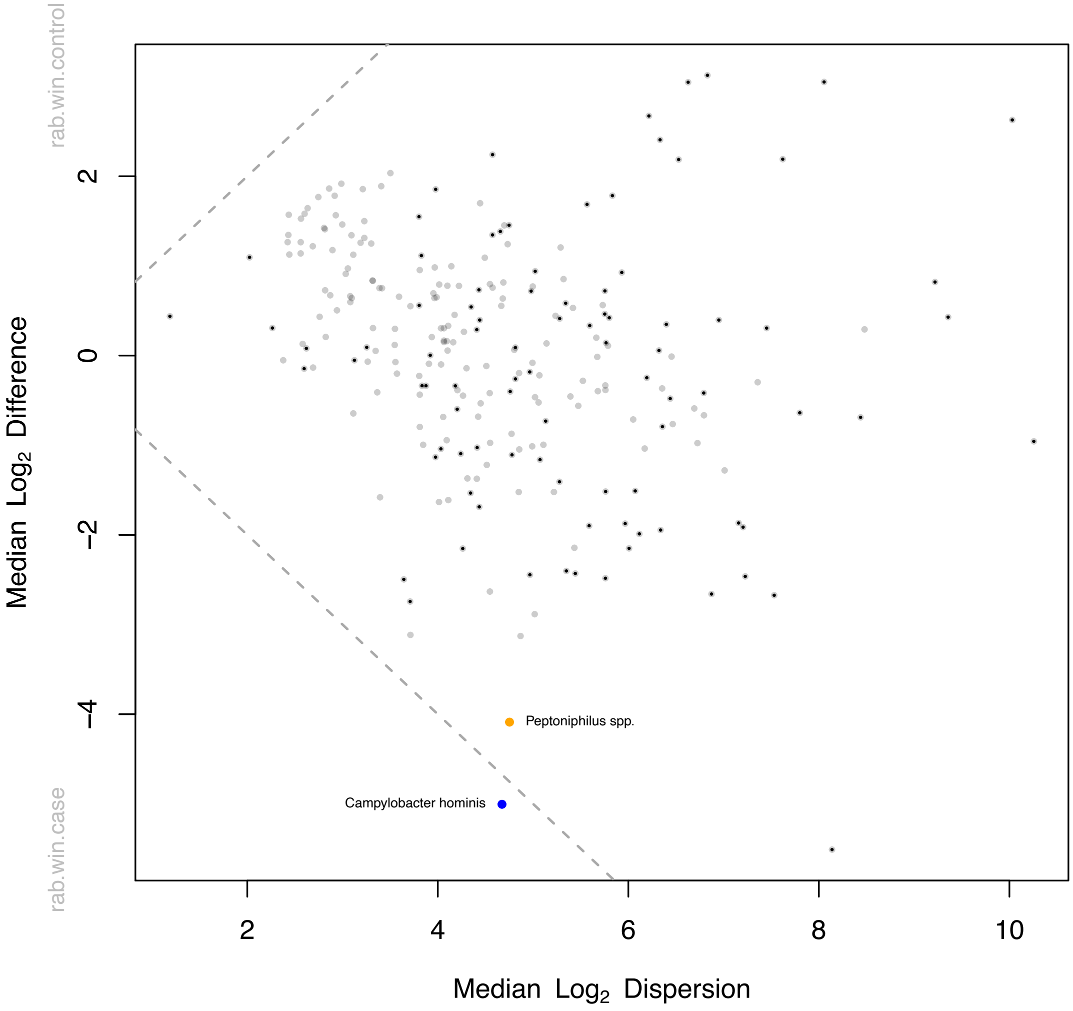


Figure 3.6: Effect plot of the dereplicated genomic bins. Median Log_2 Dispersion is the within group observed variability in the relative abundances of a genomic bin. Median Log_2 Difference is the between group variability in the relative abundances of a genomic bin. Positive values of a Median Log_2 Difference suggest and enrichment in the control samples and negative values suggest an enrichment of a bin in mothers who gave birth to infants with spina bifida (labeled with rab.win.control and rab.win.case on the axis). Bins for *Campylobacter hominis* and *Peptoniphilus*, which are enriched in the case samples, were highlighted with blue and orange points, respectively, on the plot. Dashed lines are the approximate boundary of an effect size of one.

Relative abundances of reads mapping to the ORFs of genes predicted by the gene ontology and Pfam modules of InterproScan were compared between groups as well (Table B.1). From the Pfam annotations, the most enriched gene annotation in the spina bifida group was for a putative serine esterase with an effect size of 1.16. Of note, the annotations for serine esterases are only in genomic bins from the genera *Ruminococcus* and *Porphyromonas*, so the enrichment of this gene is not attributable to the enrichment of *Campylobacter* and *Peptoniphilus* species that was seen from the taxonomic analysis. This is in contrast to the second most enriched Pfam annotation which is of a *Campylobacter* major outer membrane protein (effect size 0.95) and is only found in the genomic bin of *Campylobacter hominis*. The bins containing the gene annotation for serine esterase were of a lower enrichment with effect sizes ranging from 0.27 to 0.6. Additionally, multiple Pfam annotations for STT3 oligosaccharyl transferase genes were found to be enriched at an effect size of 0.87, which is corroborated by the enrichment of the GO term oligosaccharyl transferase activity with an effect size of 0.9. Interestingly, there is also a slight enrichment of the GO term 'regulation of conjugation' in the spina bifida group with an effect size of 0.7.

3.3.3 Differential abundances of conjugative systems in spina bifida microbiome

From the 3079136 total assembled contigs across the 33 samples, 116 of them were identified as being potentially conjugative by selective annotation with pHMMs. The UniRef90 method of conjugative systems identification was intractable for identification of conjugative elements with the computational resources available at the time. To quantify the differences in conjugative systems between groups reads were mapped to the regions localized with conjugative proteins (Figure 3.1). Unlike the bins, there is no apparent separation between groups with mostly overlapping distributions on the principal component diagram (Figure 3.7). Though, with a much lower effect size (0.54) than the genomic bin of *Campylobacter hominis*, the most enriched conjugative system in the mothers who gave birth to an infant with spina bifida belongs to *Campylobacter hominis* (Figure 3.8). Importantly, this contig was not included in the genomic bin of *Campylobacter hominis*, which is in line with the findings of Chapter 2 that the vast majority of conjugative systems are not included in metagenomic bins. Annotations of this contig with the UniRef90 database [34] show that the majority of the open reading frames are of Vir family conjugative proteins, toxin-antitoxin systems, helicases, and a glyoxalase protein.

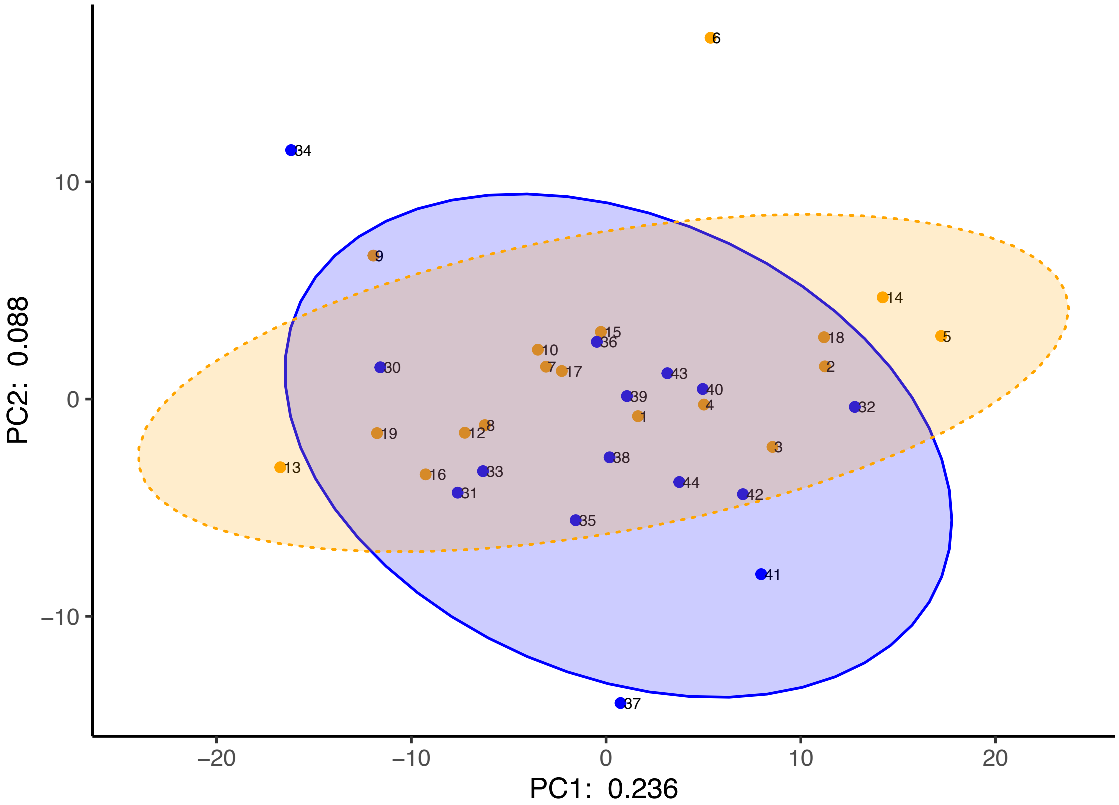


Figure 3.7: Principal component analysis diagram of the 15 spina bifida and 18 control for the relative abundances of the predicted conjugative elements. Ellipses represent a 95% confidence interval using a multivariate t-distribution about the samples belonging to each group. Orange points are samples from the control group and blue points are for the samples of mothers that gave birth to infants with spina bifida.

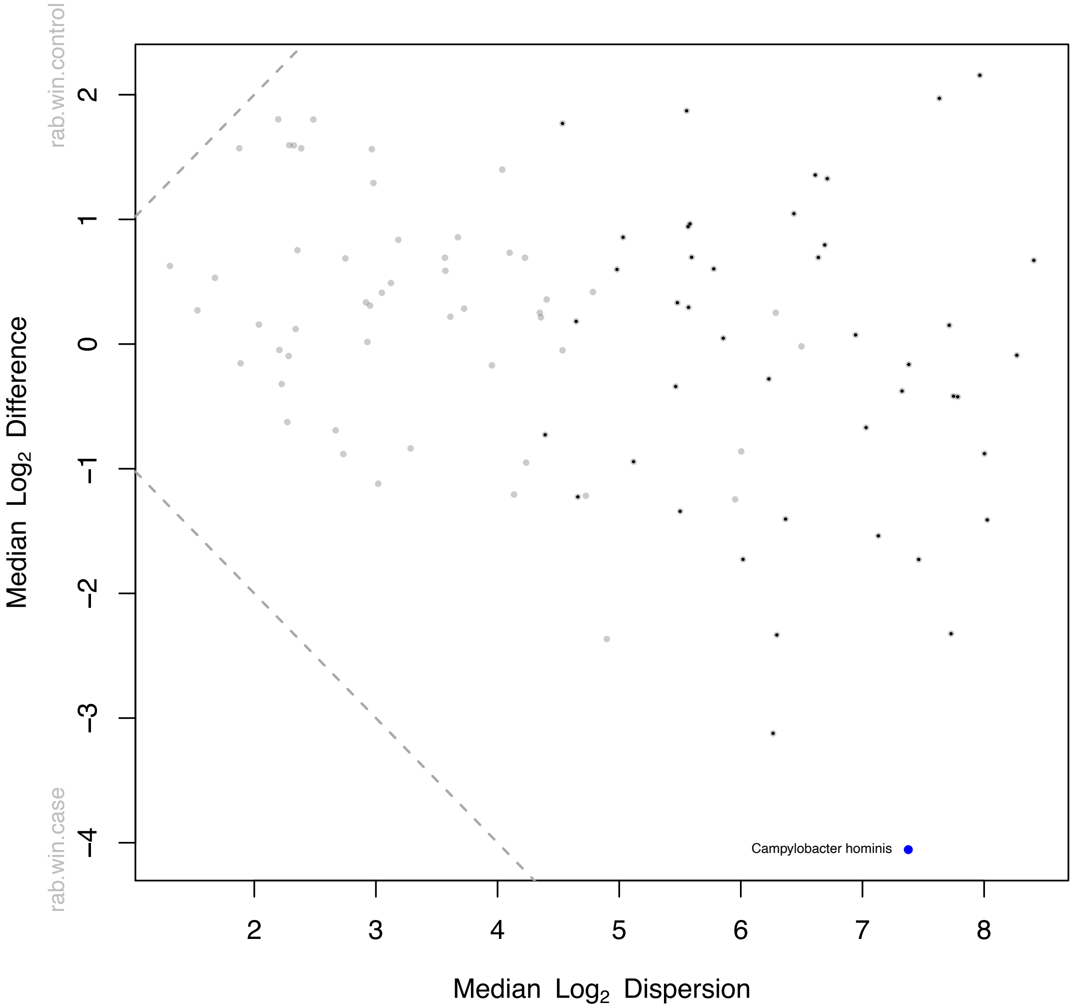


Figure 3.8: Effect plot of the conjugative elements identified with pHMMs. Median Log$_2$ Dispersion is the within group observed variability in the relative abundances of a conjugative element. Median Log$_2$ Difference is the between group variability in the relative abundances of a conjugative element. Positive values of a Median Log$_2$ Difference suggest and enrichment in the control samples and negative values suggest an enrichment of a element in mothers who gave birth to infants with spina bifida (labeled with rab.win.control and rab.win.case on the axis). A conjugative element with the taxonomic assignment of *Campylobacter hominis*, which was enriched in the case samples, was highlighted with a blue on the plot. Dashed lines are the approximate boundary of an effect size of one.

3.4 Discussion

Conjugative systems appear to be differential between cohorts in a similar manner to what has been demonstrated with the overall composition of bacterial species in the human gut between geographically-based cohorts [1]. The starkest difference is the apparent lack of abundance in the North American pre-term infant and European infant datasets of conjugative systems found in the human gut reference set of metagenomic-assembled genomes (Figure 3.2). This is not a result of a lack of conjugative systems, as evidenced by conjugative systems being successfully assembled from the same pre-term infant samples in chapter 2, but rather a bias in the databases towards well-studied cohorts such as the general adult populations from North America or Europe. This logic can be extended to populations that showed a sparser range of conjugative systems in Figure 3.2, such as the West African cohort, which are data from the microbiomes of a hunter gatherer tribe and are unlikely to be well-represented in a Western-biased database. However, these findings do illustrate that conjugative elements are broadly differential between populations. In conjunction with the findings of chapter 2 that conjugative elements as systematically excluded from metagenomic bins, it emphasizes the need to specifically include an analysis of conjugative systems if the goal is to capture the differences in the composition of the microbiome between groups. Although these differences in these elements likely reflects the differences in the overall microbiota composition, there could be biologically-relevant cargo genes that are being excluded from metagenomic bin-centric analyses.

In the gut microbiota of mothers who gave birth to infants with spina bifida, there is a consistent enrichment of the species *Campylobacter hominis* (Figure 3.6), which was differentially abundant using metagenomic binning and an assembly-free method. Also enriched in the genomic bins were annotations for serine esterases that were only found in genomic bins belonging to the genera *Ruminococcus* and *Porphyromonas*. *Campylobacter*, *Ruminococcus*, and *Porphyromonas* are all genera that have been previously associated with obesity and the resultant proinflammatory microbiota [9]. Additionally, in the assembly-free method of differential abundance analysis the species *Peptoniphilus duerdenii* was enriched in the mothers who gave birth to infants with spina bifida.

Peptoniphilus duerdenii and other *Peptoniphilus* species have been previously associated with the development of bloodstream infections, which is a risk factor for spina bifida [61].

Differential abundance analysis of the identified conjugative elements from the assemblies also revealed that the most enriched genetic element in the spina bifida cohort belonged to the species *Campylobacter hominis*. While the annotations of the conjugative element do not reveal an obvious biological mechanism for causing spina bifida, conjugative elements can be re-purposed to vectors to carry CRISPR systems that can selectively kill pathogenic bacteria. [62]. If *Campylobacter hominis* is indeed causing systemic inflammation that leads to mothers giving birth to infants, then having a vector that could serve as a vector for selectively killing *Campylobacter hominis* would be of clinical interest. Recently, one of the conjugative systems identified in the human gut reference genome set has been synthesized *de novo* and shows greater conjugation efficiency to its cognate species than to *Escherichia coli* (personal communications, Thomas Hamilton). Due to limitations in computational resources, the UniRef90 method of annotation and conjugative element identification was not possible for these sample. Given the apparent greater sensitivity of the method, additional differentially abundant conjugative elements may be present in these samples that were not detected by the pHMM method. In future confirmational studies of the association of spina bifida and the maternal gut microbiota, the more robust method should be used.

Transferable genetic elements are clinically relevant, and their systematic exclusion from genomic bins makes precise identification critical in metagenomic analyses. In the microbiomes of mothers who gave birth to infants with spina bifida, conjugative systems present represent clinical interest, but are likely to be omitted in a standard analysis. Expanding the analysis to include conjugative elements revealed a potential vector for modulation of the microbiota to reduce inflammation. Conjugative elements show distinct patterns in geographically-based cohorts, which likely extends to differences between cohorts focused on human health. A meta-analysis of studies of the human gut microbiome and human health may reveal a catalogue of potential vectors that can be used to modulate the composition in states of dysbiosis.

3.5 References

Bibliography

[1] Edoardo Pasolli, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica Armanini, Francesco Beghini, Paolo Manghi, Adrian Tett, Paolo Ghensi, Maria Carmen Collado, Benjamin L Rice, Casey DuLong, Xochitl C Morgan, Christopher D Golden, Christopher Quince, Curtis Huttenhower, and Nicola Segata. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell*, 176(3):649–662.e20, January 2019.

[2] Eman Zakaria Gomaa. Human gut microbiota/microbiome in health and diseases: a review. *Antonie Van Leeuwenhoek*, 113(12):2019–2040, December 2020.

[3] Sean M. P. Bennet, Lena Ohman, and Magnus Simren. Gut microbiota as potential orchestrators of irritable bowel syndrome. *Gut and Liver*, 9(3):318–331, May 2015.

[4] Shahla Abdollahi-Roodsaz, Steven B. Abramson, and Jose U. Scher. The metabolic role of the gut microbiota in health and rheumatic disease: mechanisms and interventions. *Nature Reviews Rheumatology*, 12(8):446–455, August 2016.

[5] Erin R. Lane, Timothy L. Zisman, and David L. Suskind. The microbiota in inflammatory bowel disease: current and therapeutic insights. *Journal of Inflammation Research*, 10:63–73, 2017.

[6] Naga S. Betrapally, Patrick M. Gillevet, and Jasmohan S. Bajaj. Changes in the Intestinal Microbiome and Alcoholic and Nonalcoholic Liver Diseases: Causes or Effects? *Gastroenterology*, 150(8):1745–1755.e3, June 2016.

[7] M. J. Villanueva-Millán, P. Pérez-Matute, and J. A. Oteo. Gut microbiota: a key player in health and disease. A review focused on obesity. *Journal of Physiology and Biochemistry*, 71(3):509–525, September 2015.

[8] Liping Zhao, Feng Zhang, Xiaoying Ding, Guojun Wu, Yan Y. Lam, Xuejiao Wang, Huaqing Fu, Xinhe Xue, Chunhua Lu, Jilin Ma, Lihua Yu, Chengmei Xu, Zhongying Ren, Ying Xu,

Songmei Xu, Hongli Shen, Xiuli Zhu, Yu Shi, Qingyun Shen, Weiping Dong, Rui Liu, Yunxia Ling, Yue Zeng, Xingpeng Wang, Qianpeng Zhang, Jing Wang, Linghua Wang, Yanqiu Wu, Benhua Zeng, Hong Wei, Menghui Zhang, Yongde Peng, and Chenhong Zhang. Gut bacteria selectively promoted by dietary fibers alleviate type 2 diabetes. *Science (New York, N.Y.)*, 359(6380):1151–1156, March 2018.

[9] Emmanuelle Le Chatelier, Trine Nielsen, Junjie Qin, Edi Prifti, Falk Hildebrand, Gwen Falony, Mathieu Almeida, Manimozhiyan Arumugam, Jean-Michel Batto, Sean Kennedy, Pierre Leonard, Junhua Li, Kristoffer Burgdorf, Niels Grarup, Torben Jørgensen, Ivan Brandslund, Henrik Bjørn Nielsen, Agnieszka S Juncker, Marcelo Bertalan, Florence Levenez, Nicolas Pons, Simon Rasmussen, Shinichi Sunagawa, Julien Tap, Sebastian Tims, Erwin G Zoetendal, Søren Brunak, Karine Clément, Joël Doré, Michiel Kleerebezem, Karsten Kristiansen, Pierre Renault, Thomas Sicheritz-Ponten, Willem M de Vos, Jean-Daniel Zucker, Jeroen Raes, Torben Hansen, MetaHIT consortium, Peer Bork, Jun Wang, S Dusko Ehrlich, and Oluf Pedersen. Richness of human gut microbiome correlates with metabolic markers. *Nature*, 500(7464):541–6, August 2013.

[10] Francesca Pistollato, Sandra Sumalla Cano, Iñaki Elio, Manuel Masias Vergara, Francesca Giampieri, and Maurizio Battino. Role of gut microbiota and nutrients in amyloid formation and pathogenesis of Alzheimer disease. *Nutrition Reviews*, 74(10):624–634, October 2016.

[11] Andrew J. Copp, N. Scott Adzick, Lyn S. Chitty, Jack M. Fletcher, Grayson N. Holmbeck, and Gary M. Shaw. Spina bifida. *Nature Reviews. Disease Primers*, 1:15007, April 2015.

[12] Jagteshwar Grewal, Suzan L. Carmichael, Chen Ma, Edward J. Lammer, and Gary M. Shaw. Maternal periconceptional smoking and alcohol consumption and risk for select congenital anomalies. *Birth Defects Research. Part A, Clinical and Molecular Teratology*, 82(7):519–526, July 2008.

[13] Phillip A. Engen, Stefan J. Green, Robin M. Voigt, Christopher B. Forsyth, and Ali Keshavarzian. The Gastrointestinal Microbiome: Alcohol Effects on the Composition of Intestinal Microbiota. *Alcohol Research: Current Reviews*, 37(2):223–236, 2015.

[14] Ziv Savin, Shaye Kivity, Hagith Yonath, and Shoenfeld Yehuda. Smoking and the intestinal microbiome. *Archives of Microbiology*, 200(5):677–684, July 2018.

[15] Mahsa M. Yazdy, Allen A. Mitchell, Simin Liu, and Martha M. Werler. Maternal dietary glycaemic intake during pregnancy and the risk of birth defects. *Paediatric and Perinatal Epidemiology*, 25(4):340–346, July 2011.

[16] P. N. Kirke, A. M. Molloy, L. E. Daly, H. Burke, D. G. Weir, and J. M. Scott. Maternal plasma folate and vitamin B12 are independent risk factors for neural tube defects. *The Quarterly Journal of Medicine*, 86(11):703–708, November 1993.

[17] J. G. Ray and H. J. Blom. Vitamin B12 insufficiency and the risk of fetal neural tube defects. *QJM: monthly journal of the Association of Physicians*, 96(4):289–295, April 2003.

[18] G. M. Shaw, E. M. Velie, and D. M. Schaffer. Is dietary intake of methionine associated with a reduction in risk for neural tube defect-affected pregnancies? *Teratology*, 56(5):295–299, November 1997.

[19] Gary M. Shaw, Richard H. Finnell, Henk J. Blom, Suzan L. Carmichael, Stein Emil Vollset, Wei Yang, and Per M. Ueland. Choline and risk of neural tube defects in a folate-fortified population. *Epidemiology (Cambridge, Mass.)*, 20(5):714–719, September 2009.

[20] C. J. Schorah, J. Wild, R. Hartley, S. Sheppard, and R. W. Smithells. The effect of periconceptional supplementation on blood vitamin concentrations in women at recurrence risk for neural tube defect. *The British Journal of Nutrition*, 49(2):203–211, March 1983.

[21] E. M. Velie, G. Block, G. M. Shaw, S. J. Samuels, D. M. Schaffer, and M. Kulldorff. Maternal supplemental and dietary zinc intake and the occurrence of neural tube defects in California. *American Journal of Epidemiology*, 150(6):605–616, September 1999.

[22] Ian Rowland, Glenn Gibson, Almut Heinken, Karen Scott, Jonathan Swann, Ines Thiele, and Kieran Tuohy. Gut microbiota functions: metabolism of nutrients and other food components. *European Journal of Nutrition*, 57(1):1–24, February 2018.

[23] Melinda A. Engevik, Christina N. Morra, Daniel Röth, Kristen Engevik, Jennifer K. Spinler, Sridevi Devaraj, Sue E. Crawford, Mary K. Estes, Markus Kalkum, and James Versalovic.

Microbial Metabolic Capacity for Intestinal Folate Production and Modulation of Host Folate Receptors. *Frontiers in Microbiology*, 10:2305, 2019.

[24] Tae Hee Kim, Jimao Yang, Pauline B. Darling, and Deborah L. O'Connor. A large pool of available folate exists in the large intestine of human infants and piglets. *The Journal of Nutrition*, 134(6):1389–1394, June 2004.

[25] Petra Louis and Harry J. Flint. Formation of propionate and butyrate by the human colonic microbiota. *Environmental Microbiology*, 19(1):29–41, January 2017.

[26] Amanda J. Cox, Nicholas P. West, and Allan W. Cripps. Obesity, inflammation, and the gut microbiota. *The Lancet. Diabetes & Endocrinology*, 3(3):207–215, March 2015.

[27] Ravinder Nagpal, Tiffany M. Newman, Shaohua Wang, Shalini Jain, James F. Lovato, and Hariom Yadav. Obesity-Linked Gut Microbiome Dysbiosis Associated with Derangements in Gut Permeability and Intestinal Cellular Homeostasis Independent of Diet. *Journal of Diabetes Research*, 2018:3462092, 2018.

[28] Alexandre Almeida, Alex L Mitchell, Miguel Boland, Samuel C Forster, Gregory B Gloor, Aleksandra Tarkowska, Trevor D Lawley, and Robert D Finn. A new genomic blueprint of the human gut microbiota. *Nature*, 568(7753):499–504, April 2019.

[29] Sergey Nurk, Dmitry Meleshko, Anton Korobeynikov, and Pavel A Pevzner. metaSPAdes: a new versatile metagenomic assembler. *Genome Res*, 27(5):824–834, May 2017.

[30] Dongwan D Kang, Feng Li, Edward Kirton, Ashleigh Thomas, Rob Egan, Hong An, and Zhong Wang. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, 7:e7359, 2019.

[31] Donovan H Parks, Michael Imelfort, Connor T Skennerton, Philip Hugenholtz, and Gene W Tyson. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*, 25(7):1043–55, July 2015.

[32] Matthew R Olm, Christopher T Brown, Brandon Brooks, and Jillian F Banfield. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J*, 11(12):2864–2868, December 2017.

[33] Doug Hyatt, Gwo-Liang Chen, Philip F Locascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11:119, March 2010.

[34] Baris E Suzek, Yuqi Wang, Hongzhan Huang, Peter B McGarvey, Cathy H Wu, and UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–32, March 2015.

[35] Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*, 12(1):59–60, January 2015.

[36] Andrew Brantley Hall, Moran Yassour, Jenny Sauk, Ashley Garner, Xiaofang Jiang, Timothy Arthur, Georgia K Lagoudas, Tommi Vatanen, Nadine Fornelos, Robin Wilson, Madeline Bertha, Melissa Cohen, John Garber, Hamed Khalili, Dirk Gevers, Ashwin N Ananthakrishnan, Subra Kugathasan, Eric S Lander, Paul Blainey, Hera Vlamakis, Ramnik J Xavier, and Curtis Huttenhower. A novel Ruminococcus gnavus clade enriched in inflammatory bowel disease patients. *Genome Med*, 9(1):103, November 2017.

[37] Molly K Gibson, Bin Wang, Sara Ahmadi, Carey-Ann D Burnham, Phillip I Tarr, Barbara B Warner, and Gautam Dantas. Developmental dynamics of the preterm infant gut microbiota and antibiotic resistome. *Nat Microbiol*, 1:16024, March 2016.

[38] Fredrik H Karlsson, Valentina Tremaroli, Intawat Nookaew, Göran Bergström, Carl Johan Behre, Björn Fagerberg, Jens Nielsen, and Fredrik Bäckhed. Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature*, 498(7452):99–103, June 2013.

[39] Erica C Pehrsson, Pablo Tsukayama, Sanket Patel, Melissa Mejia-Bautista, Giordano Sosa-Soto, Karla M Navarrete, Maritza Calderon, Lilia Cabrera, William Hoyos-Arango, M Teresita Bertoli, Douglas E Berg, Robert H Gilman, and Gautam Dantas. Interconnected microbiomes and resistomes in low-income human habitats. *Nature*, 533(7602):212–6, May 2016.

[40] Simone Rampelli, Stephanie L Schnorr, Clarissa Consolandi, Silvia Turroni, Marco Sev-

ergnini, Clelia Peano, Patrizia Brigidi, Alyssa N Crittenden, Amanda G Henry, and Marco Candela. Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota. *Curr Biol*, 25(13):1682–93, June 2015.

[41] Krithivasan Sankaranarayanan, Andrew T Ozga, Christina Warinner, Raul Y Tito, Alexandra J Obregon-Tito, Jiawu Xu, Patrick M Gaffney, Lori L Jervis, Derrell Cox, Lancer Stephens, Morris Foster, Gloria Tallbull, Paul Spicer, and Cecil M Lewis. Gut Microbiome Diversity among Cheyenne and Arapaho Individuals from Western Oklahoma. *Curr Biol*, 25(24):3161–9, December 2015.

[42] Junjie Qin, Yingrui Li, Zhiming Cai, Shenghui Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, Wenwei Zhang, Yuanlin Guan, Dongqian Shen, Yangqing Peng, Dongya Zhang, Zhuye Jie, Wenxian Wu, Youwen Qin, Wenbin Xue, Junhua Li, Lingchuan Han, Donghui Lu, Peixian Wu, Yali Dai, Xiaojuan Sun, Zesong Li, Aifa Tang, Shilong Zhong, Xiaoping Li, Weineng Chen, Ran Xu, Mingbang Wang, Qiang Feng, Meihua Gong, Jing Yu, Yanyan Zhang, Ming Zhang, Torben Hansen, Gaston Sanchez, Jeroen Raes, Gwen Falony, Shujiro Okuda, Mathieu Almeida, Emmanuelle LeChatelier, Pierre Renault, Nicolas Pons, Jean-Michel Batto, Zhaoxi Zhang, Hua Chen, Ruifu Yang, Weimou Zheng, Songgang Li, Huanming Yang, Jian Wang, S Dusko Ehrlich, Rasmus Nielsen, Oluf Pedersen, Karsten Kristiansen, and Jun Wang. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55–60, October 2012.

[43] Guoyan Zhao, Tommi Vatanen, Lindsay Droit, Arnold Park, Aleksandar D Kostic, Tiffany W Poon, Hera Vlamakis, Heli Siljander, Taina Härkönen, Anu-Maaria Hämäläinen, Aleksandr Peet, Vallo Tillmann, Jorma Ilonen, David Wang, Mikael Knip, Ramnik J Xavier, and Herbert W Virgin. Intestinal virome changes precede autoimmunity in type I diabetes-susceptible children. *Proc Natl Acad Sci U S A*, 114(30):E6166–E6175, July 2017.

[44] Brian Bushnell, Jonathan Rood, and Esther Singer. BBMerge - Accurate paired shotgun read merging via overlap. *PLoS One*, 12(10):e0185056, 2017.

[45] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for

Illumina sequence data. *Bioinformatics*, 30(15):2114–20, August 2014.

[46] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 9(4):357–9, March 2012.

[47] A Murat Eren, Özcan C Esen, Christopher Quince, Joseph H Vineis, Hilary G Morrison, Mitchell L Sogin, and Tom O Delmont. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*, 3:e1319, 2015.

[48] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017. Publisher: The Open Journal.

[49] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–9, August 2009.

[50] S R Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–63, 1998.

[51] Sean R Eddy. Accelerated Profile HMM Searches. *PLoS Comput Biol*, 7(10):e1002195, October 2011.

[52] Francesco Asnicar, Andrew Maltez Thomas, Francesco Beghini, Claudia Mengoni, Serena Manara, Paolo Manghi, Qiyun Zhu, Mattia Bolzan, Fabio Cumbo, Uyen May, Jon G. Sanders, Moreno Zolfo, Evguenia Kopylova, Edoardo Pasolli, Rob Knight, Siavash Mirarab, Curtis Huttenhower, and Nicola Segata. Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. *Nature Communications*, 11(1):2500, December 2020.

[53] F A Bastiaan von Meijenfeldt, Ksenia Arkhipova, Diego D Cambuy, Felipe H Coutinho, and Bas E Dutilh. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol*, 20(1):217, October 2019.

[54] Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1):D457–462, January 2016.

[55] Philip Jones, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Gift Nuka, Sebastien Pesseat, Antony F. Quinn, Amaia Sangrador-Vegas, Maxim Scheremetjew, Siew-Yit Yong, Rodrigo Lopez, and Sarah Hunter. InterProScan 5: genome-scale protein function classification. *Bioinformatics (Oxford, England)*, 30(9):1236–1240, May 2014.

[56] Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik L L Sonnhammer, Silvio C E Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, Robert D Finn, and Alex Bateman. Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1):D412–D419, January 2021.

[57] The Gene Ontology Consortium, Seth Carbon, Eric Douglass, Benjamin M Good, Deepak R Unni, Nomi L Harris, Christopher J Mungall, Siddartha Basu, Rex L Chisholm, Robert J Dodson, Eric Hartline, Petra Fey, Paul D Thomas, Laurent-Philippe Albou, Dustin Ebert, Michael J Kesling, Huaiyu Mi, Anushya Muruganujan, Xiaosong Huang, Tremayne Mushayahama, Sandra A LaBonte, Deborah A Siegele, Giulia Antonazzo, Helen Attrill, Nick H Brown, Phani Garapati, Steven J Marygold, Vitor Trovisco, Gil dos Santos, Kathleen Falls, Christopher Tabone, Pinglei Zhou, Joshua L Goodman, Victor B Strelets, Jim Thurmond, Penelope Garmiri, Rizwan Ishtiaq, Milagros Rodríguez-López, Marcio L Acencio, Martin Kuiper, Astrid Lægreid, Colin Logie, Ruth C Lovering, Barbara Kramarz, Shirin C C Saverimuttu, Sandra M Pinheiro, Heather Gunn, Renzhi Su, Katherine E Thurlow, Marcus Chibucos, Michelle Giglio, Suvarna Nadendla, James Munro, Rebecca Jackson, Margaret J Duesbury, Noemi Del-Toro, Birgit H M Meldal, Kalpana Paneerselvam, Livia Perfetto, Pablo Porras, Sandra Orchard, Anjali Shrivastava, Hsin-Yu Chang, Robert Daniel Finn, Alexander Lawson Mitchell, Neil David Rawlings, Lorna Richardson, Amaia Sangrador-Vegas, Judith A Blake, Karen R Christie, Mary E Dolan, Harold J Drabkin, David P Hill, Li Ni, Dmitry M Sitnikov, Midori A Harris, Stephen G Oliver, Kim Rutherford, Valerie Wood, Jaqueline Hayles, Jürg Bähler, Elizabeth R Bolton, Jeffery L De Pons, Melinda R Dwinell, G Thomas Hayman, Mary L Kaldunski, Anne E Kwitek, Stanley J F Laulederkind, Cody

Plasterer, Marek A Tutaj, Mahima Vedi, Shur-Jen Wang, Peter D'Eustachio, Lisa Matthews, James P Balhoff, Suzi A Aleksander, Michael J Alexander, J Michael Cherry, Stacia R Engel, Felix Gondwe, Kalpana Karra, Stuart R Miyasato, Robert S Nash, Matt Simison, Marek S Skrzypek, Shuai Weng, Edith D Wong, Marc Feuermann, Pascale Gaudet, Anne Morgat, Erica Bakker, Tanya Z Berardini, Leonore Reiser, Shabari Subramaniam, Eva Huala, Cecilia N Arighi, Andrea Auchincloss, Kristian Axelsen, Ghislaine Argoud-Puy, Alex Bateman, Marie-Claude Blatter, Emmanuel Boutet, Emily Bowler, Lionel Breuza, Alan Bridge, Ramona Britto, Hema Bye-A-Jee, Cristina Casals Casas, Elisabeth Coudert, Paul Denny, Anne Estreicher, Maria Livia Famiglietti, George Georghiou, Arnaud Gos, Nadine Gruaz-Gumowski, Emma Hatton-Ellis, Chantal Hulo, Alexandr Ignatchenko, Florence Jungo, Kati Laiho, Philippe Le Mercier, Damien Lieberherr, Antonia Lock, Yvonne Lussi, Alistair MacDougall, Michele Magrane, Maria J Martin, Patrick Masson, Darren A Natale, Nevila Hyka-Nouspikel, Sandra Orchard, Ivo Pedruzzi, Lucille Pourcel, Sylvain Poux, Sangya Pundir, Catherine Rivoire, Elena Speretta, Shyamala Sundaram, Nidhi Tyagi, Kate Warner, Rossana Zaru, Cathy H Wu, Alexander D Diehl, Juancarlos N Chan, Christian Grove, Raymond Y N Lee, Hans-Michael Muller, Daniela Raciti, Kimberly Van Auken, Paul W Sternberg, Matthew Berriman, Michael Paulini, Kevin Howe, Sibyl Gao, Adam Wright, Lincoln Stein, Douglas G Howe, Sabrina Toro, Monte Westerfield, Pankaj Jaiswal, Laurel Cooper, and Justin Elser. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research*, 49(D1):D325–D334, January 2021.

[58] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000.

[59] Andrew D Fernandes, Jennifer Ns Reid, Jean M Macklaim, Thomas A McMurrough, David R Edgell, and Gregory B Gloor. Unifying the analysis of high-throughput sequencing datasets:

characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, 2:15, 2014.

[60] Duy Tin Truong, Eric A Franzosa, Timothy L Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods*, 12(10):902–3, October 2015.

[61] G. M. Shaw, K. Todoroff, E. M. Velie, and E. J. Lammer. Maternal illness, including fever and medication use as risk factors for neural tube defects. *Teratology*, 57(1):1–7, January 1998.

[62] Thomas A Hamilton, Gregory M Pellegrino, Jasmine A Therrien, Dalton T Ham, Peter C Bartlett, Bogumil J Karas, Gregory B Gloor, and David R Edgell. Efficient inter-species conjugative transfer of a CRISPR nuclease for targeted bacterial killing. *Nat Commun*, 10(1):4544, October 2019.

Chapter 4

Pangenomic analysis of five new strains belonging to the manganese-oxidizing genus *Manganitrophus*

4.1 Introduction

Bacteria harbour a great diversity of metabolic pathways that are being discovered continually as previously unculturable bacteria are cultured using novel methods or sequenced metagenomically from environmental samples. Novel metabolic pathways from extreme environments have become of interest as they can play a key role in efforts to engineer communities of bacteria for the purposes of bioremediation. Through anthropogenic pollution of the environment by industrial processes, large quantities of heavy metals enter the environment causing broad ranges of toxicity to plants and animals. Technology and methods have been developed that are abiotic, however, they are expensive and inefficient in comparison to the use of biological processes [1]. Plants have been utilized to remove heavy metals from the environment, a process called phytoremediation [2], but the slow doubling time of plants make them difficult to scale and a sub-optimal candidate for remediation of the environment. Bacteria have quick doubling times and have been found to be able to counter the pollution of heavy metals such as lead, chromium, cadmium, and arsenic [3, 4, 5, 6, 7]. The cell surfaces of some species of bacteria are rich in transporters that allow heavy metals to cross the cell membrane [8]. This permits bacteria to uptake large quantities of heavy metals into their cytoplasm and out of the environment, a process known as bioaccumulation [9]. The machinery to transport and bioaccumulate heavy metals are often carried on plasmids by bacteria that can be dispersed throughout the environment to other bacteria [10]. Once inside the cell, the bacteria can metabolize the toxic heavy metal into a less toxic product through biotransformation, bioleaching, or biomineralization [11, 12]. In addition, bacteria are able to secrete molecules such as extracellular polymeric substances [13, 14], metallothioneins [15], or siderophores [16] that can sequester the heavy metals.

Heavy metal waste is commonly produced through oil refinement, which also expels large volumes of contaminant hydrocarbons into the waste water and it has been hypothesized that bacteria may have the metabolic capabilities of degrading these complex hydrocarbons. Indeed, a number of bacteria have been identified as being capable of using polyaromatic hydrocarbons (PAHs) as a sole energy source [17]. Similarly to heavy metal removal, the use of microbes to remediate oil spills is believed to be more environmentally friendly and more efficient [18]. Pathways for bioremediation of heavy metals, such as biomineralization, are also used for the sequestration of PAHs like naphthalene [19]. More well-studied than biomineralization of PAHs is the oxidation and degradation of PAHs where the bacteria utilize the high-energy bonds of hydrocarbons to fuel their metalbolism. Species from the genera *Sphingomonas*, *Rhodococcus*, and *Alcanivorax* are well-known metabolizers of PAHs and other hydrocarbons [20, 21, 22]. Similar to heavy metal bioremediation, the genes critical for the metabolism and sequestration of hydrocarbons are commonly found on plasmids [23]. Metabolism of PAHs is primed by the aromatic ring cleavage with oxygenases and dioxygenases, which results in products that can be metabolized by the tricarboxylic acid cycle [24, 25, 26]. Importantly, hydrocarbon degradation and heavy metal contamination may potentially be linked. Ammonia oxidation is inhibited by environmental chromium contamination [27], and general impairment of metabolism could extend to hydrocarbon oxidation as well. This may be a contributing reason as to why microbial consortium are much more effective at metabolizing hydrocarbons than individual bacteria [28]. There are members of the community that can sequester and metabolize heavy metals present in the wastewater, which allows the hydrocarbon-degrading bacteria to perform metabolism uninhibited. Efforts have been made to engineer *Deinococcus radiodurans* that can metabolize both heavy metal and hydrocarbons [29], but because cometabolism is so important for PAH degradation [30], it would still likely be more efficient to introduce a consortia than an optimized isolate.

A clade of bacteria that has garnered interest for bioremediation, and the unique metabolic pathways present in its member species, has been the phylum *Nitrospirota* (formerly known as *Nitrospirae*). The primary bioremediation function that is found in the *Nitrospirota* phylum is complete ammonia oxidation in the genus *Nitrospira* [31, 32, 33, 34, 35]. Though some *Nitrospira* species are only able to oxidize nitrite, many are able to oxidize both nitrite and ammonia [36]. This capacity to completely oxidize ammonia is of environmental interest for bioremediation because ammonia runoff from industrial farming causes toxic algal blooms, which may be preventable with the metabolism of ammonia by *Nitrospira*. Other interesting metabolic functions of *Nitrospirota* species include the reduction of sulfate [37, 38, 39, 40, 41], disproportionation of inorganic sulfur [42], oxidation of iron [43], and the capactity to align themselves with the earth's magnetic fields [44]. Recently, a novel genus of *Nitrospirota* was discovered that has the ability to oxidize manganese as its sole energy source [45, 46]. Through the reverse tricarboxylic acid cycle, members of the genus *Manganitrophus* are able to utilize the potential of manganese electrons to drive their autotrophic growth [45]. Given that manganese is a very common element in the earth's crust, it follows that there should be a widespread abundance of bacteria that are able to utilize this biochemical process to fuel their growth.

Here, we present five metagenomic-assembled genomes from a novel candidate species belonging to the genus *Manganitrophus*. These novel strains were assembled from long-read metagenomic data collected from biofilms growing on granular activated charcoal filters from an oil refinery's wastewater treatment plant in Sarnia, Canada. The novel strains can be placed within the same genus as the recently described *Candidatus Manganitrophus noduliformans* and *Candidatus Manganitrophus morganii* species through a phylogenetic tree based on the single-copy core genes. The candidate species share the same gene clusters related to manganese oxidation and hydrocarbon degradation, but the novel strains also have a increased genetic capacity to transport heavy metals across their cellular membrane. By publishing these genomes and comparing them to the other members of their genus, insight into how these manganese-oxidizing bacteria adapt to their harsh environment has been gained.

4.2 Methods

4.2.1 Isolation and sequencing of DNA from granular activated charcoal biofilms

To a 50mL falcon tube containing 10 g of granular activated carbon (GAC) sample, 10 mL of lysis buffer (10 mM Tris-HCl, 100 mM NaCl, 25 mM EDTA, 0.5% (w/v) SDS) and 2.5 mg of lysozyme was added and mixed by slowly rotating the tube horizontally to minimize granule movement. The mixture was incubated for 1 hour at 37°C, while slowly mixing every 15 minutes. 5µL of RNAse (20 mg/mL) was subsequently added to the mix and incubated at 37°C for another 30 minutes, with slow mixing every 15 minutes. Finally, 100 µL of Proteinase K (800 units/mL) was added and the mixture was incubated at 57°C for 1 hour and 30 minutes. Following incubation, the sample was spun at 3000g for 5 minutes to spin down small GAC particles that are suspended in the lysate before being decanted into a new 50 mL falcon tube. Then, 1 volume of 25:24:1 phenol:chloroform:isomayl alcohol was added and mixed by rocking for 8 minutes, before being spun down at 3000g for 3 minutes. The aqueous phase was transferred to a new 50 mL falcon tube using wide bore pipette tips for a chloroform wash. To the aqueous phase, 1 volume of chloroform was added and mixed by gentle rocking for 8 minutes, before being spun down at 3000g for 3 minutes. The aqueous phase was transferred to a new tube and the chloroform wash step was repeated. Next, 1/10 volume of sodium acetate (3 M, pH 5.2) was added and mixed by inversion. 2

volumes of ice cold 100% ethanol was then added and mixed by inversion. High molecular weight DNA that precipitated was spooled out into a 1.5 mL Eppendorf tube containing 200 µL of 75% ethanol using a Pasteur pipette that was melted into hook. The DNA was spun down at 10000g for 3 minutes. The ethanol was removed without disturbing the pellet and another 200 µL of 75% ethanol was added, before being spun at 10000 g for 3 minutes. After spinning, the ethanol was removed and the pellet was left to air dry for 2 minutes. Depending on the size of the pellet, the DNA was resuspended with 200 µL to 1 mL of Tris-HCl (10 mM, pH 8) overnight.

The recovered DNA was size selected to retain fragments greater than 40kb using the Circulomics Short Read Eliminator XL kit according to the manufacturer's protocol. The sequencing library was prepared from the size selected DNA using the Oxford Nanopore ligation sequencing kit (SQK-LSK110) and its associated protocol, with a few changes. In the DNA repair and end prep steps, the DNA was incubated in the thermal cycler for 20 minutes at 20 °C and 20 minutes at 65 °C. Instead of AMPure XP beads, Omega Bio-Tek Mag-Bind beads were used in clean up steps. The library was sequenced on Oxford Nanopore's MinION platform, using a 9.4.1 flow cell. Basecalling was performed with Guppy 5.0.16 in super accuracy mode (dna_r9.4.1_450bps_sup).

4.2.2 Metagenomic assembly and identification of strain GAC1

Prior to assembly, reads were trimmed using Nanofilt [47] with settings '-q 10 -l 500' to filter out reads with an average quality score below 10 or a read length less than 500 base pairs. The filtered reads were then metagenomically assembled using Flye [48] with settings '–nano-raw –meta -g 5m'. The reads were aligned to the contigs using Minimap2 [49] and then the alignments were filtered using GERENUQ [50] to avoid polishing with poorly aligned sequences or chimeric reads erroneously generated during sequencing. Polishing using Racon [51] with settings '-m8 -x -6 -g -8 -w 500' and subsequently using Medaka [52] with settings '-m r941_min_sup_g507' was performed to correct sequence errors in the assembly. Taxonomic assignment of all contigs present in the polished assembly was performed using CAT [53] with the database release 'CAT_prepare_20210107'. Conjugative elements were identified using a previously established method using profile hidden

Markov model integration in Anvi'o [54, 55]. Contigs from the first assembly with the taxonomic assignment of *Candidatus Manganitrophus noduliformans* were extracted for further analysis.

4.2.3 Mapped assembly of subsequent strains

For the remaining four strains of *Manganitrophus*, the genomes could not be resolved from a standard metagenomic assembly. To successfully assemble these strains, the reads from these samples were aligned to the genome of the GAC1 strain using Minimap2 [49], filtered using GERENUQ [50] to select for reads with full-length alignments, then assembled using Flye [56] on standard settings. For strains 3 and 5, which did not yield circular, single-contig assemblies through this mapped assembly strategy, contigs were selected using Bandage [57] that formed a circular subgraph on the assembly graph. Subsequently, we repeated this process using the circular conjugative plasmid identified belonging to the genus *Manganitrophus* from sample GAC1. Plasmid sequences were retrieved for plasmid-specific analyses from *Candidatus Manganitrophus noduliformans* and *Candidatus Manganitrophus morganii* SA1 by aligning their published genomes to the GAC1 plasmid sequence and selecting for strongly aligning sequences through manual inspection of the alignment length and match scores. Circularization of chromosomes and plasmids was confirmed by aligning reads using minimap2 [49] to the contigs and identifying a read tiling path along the sequence with reads over 5000bp in length that overlap by at least 500bp–with an additional read that maps to the start and end of the genetic sequence.

4.2.4 Phylogenomics and genome annotation

Three *Manganitrophus* genomes were accessed from the NCBI database for direct pangenomic comparison from BioProject IDs PRJNA562312 and PRJNA776098. Additionally, 15 reference-level genomes of the phylum *Nitrospirota* were downloaded from the NCBI Assembly Database [58] to help construct the unrooted tree. CheckM [59] was run on all genomes to assess completion, contamination, GC%, and to generate a concatenated protein alignment of the single copy core bacterial genes for phylogenetic analysis. Construction of the phylogenetic tree was performed using

FastTree [60] integration in Anvi'o [55] and visualization of the unrooted tree was conducted with iTOL [61]. Average nucleotide identity of the 8 *Manganitrophus* genomes was performed using pyANI [62]. FastANI [63] was utilized to visualize the alignment between the GAC1 conjugative element and the published *Candidatus Manganitrophus noduliformans* genome. The eight *Manganitrophus* genomes were imported into Anvi'o, annotated with NCBI COGs [64] and KEGG [65], and built into a pangenome of gene clusters. Gene cluster bins were assigned manually on the basis of presence in the species and strains. The genomes were further annotated by aligning the open reading frames predicted with Prodigal to the UniRef50 [66] and AromaDeg [67] databases using the Diamond [68, 69] aligner with settings '–id 50 –query-cover 50 -f 6 –salltitles -p 10 -k 1' and '–very-sensitive -f 6 –salltitles -p 10 -k 1', respectively. Additionally, the predicted protein sequences were queried to the CANT-HYD [70] database using HMMER3 [71] and CRISPR arrays were predicted in the nucleotide sequences using PilerCR [72]. Pangemomic clustering and annotation was repeated for the separated plasmid sequences. Plasmid origins of replication were identified using the Ori-Finder2 web tool [73]. Plasmids were additionally annotated through an alignment to UniRef90 database [66] to identify proteins involved in conjugation.

4.3 Results

4.3.1 Assembly of five novel genomes belonging to the Manganitrophus genus from a biofilm growing on charcoal filters

The genus *Manganitrophus* is a recently described clade of bacteria from the phylum *Nitrospirota* that has been shown to have the capability to oxidize Manganese (II) carbonates as a sole source of energy [45]. The first genome published was of *Candidatus Manganitrophus noduliformans* which was serendipitously enriched from Californian tap water and the next two genomes were of *Candidatus Manganitrophus morganii* SA1 and SB1 which were enriched from samples collected from South Africa and California growing on a rock surface and an iron oxide mat, respectively. In this study, the novel genomes were identified via metagenomic assembly from a biofilm growing on a granular activated charcoal (GAC) filter from Suncor Energy's Sarnia, Ontario, Canada oil refinery. High molecular weight DNA was isolated from the biofilm and sequenced on an Oxford Nanopore Minion (9.41 flow cell). From the first sample, a circularized genomic element was taxonomically assigned as belonging to the clade *Candidatus Manganitrophus noduliformans*. However, genomes from subsequent samples were not able to be successfully assembled into circularized genomes using metaFlye due to the complexity of the communities and the apparent intrasample heterogeneity. To obtain circularized or highly contiguous genomes of the novel species for the other four samples, the long reads were aligned to the complete and circular genome of sample GAC1, filtered, and reassembled as a 'pseudo-isolate', which yielded two additional circularized genomes and two highly contiguous genomes (Table 4.1). Circularization of chromosomes from samples GAC1, GAC2, and GAC4 were confirmed through an overlapping read tiling path.

Table 4.1: Genome quality and summary metrics of the newly assembled and previously published *Manganitrophus* strains including plasmid sequences as evaluated by CheckM.

	Completion	Contamination	Strain Heterogeneity	GC Content	Size (bp)	# Contigs
M. noduliformans	98.58	4.55	0	56.41	5171380	22
M. morganii SA1	97.67	3.64	0.00	56.10	4257136	8
M. morganii SB1	97.67	3.64	0.00	56.20	4287287	1
M. GAC1	95.40	2.73	0.00	56.13	4562547	2
M. GAC2	96.58	3.64	0.00	56.10	4602354	6
M. GAC3	89.38	16.74	78.26	56.29	5306623	31
M. GAC4	90.93	2.78	0.00	56.37	4553225	3
M. GAC5	96.49	42.50	84.13	56.05	5615601	18

4.3.2 Assembly of five novel conjugative plasmids

In addition to the 5 chromosomal genomes of a novel *Candidatus Manganitrophus bacteria*, five conjugative plasmids were also identified. A circularized genetic element of roughly 330kb in length was identified from the first GAC sample that was assigned the same taxonomy as the chromosomal DNA. Ori-Finder2 [73] predicted bacterial origin of replication sequences within the chromosome of GAC1 as well as within the separate circularized sequence presumed to be a plasmid. To confirm that this circular genetic element found within the metagenome was indeed a plasmid belonging to the novel *Candidatus Manganitrophus* species, the plasmid sequence was aligned to the published *Candidatus Manganitrophus noduliformans* genome with FastANI, which showed that multiple contigs from the published isolate genome were of plasmid origin (Figure A.1). This finding suggests that this large, conjugative plasmid is well-maintained within this genus. Indeed, the novel plasmid also aligned with contigs from the published *Candidatus Manganitrophus morganii* SA1 genome. However, the published genome of *Candidatus Manganitrophus morganii* SB1 only contained a single contig that did not have any alignments to the plasmid. For samples 2-5, circularized contigs could not be obtained from a standard metagenomic assembly, so the mapped assembly strategy was used to obtain two additional circularized plasmids and two highly contiguous plasmid sequences (Table 4.2). Circularization of plasmids from samples GAC1, GAC3, and GAC5 were confirmed through an overlapping read tiling path. All of the assembled and identified plasmids had multiple open reading frames annotated by UniRef90 as components for type IV conjugative transfer. Additionally, carried as cargo on these plasmids are a multitude of genes related to the transport and biotransformation of metals such as cobalt, copper, and mercury.

Table 4.2: Summary information of identified conjugative plasmids. Summary information on GC content, size, and contigs were obtained from CheckM and the number of conjugation genes was determined by annotation of the genome with the UniRef50 database.

	GC Content	Size (bp)	# Conjugation genes	# Contigs
M. noduliformans	54.71	342241	5	4
M. morganii SA1	53.55	288825	5	1
M. GAC1	53.57	328837	8	1
M. GAC2	53.85	364954	6	5
M. GAC3	53.74	330790	6	1
M. GAC4	53.74	318819	6	2
M. GAC5	53.74	295087	7	1

4.3.3 Placement of novel genomes into the genus Manganitrophus

The first indication that the novel genomes belonged to the genus *Manganitrophus* was the taxonomic assignment provided by the Contig Annotation Tool of *Candidatus Manganitrophus noduliformans*. All representative genomes from the *Nitrospirota* phylum were downloaded from the

NCBI assembly database to confirm the placement of the newly assembled genomes in the genus *Manganitrophus*. The phylogenetic tree was based on the protein alignments of the concatenated single-copy core genes predicted by CheckM and constructed using FastTree integration in Anvi'o. In an unrooted phylogenetic tree, the five novel genomes were very closely associated on a branch with *Candidatus Manganitrophus noduliformans*, *Candidatus Manganitrophus morganii* SA1 and *Candidatus Manganitrophus morganii* SB1 (Figure 4.1). To further elucidate the fine-grain relationships between the *Manganitrophus* species, the eight genomes were compared using pyANI to obtain average nucleotide identity values. The five novel strains show high similarity to one another, but form a separate branch when hierarchically clustered with *M. morganii* and *M. noduliformans* (Figure 4.2). Based on the phylogeny and ANI clustering, we propose our assembled genomes belonging into a novel, uncharacterized species in the *Manganitrophus* genus.

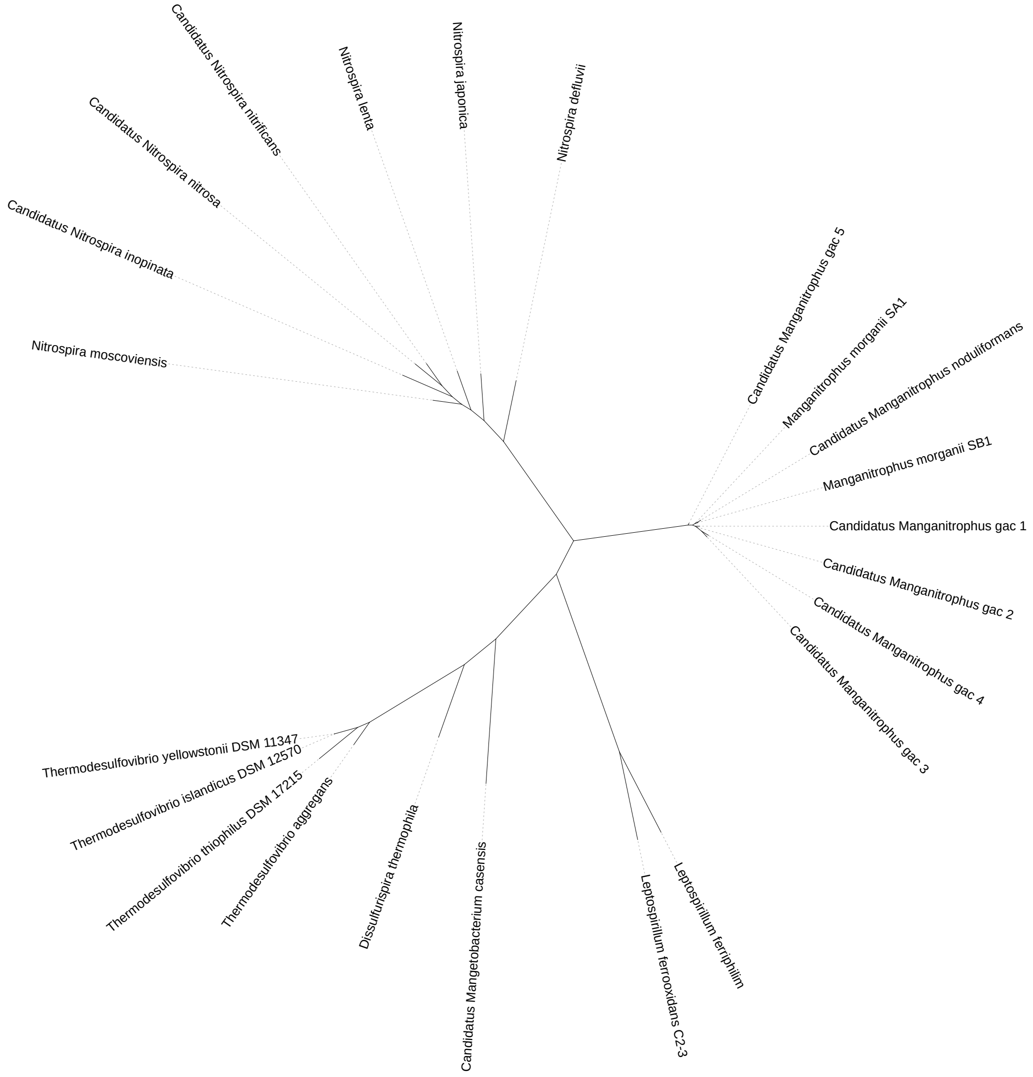

Candidatus Nitrospira nitrificans
Nitrospira lenta
Nitrospira japonica
Nitrospira defluvii
Candidatus Nitrospira nitrosa
Candidatus Nitrospira inopinata
Nitrospira moscoviensis
Candidatus Manganitrophus gac 5
Manganitrophus morganii SA1
Candidatus Manganitrophus noduliformans
Manganitrophus morganii SB1
Candidatus Manganitrophus gac 1
Candidatus Manganitrophus gac 2
Candidatus Manganitrophus gac 4
Candidatus Manganitrophus gac 3
Thermodesulfovibrio yellowstonii DSM 11347
Thermodesulfovibrio islandicus DSM 12570
Thermodesulfovibrio thiophilus DSM 17215
Thermodesulfovibrio aggregans
Dissulfurispira thermophila
Candidatus Mangetobacterium casensis
Leptospirillum ferrooxidans C2-3
Leptospirillum ferriphilim

Figure 4.1: Unrooted phylogenetic tree of the phylum *Nitrospirota* generated by FastTree using alignments of bacterial single copy genes produced by CheckM.

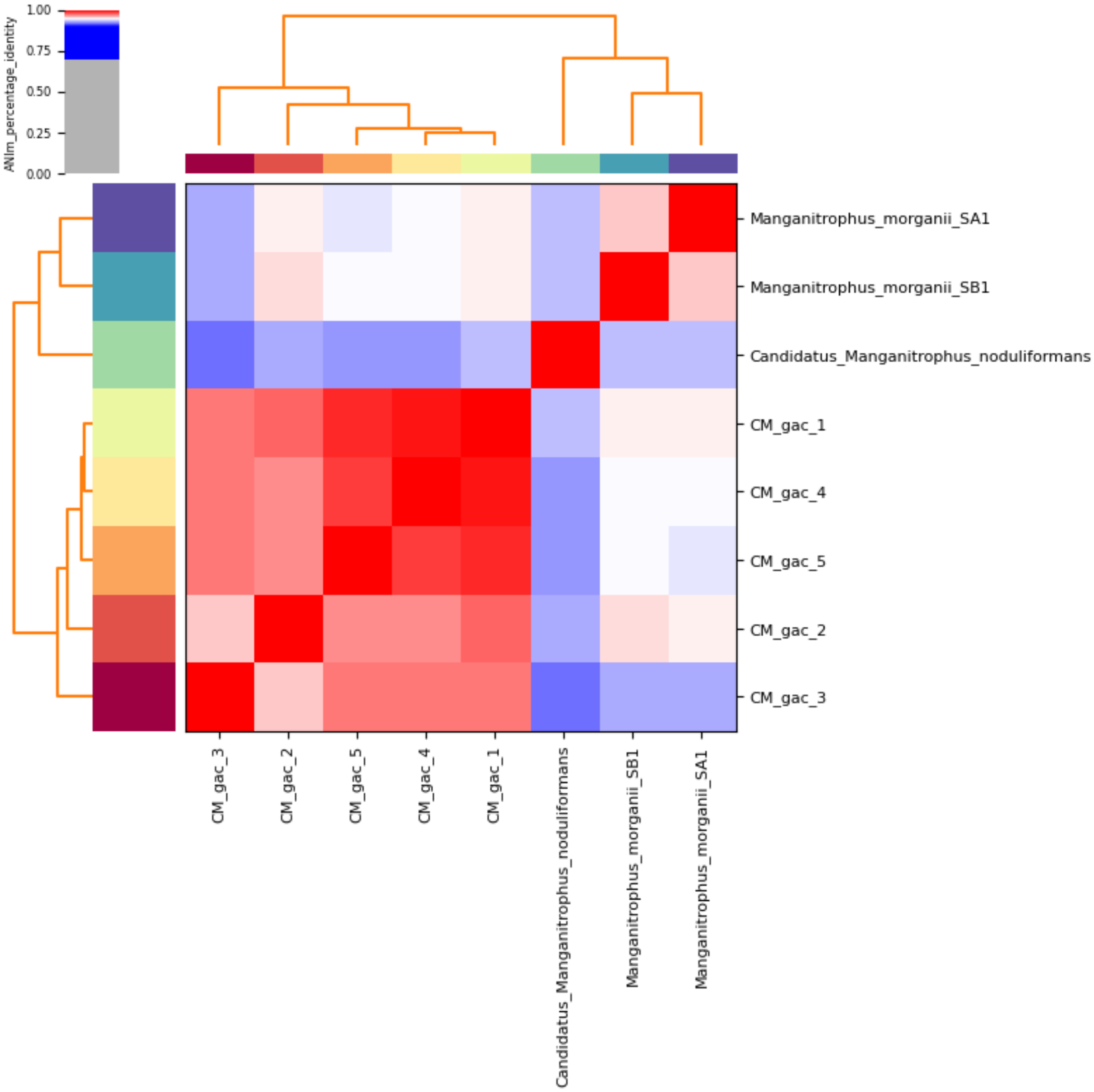


Figure 4.2: Average nucleotide identity heat map and hierarchical clustering of the eight Manganitrophus genomes calculated by pyANI.

4.3.4 Establishing the core and accessory genes in the *Manganitrophus* pangenome

The Anvi'o suite of pangenomic tools were used to cluster orthologous genes within the eight *Manganitrophus* genomes, assign the gene clusters into bins based on presence in certain genomes, and to annotate the gene clusters with KEGG and NCBI COG. The result of the gene cluster can be visualized with Anvi'o (Figure 4.3). The largest group of gene clusters that are found consistently in all eight of the strains are referred to as the stable core. It is unlikely that the orthologues present in these gene clusters are responsible for the differences between the species. The stable core contains all the single-copy core genes that are identified by CheckM. There is also another large bin that is labelled as the ‘Unstable Core’. These gene clusters are mostly found in all of the strains, but contain some degree of variation and absence in strains that do not follow the species boundaries. These two bins of gene clusters form what is referred to as the core gene set and the remaining gene clusters form the accessory gene set.

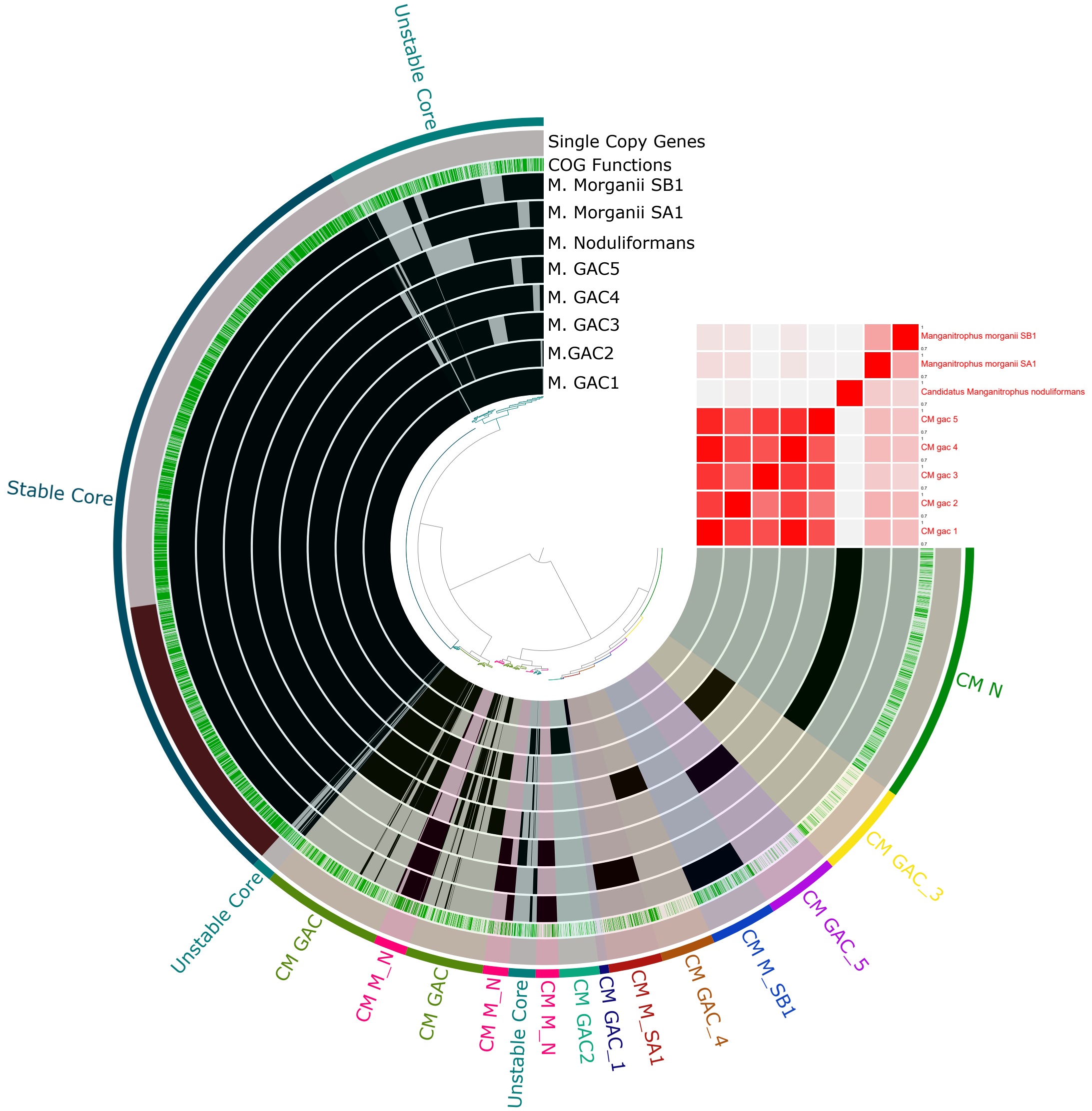


Figure 4.3: Pangenomic visualization of the full genomic content of the genus *Manganitrophus*. Black shading of circular track represents the presence of a gene cluster in a given strain. Bins of gene clusters on outer ring were determined manually to denote the pattern of presence in the groups of strains or strains. CM GAC are genes clusters primarily present in the GAC strains, CM M_N are primarily in the *Manganitrophus morganii* and *noduliformans* strains, and the CM N, M_SB1, M_SA1, GAC_1, GAC2, GAC_3, GAC_4, GAC_5 groupings are gene clusters unique to the strains the groupings are named after. The intensity of the red shading represents the average nucleotide identity with a deeper shade of red being a greater level of similarity.

The accessory genes are more likely to be involved in the differences between species, given that they are not conserved throughout the genus or species, and can permit some insight into the adaptation of the species to their given environments. By using the COG annotations of the gene clusters, proportions of gene clusters belonging to different functional categories can be compared between the core and accessory genome (Figure 4.4). In the core genome, well-conserved functional categories such as translation and protein metabolism make up a greater proportion of the core genes that in the accessory genome. In contrast, protein categories such as defence mechanisms, cell wall biogenesis, and inorganic ion transport and metabolism make up an out-sized proportion of the accessory genome. The last of which is of particular interest given the known ability of the *Manganitrophus* genus to oxidize manganese as the sole source of energy. Interestingly, the manganese oxidases responsible for this process are found in both the core and accessory genome. Manganese oxidase genes, as annotated by KEGG, are found in both the stable and unstable core gene cluster bins and the CM_GAC bin, which are the genes found primarily in only the novel assembled genomes. Additionally, c-type cytochrome genes are found in the core genome as well, and were noted as being a key component in the oxidation of manganese [45].

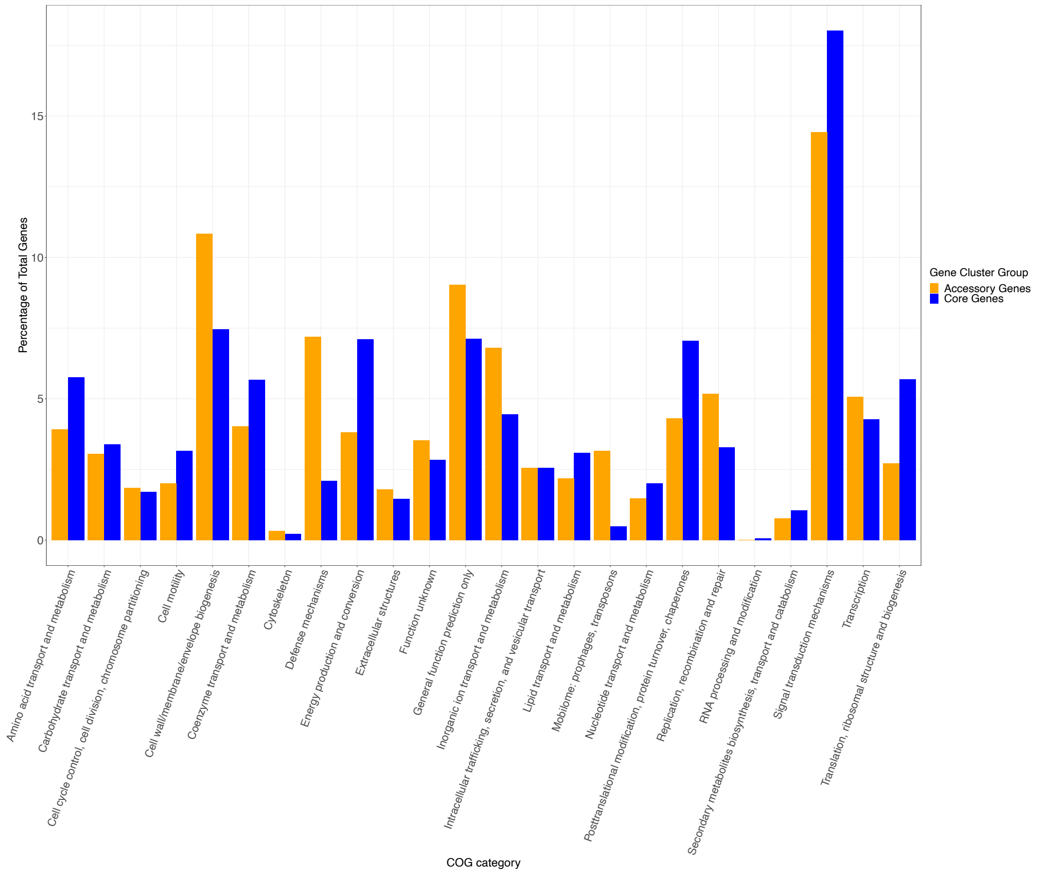


Figure 4.4: Proportions of core and accessory gene clusters found in the full genomes annotated as each COG functional group. Percentage calculated as the number of gene clusters annotated as COG functional group in the core or accessory genome relative to the total number of gene clusters in either the core or accessory genome.

4.3.5 Establishing the pangenome of the Manganitrophus plasmid

Though the plasmid was included in the pangenome displayed in Figure 4.3, it is important to separately analyze an independently replicating genetic element that in theory could be exchanged with other bacteria. Like the full pangenome, there are a number of gene clusters that are conserved within the genus (Figure 4.5). Also in a similar manner to the full pangenome, there are a number of genes that appear to be exclusive, or enriched, in the novel genomes assembled from the GAC community. In contrast to the full pangenome, cell wall biogenesis is enriched in the gene clusters of the core genome (Figure 4.6). Also, comparatively more represented in the core genome are gene clusters with functions of protein turnover and replication. In the accessory genome is an enrichment of genes related to the mobilome, signal transduction, and motility, which is suggestive that there may be some differences between the plasmids in terms of conjugative machinery because orthologues for these gene categories are not necessarily found in each plasmid sequence. In the full pangenome inorganic ion transport was shown to be more represented in the accessory than the core. However, this difference does not appear to be due to the differences in ion transport carried as cargo on the conjugative megaplasmid, which commonly harbour genes related to heavy metal transport [10]. Indeed, there are UniRef90 alignments on these plasmids for transporters and binding proteins of metals such as cadmium, nickel, mercury, and copper. Copper binding proteins and transporters are found among all of the species, but cadmium transporters are only annotated on the plasmids of the GAC strains.

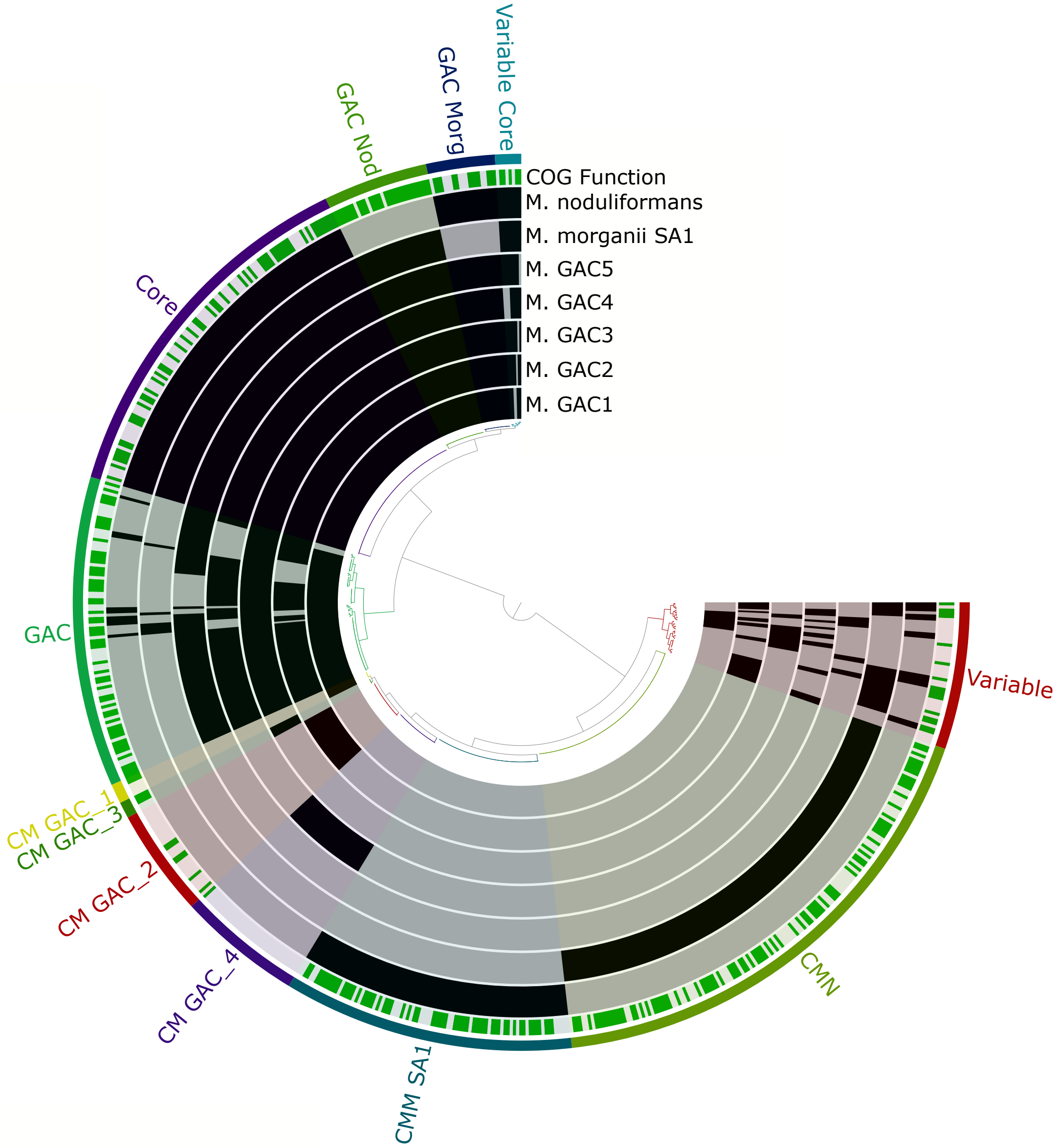


Figure 4.5: Pangenomic visualization of the plasmid content of the genus Manganitrophus. Black shading of circular track represents the presence of a gene cluster in a given strain. Bins of gene clusters on outer ring were determined manually to denote the pattern of presence in the groups of strains or strains.

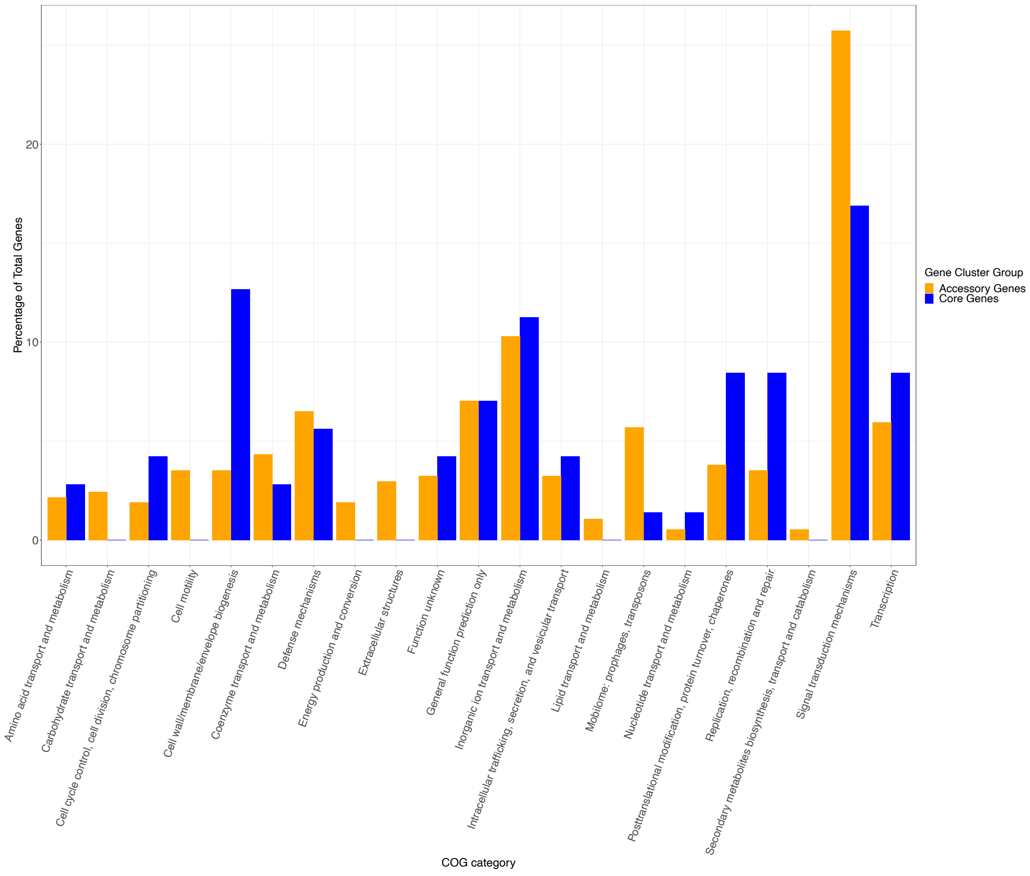


Figure 4.6: Percentage of core and accessory gene clusters found in the plasmids annotated as each COG functional group. Percentage calculated as the number of gene clusters annotated as COG functional group in the core or accessory genome relative to the total number of gene clusters in either the core or accessory genome.

4.3.6 Gene clusters characteristic to novel bacteria growing on GAC filters

Though there are 682 gene clusters in the 'CM_GAC' gene cluster bin, that does not mean that there are 682 unique genes. For instance, manganese oxidase genes are found in the 'CM_GAC' gene cluster bin but also the 'Stable Core' and 'Unstable Core' gene cluster bins. There is a high degree of overlap between the annotations for the gene clusters that appear divergent. This was addressed by searching for unique KEGG annotations of the gene clusters 'CM_GAC', 'CM_GAC2', 'CM_GAC_1', 'CM_GAC_3', 'CM_GAC_4', 'CM_GAC_5' (Table 4.3). Many of the genes are unique to these gene cluster bins, which suggests there is a difference with how these novel bacteria exchange DNA and other substrates with other bacteria in their environment due to the presence of unique conjugation genes. Also unique to the novel bacteria are a number of genes related to metal resistance. Zn2+/Cd2+-exporting ATPase, nickel/cobalt transporter (NicO) family protein, and copper resistance protein C are among the KEGG annotations unique to the GAC bacteria. These genes, among others, suggest that these bacteria are adapted to the harsh environment in the refinery effluent that is enriched with heavy metals and toxic compounds.

Table 4.3: Unique COG annotations for each *Manganitrophus* GAC strain sorted into their respective COG categories.

COG Category	GAC1	GAC2	GAC3	GAC4	GAC5
Signal transduction mechanisms	20	13	31	24	24
Inorganic ion transport and metabolism	8	5	6	7	4
Intracellular trafficking, secretion, and vesicular transport	9	7	7	8	8
Extracellular structures	11	8	8	9	9
Nucleotide transport and metabolism	5	8	5	5	6
Transcription	5	4	6	5	5
General function prediction only	4	2	4	3	4
Cell wall/membrane/envelope biogenesis	9	6	6	7	6
Coenzyme transport and metabolism	2	2	2	2	2
Posttranslational modification, protein turnover, chaperones	6	4	3	0	2
Replication, recombination and repair	4	1	4	7	2
Function unknown	8	9	6	8	6
Mobilome: prophages, transposons	3	4	3	1	1
Cell motility	2	1	1	1	1
Carbohydrate transport and metabolism	3	4	2	3	2
Amino acid transport and metabolism	5	5	3	4	3
Defense mechanisms	2	2	1	2	1
Secondary metabolites biosynthesis, transport and catabolism	1	1	1	1	1
Translation, ribosomal structure and biogenesis	1	1	2	1	0
Lipid transport and metabolism	1	1	0	1	0
Energy production and conversion	0	0	1	0	1

In addition, the annotations unique to *M. noduliformans* and *M. morganii* were explored. Of note, there are separate type IV secretion system protein and plasmid separation proteins, which coincides with the notion that the original *M. noduliformans* genome has an additional conjugative plasmid in comparison to the other seven genomes. In Table 4.1, it is clear that *M. noduliformans* is much larger than the other non-redundant genomes that have been surveyed. Additionally, there are CRISPR systems unique to these genomes, which are represented in the defense mechanism enrichment in the accessory genome in Figure 4.4. CRISPR arrays are only predicted on two contigs belonging to *M. noduliformans* and may be unique to that species.

4.3.7 Hydrocarbon metabolism in the genus Manganitrophus

The ability of *Manganitrophus* species to degrade toxic compounds is of special interest because of the environment that these novel genomes are derived from. The wastewater that is filtered through the GAC systems are rich in naphthenic acids, asphaltenes, and other hydrocarbons, and it is presumed that the bacteria growing on these filters play a role in the removal of these compounds through degradation. Given the differences in environments between the novel species and *M. noduliformans* or *M. morganii*, which were recovered from tap water and sediments, one would expect differences in the capacity to metabolize hydrocarbons. However, there is a near complete overlap in the hydrocarbon metabolism genes predicted by CANT-HYD and AromaDeg (Table 4.4). Within the genus *Manganitrophus*, there appears to be a capacity to metabolize phthalate, gentisate, salicylate, monocyclic aromatic hydrocarbons, and biphenyls (as predicted through the AromaDeg database). Predicted by CANT-HYD, are genes responsible for the metabolism of benzene, alkane, p-cymene, phenol, dibenzothiophene, ethylbenzene, naphthalene, monoaromatics, polyaromatics, propane, butane, and toluene throughout the genus. Metabolism of hydrocarbons as a source of energy is evidently a well-conserved function in this genus despite the disparate environments they were discovered in.

Table 4.4: Summary of hyrdocarbon metabolism annotation through CANT-HYD and AromaDeg.

	CANT-HYD annotations	AromaDeg annotations
M. noduliformans	16	12
M. morganii SA1	16	14
M. morganii SB1	15	13
M. GAC1	17	13
M. GAC2	17	14
M. GAC3	15	14
M. GAC4	17	14
M. GAC5	16	14

4.3.8 Evidence of intrasample heterogeneity in genomes

As evidenced by the incomplete and contaminated assemblies of strains GAC3 and GAC5 (Table 4.1), there is intrasample strain heterogeneity in the novel genomes. To some extent, the heterogeneity can be visualized through the assembly graphs of GAC3 and GAC5 (Figures 4.7 and A.2). There are multiple bubbles on the assembly graphs that show potential strains within the closed-loop assembly graphs. In the strain GAC3, the bubbles on the graph exist as a set of two paths with one path usually at about 20x coverage and the other path as 30x coverage. GAC5 has 3 apparent bubbles on the assembly graph where the two paths are at roughly 100x and 200x coverage compared to the mean coverage of roughly 300x, which is strong evidence of two strains within this sample (Figure 4.7). By separating the strains from sample GAC5, CheckM can be performed to assess the completion, contamination, and strain heterogeneity for each of the two proposed strains identified on the assembly graph. Strain 1, the red/black path on Figure 4.7, showed high completion and low contamination with no strain heterogeneity, whereas strain 2 (the blue/black path) showed lower completion, higher contamination, and strain heterogeneity (Table 4.5). The quality metrics show that strain 1 is a high-quality genome of the novel *Manganitrophus* species and that strain 2 still has too much heterogeneity to be considered a high-quality representation of the species. Performing a mapped assembly to each of the strains did not result in a higher quality assembly, likely due to the high similarity in sequences resulting in cross mapping and retention of the other strain in the filtered reads. Unfortunately due to the similar coverage of the two strains in the assembly graph of GAC3, the strains present could not be resolved in the same manner as GAC5.

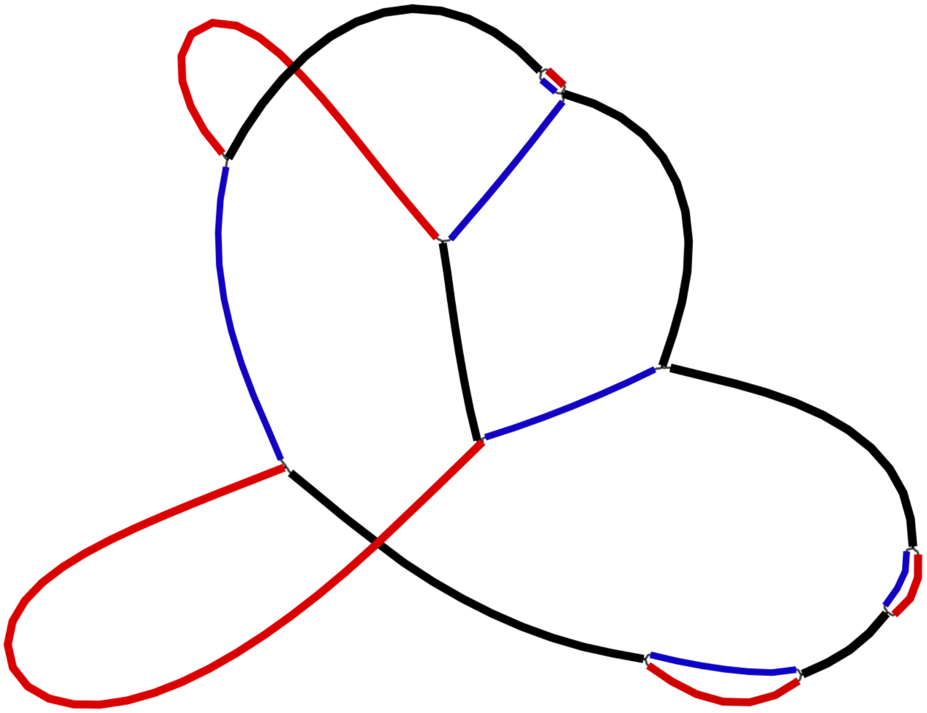


Figure 4.7: Assembly graph of sample GAC5 with strains highlighted. Edges of the graph in black are shared between the strains, edges in red are those belonging to strain 1, and edges in blue are those belonging to strain 2.

Table 4.5: Quality and summary metrics strains resolved from sample GAC5 as evaluated by CheckM.

	Completion	Contamination	Strain Heterogeneity	GC Content	Size (bp)	# Contigs
Strain 1	93.58	2.73	0	56.36	4168356	11
Strain 2	74.03	5.45	50.00	56.75	3089018	11

4.4 Discussion

In summary, five novel sequences recovered from metagenomic datasets were analyzed in the context of their genus. These sequences are of particular interest because they expand the range of non-core functions for the recently described genus *Manganitrophus*, whose biochemical pathway for manganese oxidation for energy has been thoroughly analyzed [45, 46]. Through an unrooted phylogenetic tree built on alignments of the single-copy core genes, the novel sequences were placed onto the same branch of the tree as the three known genomes of the *Manganitrophus* genus (Figure 4.1). The closeness on the tree was suggestive of a high degree of similarity, which is confirmed by ANI calculations that places the published genomes between 91-95% identity to the novel sequences (Figure 4.2). Indeed, many genes are highly conserved within the genus included genes for manganese oxidase, which is found in the core, variable core, and GAC strain-specific bins of gene clusters in the pangenome (Figure 4.3). Based on the conservation of manganese oxidase and the high degree of sequence identity, it is likely that these novel strains are able to utilize the oxidation of manganese for energy in the same manner that *Candidatus Manganitrophus noduliformans* was shown to be capable of in culture [45]. Another metabolic pathway that is conserved among all members of the *Manganitrophus* genus, which is of interest for bioremediation, are hydrocarbon degradation pathways. Many genes are predicted through alignments to the CANT-HYD and AromaDeg databases [67, 70], which suggest the capacity for these species to degrade aromatic and non-aromatic hydrocarbons, which are byproducts of oil extraction and processing. Given that the novel strains presented were found in a consortium that grows on the carbon filters at a wastewater treatment facility for an oil refinery, it is possible that the metabolism of hydrocarbons is a major source of energy for these bacteria. One of the metabolic pathways that the novel strains diverge greatly from the other *Manganitrophus* species is the ability to transport heavy metal ions. Found exclusively in the GAC strains are genes for zinc, cobalt, cadmium, nickel, and copper transport, which has implications for bioremediation as many of these metals are common environmental contaminants in the refinery effluent [4]. Additionally, heavy metal contamination is known to inhibit biochemical processes such as ammonia oxidation [27]. The ability of this species to transport and sequester heavy metals from a heavily contaminated environment may work synergistically with the other members of the biofilm to metabolize the wide array of hydrocarbons uninhibited. Furthermore, the refinery effluent is anoxic, so these bacteria need an energy source not tied to oxygen, which the anaerobic metabolism of hydrocarbons can provide [74].

Culturable bacteria make up the bulk of the databases due to their ease of study [75], which makes the use of metagenomic analyses to study the more-abundant, uncultured bacteria (as seen in human gut microbiome studies) greatly important [76]. To obtain circularized, or highly contiguous, genomes from the five metagenomic samples of bacteria growing on the GAC filters, methods beyond standard metagenomic assembly had to be developed and employed. From the standard long-read metagenomic assembly of the samples, one genome could be successfully assembled into a circular contig. However, for the other samples, metagenomic assembly resulted in a heavily fragmented *Manganitrophus* genome. The complexity of the community growing on the GAC filters and the apparent strain heterogeneity of the novel species necessitated the use of methods that can overcome these obstacles. Drawing inspiration from reference-guided assemblies and Jorg [77], which attempts to assemble circularize sequences using short reads that map to bins, long reads were mapped to the successfully circularized GAC1 strain and filtered for reads with full-length alignments using GERENUQ [50], which will discard reads that are poor alignments to the genome due to cross mapping from other species or chimeric reads that were erroneously sequenced. Using this filtered set of reads, the two additional circularized sequences and two highly contiguous sequences could be assembled. The chromosomal sequences in samples GAC3 and GAC5 could not be completed due to the high level of strain heterogeneity in the samples. In sample GAC5, the two competing strains of the *Manganitrophus* genome can be visualized on the assembly graph (Figure 4.7). By selecting for the contigs with a consistent coverage pattern, the strains of GAC5 can be resolved, in part. Strain 1 of GAC5 has high completion and low redundancy and mirrors the three circularized strains in size. Knowing the circularized sequences of the bacterial genomes permitted pangenomic comparisons between strains of the novel strains and to the published *Candidatus Manganitrophus noduliformans* and *Candidatus Manganitrophus morganii* genomes.

In addition to the chromosomal sequences of the novel *Manganitrophus* genomes, the sequences of a large conjugative plasmid belonging to the strain was recovered from each sample. Similar to the chromosomal sequences, the circularized sequence of the GAC1 plasmid was assembled and identified from the metagenomic assembly of the sample. Also similarly, the subsequent mapped assembly of the plasmid sequence with the four other samples yielded two circularized and two highly-contiguous plasmids. Furthermore, plasmid sequences from *Candidatus Manganitrophus noduliformans* and *Candidatus Manganitrophus morganii* SB1 were identified through alignment to the GAC1 plasmid. Given that plasmids commonly harbour genes related to hydrocarbon degradation and heavy metal regulation [23, 30, 10], it is essential that plasmids can be identified and separated from their cognate chromosome. In addition to the conserved plasmid identified in seven of the eight genomes, the analysis suggests that the published *Candidatus Manganitrophus noduliformans* has an additional conjugative plasmid that harbours a CRISPR system. Plasmids are known to regularly harbour these gene-editing systems [78] and this sequence is likely responsible for the enrichment in defense mechanisms in the accessory genome seen in Figure 4.4. Plasmids represent a potentially transferable genetic element, so knowing their sequences can give insight into the potential metabolic pathways that can be exchanged by these species within a biofilm.

4.5 Conclusion

An ever-increasing number of unculturable bacteria are being sequenced whose metabolism pushes the limit on what was thought to be possible. Five additional genomes have been added to the genus that is able to oxidize manganese for energy and by doing so has increased the resolution of the genomes within the genus that came before. With these additional genomes, it is now known that there is a large megaplasmid of over 300kb in length that is well-conserved among most, if not all members of the genus. Additionally, it is likely that these bacteria utilize hyrdocarbons as another energy source given the high number of predicted genes involved in the process and the environment from which the novel strains were sequenced within. Isolation of the novel species and testing of the metabolic potential for heavy metals and hydrocarbons could shed further light on the role of this unique species growing on the biofilms that form on granular activated charcoal filters in an oil refinery.

4.6 References

Bibliography

[1] Raymond A. Wuana and Felix E. Okieimen. Heavy Metals in Contaminated Soils: A Review of Sources, Chemistry, Risks and Best Available Strategies for Remediation. *ISRN Ecology*, 2011:1–20, October 2011.

[2] I Pulford. Phytoremediation of heavy metal-contaminated land by trees—a review. *Environment International*, 29(4):529–540, July 2003.

[3] Sonia M. Tiquia-Arashiro. Lead absorption mechanisms in bacteria as strategies for lead bioremediation. *Applied Microbiology and Biotechnology*, 102(13):5437–5444, July 2018.

[4] Shahid Sher and Abdul Rehman. Use of heavy metals resistant bacteria—a strategy for arsenic bioremediation. *Applied Microbiology and Biotechnology*, 103(15):6007–6021, August 2019.

[5] Bhupendra Pushkar, Pooja Sevak, Sejal Parab, and Nikita Nilkanth. Chromium pollution and its bioremediation mechanisms in bacteria: A review. *Journal of Environmental Management*, 287:112279, June 2021.

[6] Padma Seragadam, Abhilasha Rai, Kartik Chandra Ghanta, Badri Srinivas, Sandip Kumar Lahiri, and Susmita Dutta. Bioremediation of hexavalent chromium from wastewater using bacteria-a green technology. *Biodegradation*, 32(4):449–466, August 2021.

[7] Abd Elnaby Hanan, M Abou Elela Gehan, and A El Sersy Nermeen. Cadmium resisting bacteria in Alexandria Eastern Harbor (Egypt) and optimization of cadmium bioaccumulation by Vibrio harveyi. *African Journal of Biotechnology*, 10(17):3412–3423, April 2011.

[8] F. Pagnanelli, M. Petrangeli Papini, L.Toro,, M. Trifoni, and F. Vegliò. Biosorption of Metal Ions on *Arthrobacter sp.* : Biomass Characterization and Biosorption Modeling. *Environmental Science & Technology*, 34(13):2773–2778, July 2000.

[9] Lina Velásquez and Jenny Dussan. Biosorption and bioaccumulation of heavy metals on dead and living biomass of Bacillus sphaericus. *Journal of Hazardous Materials*, 167(1-3):713–

716, August 2009.

[10] Surajit Das, Hirak R. Dash, and Jaya Chakraborty. Genetic basis and importance of metal resistant genes in bacteria for bioremediation of contaminated environments with toxic metal pollutants. *Applied Microbiology and Biotechnology*, 100(7):2967–2984, April 2016.

[11] Varenyam Achal, Xiangliang Pan, Qinglong Fu, and Daoyong Zhang. Biomineralization based remediation of As(III) contaminated soil by Sporosarcina ginsengisoli. *Journal of Hazardous Materials*, 201-202:178–184, January 2012.

[12] Jinghong Zhang, Xu Zhang, Yongqing Ni, Xiaojuan Yang, and Hongyu Li. Bioleaching of arsenic from medicinal realgar by pure and mixed cultures. *Process Biochemistry*, 9(42):1265–1271, 2007.

[13] Pratima Gupta and Batul Diwan. Bacterial Exopolysaccharide mediated heavy metal removal: A Review on biosynthesis, mechanism and remediation strategies. *Biotechnology Reports*, 13:58–71, March 2017.

[14] W. C. Leung, M.-F. Wong, H. Chua, W. Lo, P. H. F. Yu, and C. K. Leung. Removal and recovery of heavy metals by bacteria isolated from activated sludge treating industrial effluents and municipal wastewater. *Water Science and Technology*, 41(12):233–240, June 2000.

[15] H. A. Elliott, M. R. Liberati, and C. P. Huang. Competitive Adsorption of Heavy Metals by Soils. *Journal of Environmental Quality*, 15(3):214–219, July 1986.

[16] Milind Mohan Naik and Santosh Kumar Dubey. Lead-enhanced siderophore production and alteration in cell morphology in a Pb-resistant Pseudomonas aeruginosa strain 4EA. *Current Microbiology*, 62(2):409–414, February 2011.

[17] Robert A. Kanaly and Shigeaki Harayama. Biodegradation of High-Molecular-Weight Polycyclic Aromatic Hydrocarbons by Bacteria. *Journal of Bacteriology*, 182(8):2059–2067, April 2000.

[18] Francesca Mapelli, Alberto Scoma, Grégoire Michoud, Federico Aulenta, Nico Boon, Sara Borin, Nicolas Kalogerakis, and Daniele Daffonchio. Biotechnologies for Marine Oil Spill Cleanup: Indissoluble Ties with Microorganisms. *Trends in Biotechnology*, 35(9):860–870,

September 2017.

[19] Sudip K. Samanta, Om V. Singh, and Rakesh K. Jain. Polycyclic aromatic hydrocarbons: environmental pollution and bioremediation. *Trends in Biotechnology*, 20(6):243–248, June 2002.

[20] Noriyuki Iwabuchi, Michio Sunairi, Makoto Urai, Chiaki Itoh, Hiroshi Anzai, Mutsuyasu Nakajima, and Shigeaki Harayama. Extracellular polysaccharides of Rhodococcus rhodochrous S-2 stimulate the degradation of aromatic components in crude oil by indigenous marine bacteria. *Applied and Environmental Microbiology*, 68(5):2337–2343, May 2002.

[21] E. Gontikaki, L. D. Potts, J. A. Anderson, and U. Witte. Hydrocarbon-degrading bacteria in deep-water subarctic sediments (Faroe-Shetland Channel). *Journal of Applied Microbiology*, 125(4):1040–1053, October 2018.

[22] K. Rahul, Ch. Sasikala, L. Tushar, R. Debadrita, and Ch. V.YR 2014 Ramana. Alcanivorax xenomutans sp. nov., a hydrocarbonoclastic bacterium isolated from a shrimp cultivation pond. *International Journal of Systematic and Evolutionary Microbiology*, 64(Pt_10):3553–3558. Publisher: Microbiology Society,.

[23] Andreas Stolz. Degradative plasmids from sphingomonads. *FEMS Microbiology Letters*, 350(1):9–19, January 2014.

[24] Somnath Mallick, Joydeep Chakraborty, and Tapan K. Dutta. Role of oxygenases in guiding diverse metabolic pathways in the bacterial degradation of low-molecular-weight polycyclic aromatic hydrocarbons: a review. *Critical Reviews in Microbiology*, 37(1):64–90, February 2011.

[25] S. A. Selifonov, M. Grifoll, R. W. Eaton, and P. J. Chapman. Oxidation of naphthenoaromatic and methyl-substituted aromatic compounds by naphthalene 1,2-dioxygenase. *Applied and Environmental Microbiology*, 62(2):507–514, February 1996.

[26] D. T. Gibson, S. M. Resnick, K. Lee, J. M. Brand, D. S. Torok, L. P. Wackett, M. J. Schocken, and B. E. Haigler. Desaturation, dioxygenation, and monooxygenation reactions catalyzed by naphthalene dioxygenase from Pseudomonas sp. strain 9816-4. *Journal of Bacteriology*,

177(10):2615–2621, May 1995.

[27] Cheng Yu, Xi Tang, Lu-Shan Li, Xi-Lin Chai, Ruiyang Xiao, Di Wu, Chong-Jian Tang, and Li-Yuan Chai. The long-term effects of hexavalent chromium on anaerobic ammonium oxidation process: Performance inhibition, hexavalent chromium reduction and unexpected nitrite oxidation. *Bioresource Technology*, 283:138–147, July 2019.

[28] Qingguo Chen, Jingjing Li, Mei Liu, Huiling Sun, and Mutai Bao. Study on the biodegradation of crude oil by free and immobilized bacterial consortium in marine environment. *PLOS ONE*, 12(3):e0174445, March 2017.

[29] Hassan Brim, Sara C. McFarlan, James K. Fredrickson, Kenneth W. Minton, Min Zhai, Lawrence P. Wackett, and Michael J. Daly. Engineering Deinococcus radiodurans for metal remediation in radioactive mixed waste environments. *Nature Biotechnology*, 18(1):85–90, January 2000.

[30] Hiroshi Habe and Toshio Omori. Genetics of polycyclic aromatic hydrocarbon metabolism in diverse aerobic bacteria. *Bioscience, Biotechnology, and Biochemistry*, 67(2):225–243, February 2003.

[31] Susan Jane Fowler, Alejandro Palomo, Arnaud Dechesne, Paul D. Mines, and Barth F. Smets. Comammox Nitrospira are abundant ammonia oxidizers in diverse groundwater-fed rapid sand filter communities. *Environmental Microbiology*, 20(3):1002–1015, 2018. _eprint: https://sfamjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/1462-2920.14033.

[32] Alejandro Palomo, Anders G. Pedersen, S. Jane Fowler, Arnaud Dechesne, Thomas Sicheritz-Pontén, and Barth F. Smets. Comparative genomics sheds light on niche differentiation and the evolutionary history of comammox Nitrospira. *The ISME Journal*, 12(7):1779–1793, July 2018.

[33] Maartje A.H.J. van Kessel, Daan R. Speth, Mads Albertsen, Per H. Nielsen, Huub J.M. Op den Camp, Boran Kartal, Mike S.M. Jetten, and Sebastian Lücker. Complete nitrification by a single microorganism. *Nature*, 528(7583):555–559, December 2015.

[34] Lianna Poghosyan, Hanna Koch, Adi Lavy, Jeroen Frank, Maartje A.H.J. Kessel, Mike S.M.

Jetten, Jillian F. Banfield, and Sebastian Lücker. Metagenomic recovery of two distinct comammox *Nitrospira* from the terrestrial subsurface. *Environmental Microbiology*, 21(10):3627–3637, October 2019.

[35] Hanna Koch, Maartje A. H. J. van Kessel, and Sebastian Lücker. Complete nitrification: insights into the ecophysiology of comammox Nitrospira. *Applied Microbiology and Biotechnology*, 103(1):177–189, January 2019.

[36] Hirotsugu Fujitani, Kengo Momiuchi, Kento Ishii, Manami Nomachi, Shuta Kikuchi, Norisuke Ushiki, Yuji Sekiguchi, and Satoshi Tsuneda. Genomic and Physiological Characteristics of a Novel Nitrite-Oxidizing Nitrospira Strain Isolated From a Drinking Water Treatment Plant. *Frontiers in Microbiology*, 11:545190, September 2020.

[37] Yulia A. Frank, Vitaly V. Kadnikov, Anastasia P. Lukina, David Banks, Alexey V. Beletsky, Andrey V. Mardanov, Elena I. Sen'kina, Marat R. Avakyan, Olga V. Karnachuk, and Nikolai V. Ravin. Characterization and Genome Analysis of the First Facultatively Alkaliphilic Thermodesulfovibrio Isolated from the Deep Terrestrial Subsurface. *Frontiers in Microbiology*, 7, December 2016.

[38] E. A. Henry, R. Devereux, J. S. Maki, C. C. Gilmour, C. R. Woese, L. Mandelco, R. Schauder, C. C. Remsen, and R. Mitchell. Characterization of a new thermophilic sulfate-reducing bacterium Thermodesulfovibrio yellowstonii, gen. nov. and sp. nov.: its phylogenetic relationship to Thermodesulfobacterium commune and their origins deep within the bacterial domain. *Archives of Microbiology*, 161(1):62–69, January 1994.

[39] Olfa Haouari, Marie-Laure Fardeau, Jean-Luc Cayol, Guy Fauque, Corinne Casiot, Françoise Elbaz-Poulichet, Moktar Hamdi, and Bernard Ollivier. Thermodesulfovibrio hydrogeniphilus sp. nov., a new thermophilic sulphate-reducing bacterium isolated from a Tunisian hot spring. *Systematic and Applied Microbiology*, 31(1):38–42, March 2008.

[40] Y. Sekiguchi, M. Muramatsu, H. Imachi, T. Narihiro, A. Ohashi, H. Harada, S. Hanada, and Y. Kamagata. Thermodesulfovibrio aggregans sp. nov. and Thermodesulfovibrio thiophilus sp. nov., anaerobic, thermophilic, sulfate-reducing bacteria isolated from thermophilic

methanogenic sludge, and emended description of the genus Thermodesulfovibrio. *INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY*, 58(11):2541–2548, November 2008.

[41] Jacob Sonne-Hansen and Birgitte K. Ahring. Thermodesulfobacterium hveragerdense sp.nov., and Thermodesulfovibrio islandicus sp.nov., Two Thermophilic Sulfate Reducing Bacteria Isolated from a Icelandic Hot Spring. *Systematic and Applied Microbiology*, 22(4):559–564, December 1999.

[42] Kazuhiro Umezawa, Hisaya Kojima, Yukako Kato, and Manabu Fukui. Disproportionation of inorganic sulfur compounds by a novel autotrophic bacterium belonging to Nitrospirota. *Systematic and Applied Microbiology*, 43(5):126110, September 2020.

[43] H Hippe. Leptospirillum gen. nov. (ex Markosyan 1972), nom. rev., including Leptospirillum ferrooxidans sp. nov. (ex Markosyan 1972), nom. rev. and Leptospirillum thermoferrooxidans sp. nov. (Golovacheva et al. 1992). *International Journal of Systematic and Evolutionary Microbiology*, 50(2):501–503, March 2000.

[44] S. Spring, R. Amann, W. Ludwig, K. H. Schleifer, H. van Gemerden, and N. Petersen. Dominating role of an unusual magnetotactic bacterium in the microaerobic zone of a freshwater sediment. *Applied and Environmental Microbiology*, 59(8):2397–2403, August 1993.

[45] Hang Yu and Jared R. Leadbetter. Bacterial chemolithoautotrophy via manganese oxidation. *Nature*, 583(7816):453–458, July 2020.

[46] Hang Yu, Grayson L. Chadwick, Usha F. Lingappa, and Jared R. Leadbetter. Comparative genomics on cultivated and uncultivated, freshwater and marine *Candidatus* Manganitrophaceae species implies their worldwide reach in manganese chemolithoautotrophy. preprint, Microbiology, November 2021.

[47] Wouter De Coster, Svenn D'Hert, Darrin T Schultz, Marc Cruts, and Christine Van Broeckhoven. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*, 34(15):2666–2669, August 2018.

[48] Mikhail Kolmogorov, Derek M. Bickhart, Bahar Behsaz, Alexey Gurevich, Mikhail Rayko,

Sung Bong Shin, Kristen Kuhn, Jeffrey Yuan, Evgeny Polevikov, Timothy P. L. Smith, and Pavel A. Pevzner. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods*, 17(11):1103–1110, November 2020.

[49] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, September 2018.

[50] abahcheli. GERENUQ, October 2021. original-date: 2020-10-31T15:13:33Z.

[51] Robert Vaser, Ivan Sović, Niranjan Nagarajan, and Mile Šikić. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, 27(5):737–746, May 2017.

[52] Medaka, February 2022. original-date: 2017-06-07T14:01:06Z.

[53] F. A. Bastiaan von Meijenfeldt, Ksenia Arkhipova, Diego D. Cambuy, Felipe H. Coutinho, and Bas E. Dutilh. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biology*, 20(1):217, December 2019.

[54] Benjamin R Joris, Tyler S Browne, Thomas A Hamilton, David R Edgell, and Gregory B Gloor. Separation of Cohorts on the Basis of Bacterial Type IV Conjugation Systems Identified From Metagenomic Assemblies. preprint, In Review, June 2021.

[55] A. Murat Eren, Özcan C. Esen, Christopher Quince, Joseph H. Vineis, Hilary G. Morrison, Mitchell L. Sogin, and Tom O. Delmont. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*, 3:e1319, October 2015.

[56] Mikhail Kolmogorov, Jeffrey Yuan, Yu Lin, and Pavel A. Pevzner. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, 37(5):540–546, May 2019.

[57] Ryan R. Wick, Mark B. Schultz, Justin Zobel, and Kathryn E. Holt. Bandage: interactive visualization of *de novo* genome assemblies: Fig. 1. *Bioinformatics*, 31(20):3350–3352, October 2015.

[58] Paul A. Kitts, Deanna M. Church, Françoise Thibaud-Nissen, Jinna Choi, Vichet Hem, Victor Sapojnikov, Robert G. Smith, Tatiana Tatusova, Charlie Xiang, Andrey Zherikov, Michael DiCuccio, Terence D. Murphy, Kim D. Pruitt, and Avi Kimchi. Assembly: a resource for

assembled genomes at NCBI. *Nucleic Acids Research*, 44(D1):D73–80, January 2016.

[59] Donovan H. Parks, Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, and Gene W. Tyson. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 25(7):1043–1055, July 2015.

[60] Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, 5(3):e9490, March 2010.

[61] Ivica Letunic and Peer Bork. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, 49(W1):W293–W296, July 2021.

[62] Leighton Pritchard, Rachel H. Glover, Sonia Humphris, John G. Elphinstone, and Ian K. Toth. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Analytical Methods*, 8(1):12–24, 2016.

[63] Chirag Jain, Luis M. Rodriguez-R, Adam M. Phillippy, Konstantinos T. Konstantinidis, and Srinivas Aluru. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications*, 9(1):5114, December 2018.

[64] R. L. Tatusov, M. Y. Galperin, D. A. Natale, and E. V. Koonin. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, 28(1):33–36, January 2000.

[65] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361, January 2017.

[66] B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu, and the UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, March 2015.

[67] Márcia Duarte, Ruy Jauregui, Ramiro Vilchez-Vargas, Howard Junca, and Dietmar H. Pieper. AromaDeg, a novel database for phylogenomics of aerobic bacterial degradation of aromatics. *Database*, 2014, January 2014.

[68] Benjamin Buchfink, Klaus Reuter, and Hajk-Georg Drost. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, 18(4):366–368, April 2021.

[69] Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1):59–60, January 2015.

[70] Varada Khot, Jackie Zorz, Daniel A. Gittins, Anirban Chakraborty, Emma Bell, María A. Bautista, Alexandre J. Paquette, Alyse K. Hawley, Breda Novotnik, Casey R. J. Hubert, Marc Strous, and Srijak Bhatnagar. CANT-HYD: A Curated Database of Phylogeny-Derived Hidden Markov Models for Annotation of Marker Genes Involved in Hydrocarbon Degradation. *Frontiers in Microbiology*, 12:764058, January 2022.

[71] Sean R. Eddy. Accelerated Profile HMM Searches. *PLoS computational biology*, 7(10):e1002195, October 2011.

[72] Robert C Edgar. PILER-CR: Fast and accurate identification of CRISPR repeats. *BMC Bioinformatics*, 8(1):18, December 2007.

[73] Hao Luo, Chun-Lan Quan, Chong Peng, and Feng Gao. Recent development of Ori-Finder system and DoriC database for microbial replication origins. *Briefings in Bioinformatics*, 20(4):1114–1124, July 2019.

[74] Kartik Dhar, Suresh R. Subashchandrabose, Kadiyala Venkateswarlu, Kannan Krishnan, and Mallavarapu Megharaj. Anaerobic Microbial Degradation of Polycyclic Aromatic Hydrocarbons: A Comprehensive Review. *Reviews of Environmental Contamination and Toxicology*, 251:25–108, 2020.

[75] Jörg Overmann, Birte Abt, and Johannes Sikorski. Present and Future of Culturing Bacteria. *Annual Review of Microbiology*, 71:711–730, September 2017.

[76] Jean-Christophe Lagier, Grégory Dubourg, Matthieu Million, Frédéric Cadoret, Melhem Bilen, Florence Fenollar, Anthony Levasseur, Jean-Marc Rolain, Pierre-Edouard Fournier, and Didier Raoult. Culturing the human microbiota and culturomics. *Nature Reviews. Microbiology*, 16:540–550, May 2018.

[77] Lauren M. Lui, Torben N. Nielsen, and Adam P. Arkin. A method for achieving complete

microbial genomes and improving bins from metagenomics data. *PLOS Computational Biology*, 17(5):e1008972, May 2021.

[78] Rafael Pinilla-Redondo, Jakob Russel, David Mayo-Muñoz, Shiraz A. Shah, Roger A. Garrett, Joseph Nesme, Jonas S. Madsen, Peter C. Fineran, and Søren J. Sørensen. CRISPR-Cas systems are widespread accessory elements across bacterial and archaeal plasmids. *Nucleic Acids Research*, page gkab859, October 2021.

Chapter 5

Secondary assembly of potential hydrocarbon-degrading bacterial communities yields additional complete genomes

5.1 Background

Long-read sequencing with Oxford Nanopore Technologies platforms has enabled the sequencing of reads with a nearly unbound upper limit for length. The read length advantage that third generation sequencing technologies have introduced has provided benefits to many genomics fields, including metagenomics[1]. Assembly algorithms such as metaFlye [2] were developed to specifically take advantage of the longer, yet error-prone, reads. As a result of the long reads being capable of spanning repetitive regions that would previously confound short read metagenomic assemblies, long-read metagenomic assemblies could generate far more contiguous assemblies. In fact, in recent years as improved methodology have increased the integrity of the extracted DNA and the length of the sequenced DNA, a number of studies have been published that have successfully circularized bacterial genomes from complex environments, such as the human gut and activated sludge [3, 4, 5, 6]. In perhaps the most extreme example of the potential of long reads, a total of 44 circuluar metagenome-assembled genomes (MAGs) were generated from a single sample [7]. While it is unreasonable to sequence 255 Gbp per sample due to the expense in a typical metagenomic study, this study highlights the potential of long-read sequencing to characterize complex communities down to the strain level.

Despite the promise of long-read sequencing, much of the development of metagenomic assembly protocols remains focused on short-read sequencing and there is much room for growth for tools and frameworks for utilizing long reads to their maximum potential. For short read assembly data, there are tools such as Recycler [8] and Jorg [9] that look to maximize the potential of the data by increasing the assembly quality beyond the primary assembly by inferring circular sequences from the assembly graphs or by conducting subsequent assemblies on binned contigs. These tools recognize the benefit of identifying and assembling circular contigs from bacterial metagenomic assemblies. Complete chromosomal sequences can allow for precise identification of a genome's taxa without concerns about contamination and chimerism that can occur with metagenomic binning [10], full knowledge of the synteny of genes, and quality scaffolds to examine genetic variation. In addition, the circularization of extra-chromosomal elements with these approaches helps alleviate the concerns with the exclusion of plasmids and other mobile elements from metagenomic bins [11, 12]. While long reads are able to circularize sequences from a metagenomic assembly, long reads still fail to completely assemble all sequences present in a metagenome. As such, similar approaches to those used to maximize short-read assembly data could yield better results than a simple primary metagenomic assembly.

Here we present a framework for performing a secondary assembly on binned long-read assemblies. The contigs from the primary assembly are partitioned into uncircularized and circularized contigs. Uncircularized contigs from the primary metagenomic assembly are binned with MetaBAT2 and reads are aligned to bins and filtered for long-reads that align end-to-end. This subset of reads is then assembled as a pseudo-isolate in an attempt to generate circularized or higher quality assemblies as a result. The framework was applied to nine samples from an oil refinery environment: eight samples growing on granular-activated charcoal filters and one from wastewater flocculent. A total of 48 contigs over 1Mb were assembled in the primary metagenomic assemblies and the secondary assemblies yielded another 66 contigs over this size threshold. For smaller contigs the primary assembly generated 3522 small, circular contigs, whereas the secondary assembly only yielded 536 small, circular contigs. With these findings, we demonstrate the general utility of performing secondary assemblies with long read data to obtain additional circular contigs.

5.2 Methods

5.2.1 Phenol-Chloroform DNA extraction

Eight samples of granular activated carbon (GAC) and a single sample of flocculent were collected over a period of two years from the SunCor Energy Sarnia Canada facility. High molecular weight DNA was extracted from approximately 10 g of GAC sample or 50 mL flocculent using the same method. Prior to extraction, however, the flocculent was spun down to pellet at 4000 g for 10 minutes at RT and the supernatant was discarded.

10 mL of lysis buffer (10 mM Tris-HCl, 100 mM NaCl, 25 mM EDTA, 0.5% (w/v) SDS) and 2.5 mg of lysozyme was added to a 50 mL falcon tube containing the sample (either GAC or flocculent), and was mixed by slowly rotating the tube horizontally around its vertical axis to minimize granule movement. The mixture was incubated for 1 hour at 37 °C. 5 µL of RNAse (20 mg/mL) was then added to the mix and incubated at 57 °C for 30 minutes. Finally, 100 µL of Proteinase K (800 units/mL) was added and the mixture was incubated at 57 °C for 1 hour and 30 minutes. During each incubation step, the lysate was mixed using the aforementioned mixing technique every 15 minutes.

Following incubation, the sample was spun at 3000 g for 5 minutes to spin down small GAC particles that are suspended in the lysate before being decanted into a new 50 mL falcon tube. 1

volume of 25:24:1 phenol:chloroform:isomayl alcohol was added to the lysate, mixed by rocking on a destainer for 8 minutes, then spun down at 3000 g for 3 minutes. The aqueous phase was transferred to a new 50 mL falcon tube using wide bore pipette tips, then 1 volume of chloroform was added and the mixture was rocked on a destainer for 8 minutes, before being spun down at 3000 g for 3 minutes. The aqueous phase was transferred to a new tube and the chloroform wash step was repeated. Next, 1/10 volume of NaAC (3 M, pH 5.2) was added to the transferred aqueous phase and mixed by inversion. 2 volumes of ice cold 100% ethanol was then added and mixed by inversion. High molecular weight DNA that precipitated was spooled out into a 1.5 mL Eppendorf tube containing 200 µL of 75% ethanol using a Pasteur pipette that was melted into hook. The DNA was spun down at 10000 g for 3 minutes. The ethanol was removed without disturbing the pellet and another 200 µL of 75% ethanol was added and then spun at 10000 g for 3 minutes. Following centrifugation, the ethanol was removed and the pellet was left to air dry for 2 minutes. Depending on the size of the pellet, the DNA was resuspended with 200 µL to 1 mL of Tris-HCl (10 mM, pH 8) at 4 °C overnight.

The recovered DNA was size selected to retain fragments greater than 40kb using the Circulomics Short Read Eliminator XL kit according to the manufacturer's protocol. The sequencing library was prepared from the size selected DNA using Nanopore's ligation sequencing kit (SQK-LSK110) and its associated protocol, with a few changes. In the DNA repair and end prep steps, the DNA was incubated in the thermal cycler for 20 minutes at 20 °C and 20 minutes at 65 °C. Instead of AMPure XP beads, Omega Bio-Tek Mag-Bind beads were used in bead clean up steps. The library was sequenced on Oxford Nanopore's MinION platform, using a FLO-MIN106 R9.4.1 flow cell. Basecalling was performed with Guppy 5.0.16 in super accuracy mode (dna_r9.4.1_450bps_sup).

5.2.2 Read quality control and primary assembly

Prior to assembly the reads were filtered for read length and quality with NanoFilt [13] with settings '-q 10 -l 500' to retain reads greater than 500 base pairs in length and a mean read quality score of

10 or greater. The filtered reads were assembled metagenomically using metaFlye [2] with settings '–meta -g 5mb'. Following the assembly contigs were partitioned for whether metaFlye tagged the contig as being circular and not repetitive. Non-circular contigs were passed to the binning and secondary assembly pipeline.

5.2.3 Binning and secondary assembly

Reads from all samples were aligned to the set of non-circular contigs from each sample using minimap2 [14]. Using the alignment records, the contigs were grouped into bins using MetaBAT2 [15]. To each of the bins, the reads from the corresponding sample were aligned to the bin and filtered for strong, end-to-end alignments using GERENUQ [16]. GERENUQ filters alignments to those over 1000 base pairs in length, with an overall alignment score of 1, and with at least half of the base pairs matching. Using the filtered alignments to subset the reads, a secondary assembly of the bin was performed using Flye with the setting of genome size set to the total size of the bin.

5.2.4 Assembly quality assessment and validation

The quality of the bins were assessed before and after the secondary assembly using CheckM [17]. In addition, the contigs circularized from the primary assembly and the contigs that could not be binned by MetaBAT2 were also assessed for completion and contamination. To validate the annotation of contigs as being circularized by Flye, reads were first aligned to all contigs using minimap2 and output into the PAF format. A complete tiling path along the contig using reads greater than 5000bp in length that overlap by 500bp, with a read that aligns to the start and end of the contig, was used to confirm the circularization of the sequence. Assessment of the shared contigs and bins between samples was determined by dRep [18]. Contigs and bins were clustered at a level of 95% sequence identity, which is roughly at the boundary for species-level for bacteria [19], and the number of species per cluster was evaluated.

5.2.5 Taxonomic assignment and annotation of bins and contigs

Contigs initially reported by Flye as circularized, unbinned contigs, and secondarily assembled bins were assigned taxonomy using the Contig Annotation Tool (CAT) [20] version 5.2.2 with the database version '2021-01-07_taxonomy'. Circularized contigs 1Mb or greater in size and bins that exceeded a completion of 80% and were below a contamination level of 10% were additionally taxonomically assigned using GTDB-Tk [21] version 2.1.0 with database version 'R207_v2'. Contigs and bins were imported into Anvi'o [22] version 7.0 to annotate with NCBI COGs [23]. Plasmids were identified using PlasFlow version 1.1 [24]. Conjugative systems were identified with HMMER [25] by using a curated set of profile hidden Markov models (pHMMs) to search for contigs containing a relaxase, a type IV coupling protein, and a type IV secretion system protein [12]. Hydrocarbon and aromatic metabolic genes were specifically annotated for by aligning open reading frames predicted by Prodigal version 2.6.3 [26] to two databases: AromaDeg [27] with the BLASTP module of Diamond version 2.0.4 [28] and to CANT-HYD [29] using HMMER version 3.3.2 [25].

5.3 Results

5.3.1 Assembly of oil refinery metagenomes

To reconstruct the genomes of the bacteria that colonize the granular-activated charcoal filters and flocculent of Suncor Energy's Sarnia Canada oil refinery wastewater treatment facility, two rounds of assembly were conducted on the metagenomic reads sequenced on the Nanopore MinION platform (Figure 5.1). Uncircularized contigs from the primary metaFlye assembly were binned using MetaBAT2 and reads were aligned and filtered to these bins to created a pseudo-isolate set of reads that can be secondarily assembled using Flye. The three desired outcomes of the secondary assembly of the metagenomic bins were circularization of the chromosomal DNA, circularization of extra-chromosomal DNA, or a general increase in the overall assembly quality of the bin (as measured by completion, contamination, and contiguity of the assembly).

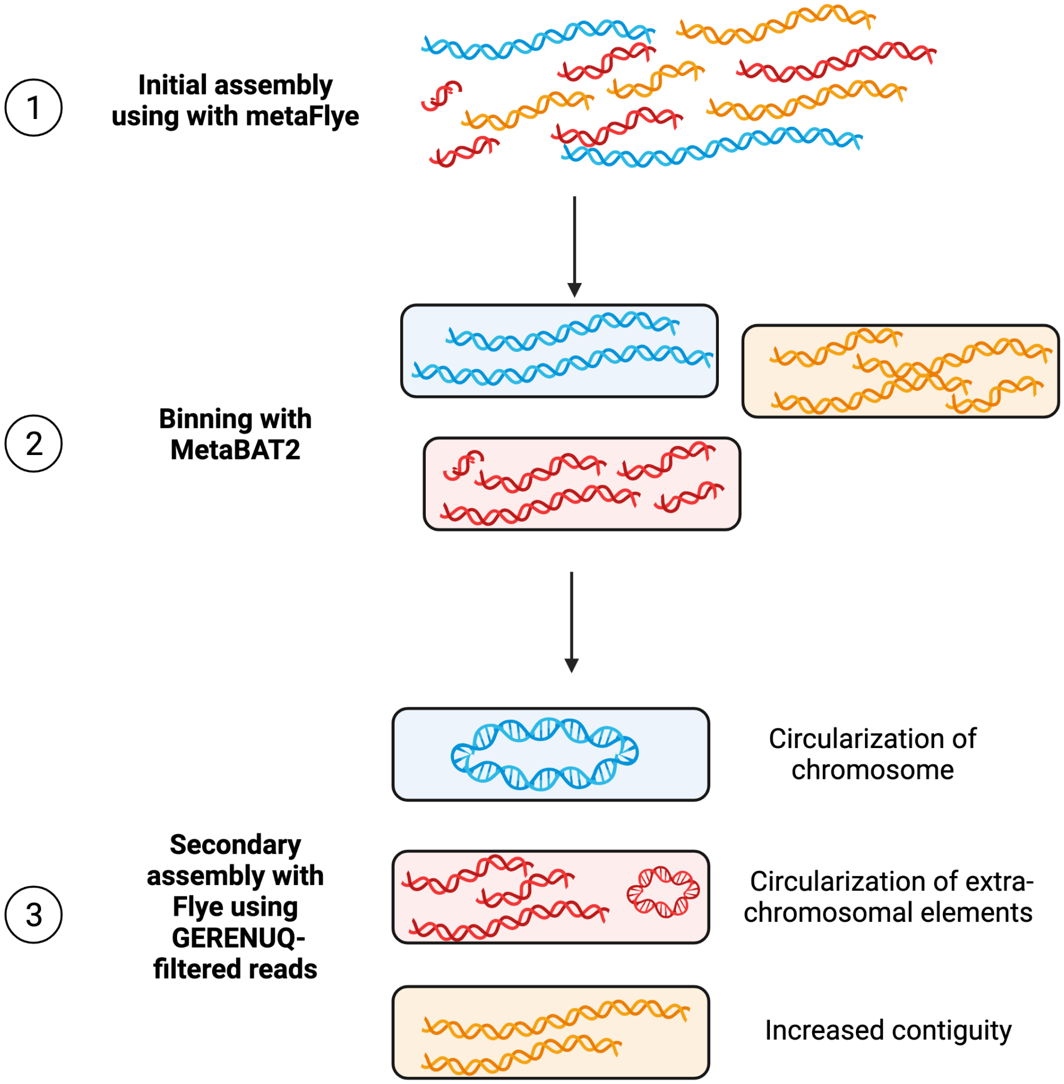

1
Initial assembly using with metaFlye
2
Binning with MetaBAT2
3
Secondary assembly with Flye using GERENUQ-filtered reads
Circularization of chromosome
Circularization of extra-chromosomal elements
Increased contiguity

Figure 5.1: Methodological overview of the assembly pipeline. Initial assembly is performed with metaFlye. Reads are aligned to the bins produced by MetaBAT2 and filtered to create a reduced set of reads to use for a secondary assembly. Secondary assembly is performed with Flye instead of metaFlye because the subset reads being assembled should belong to a single species rather than a complex community of species. There are 3 desired outcomes for the secondary assembly: circularization of chromosomal elements, circularization of extra-chromosomal elements, or an increase in the contiguity of the bin. Created with Biorender.

From every sample, except for the flocculent sample, a contig greater than 1Mb was found to be circularized by metaFlye from the initial metagenomic assembly (Table 5.1). Additionally, an average of nearly 400 sequences under 1Mb per sample were found to be circularized. Highlighting the complexity of the communities that grow on the GAC filters, an average of over 1,100 metagenomic bins per sample were formed by MetaBAT2, which were used for the subsequent secondary assembly. The secondary assembly using subsets of reads that strongly aligned to each bin yielded a number of additional circularized contigs. The total number of new contigs greater than 1Mb that were circularized by the secondary assembly of the bins (66) was greater than the number of contigs found to be circularized by the primary assembly (48). Comparatively, the number of smaller circular contigs in the secondary assembly was far fewer (536) than in the primary assembly (3522). While the secondary assembly was successful in circularizing a number of additional sequences, the overall assembly quality statistics showed a minimal to null improvement. The mean change in N50 was an increase of 40402.27, but with a standard deviation of 441235.1. Change in completion and contamination were largely centred around zero with mean values of 0.051 (SD 3.77) and 1.58 (SD 8.03).

Table 5.1: Summary statistics of the initial and secondary assemblies. Read stats are of the filtered read set used for assembly. Contigs are flagged as being circularized based on the assembly stats output by Flye.

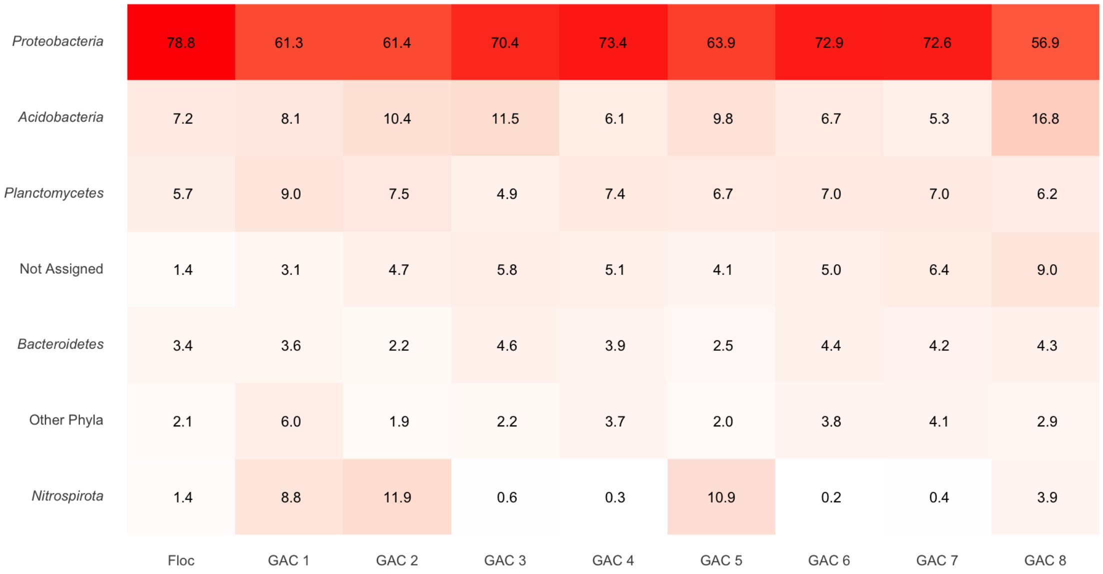
	Floc 1	GAC 1	GAC 2	GAC 3	GAC 4	GAC 5	GAC 6	GAC 7	GAC 8
Read N50	15,046	8,149	17,906	13,614	6,403	19,889	9,515	12,287	12,451
Mean quality score	12.1	13.5	14.2	14.1	13.2	13.9	13.6	13.5	13.5
Initial circular contigs ≥ 1Mb	0	1	13	8	5	5	4	4	8
Initial circular contigs < 1Mb	150	184	628	453	512	450	318	470	357
Number of bins	282	639	1068	1352	1317	834	972	769	783
Secondarily assembled circular contigs ≥ 1Mb	1	3	10	8	11	7	9	8	9
Secondarily assembled circular contigs < 1Mb	11	21	63	105	89	40	74	70	63

To validate that the contigs reported by Flye as being circular were indeed circular a tiling path of reads greater than 5kb in length that overlap by a minimum of 500bp was built for each assembled contig with an additional read that aligned to the start and end of the sequence. For sequences over 1Mb in size, both primarily and secondarily assembled, that were annotated as circularized by Flye had an 83.3% rate of being validated as being circular through the tiling path. For primarily assembled contigs less than 1Mb in size, the success rate was 85.8%, whereas, the success rate for the secondarily assembled contigs less than 1Mb was only 66.4% with the tiling path. Interestingly, across the 9 samples there were an additional 1000 contigs that met the circularization criteria with the tiling path that were not flagged as circular by Flye. This includes an additional 27 contigs that were over 1Mb in size and 574 contigs that were less than 1Mb in size from the secondary assemblies. As well, there were 399 contigs from the unbinned fraction of contigs that were less than 1Mb in size that met the tiling path criteria.

5.3.2 Taxonomic composition and stability of communities

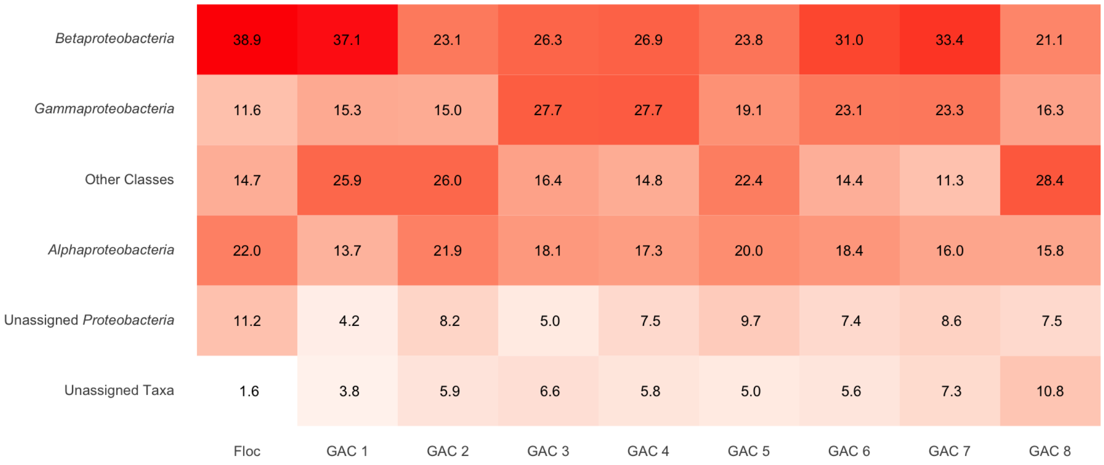
Taxonomy was assigned to each of the contigs and secondarily-assembly bins using the Contig Annotation Tool, which utilizes all protein coding sequences to determine the taxonomy rather than marker genes which are featured in bioinformatic tools such as GTDB. Because many of the contigs featured in the analysis are plasmids without association to genomes that lack the core, single-copy genes, classification by marker gene analysis would struggle to assign taxonomy for them. On the other hand, a total protein coding approach would be much more effective for mobile elements. However, the overall rate of assignment of specific taxonomy was quite low. For contigs and bins over 1Mb, the rate was roughly 12% for genus-level taxonomic assignment, and the rate for contigs less than 1Mb was roughly 14%. This low rate of taxonomic assignment could be attributable to a combination of factors including the novelty of the community, contamination of the bins and assemblies, and the low accuracy of Nanopore reads negatively affecting the open reading frames that are used for the taxonomic assignment. Therefore, higher-level taxonomic classification was used to assess the compositions of the communities. The proportions of reads mapping to contigs and bins assigned at the phylum level showed that the communities are dominated by species belonging to the phylum *Proteobacteria*, which comprised the majority of reads in every sample (Figure 5.2). Other phyla that are well-represented in these samples are *Acidobacteria*, *Planctomycetes*, *Bacteriodetes*, and *Nitrospirota*. The samples that have an apparent lower abundance of species belonging to the *Proteobacteria* have spikes in the abundance of either *Aci-*

dobacteria or *Nitrospirota*. At the class level, there appears to be variability within the classes of *Proteobacteria* with some samples being dominated by the class *Betaproteobacteria* while other samples have equal proportions of *Betaproteobacteria* and *Gammaproteobacteria* (Figure 5.3).



	Floc	GAC 1	GAC 2	GAC 3	GAC 4	GAC 5	GAC 6	GAC 7	GAC 8
Proteobacteria	78.8	61.3	61.4	70.4	73.4	63.9	72.9	72.6	56.9
Acidobacteria	7.2	8.1	10.4	11.5	6.1	9.8	6.7	5.3	16.8
Planctomycetes	5.7	9.0	7.5	4.9	7.4	6.7	7.0	7.0	6.2
Not Assigned	1.4	3.1	4.7	5.8	5.1	4.1	5.0	6.4	9.0
Bacteroidetes	3.4	3.6	2.2	4.6	3.9	2.5	4.4	4.2	4.3
Other Phyla	2.1	6.0	1.9	2.2	3.7	2.0	3.8	4.1	2.9
Nitrospirota	1.4	8.8	11.9	0.6	0.3	10.9	0.2	0.4	3.9

Figure 5.2: Heat map of the percentage of reads mapping to contigs or bins assigned at the phylum level. Taxonomy was assigned to the circularized contigs and bins using CAT and truncated to the phylum level. Reads were aligned to the contigs and bins using minimap2 and quantified using SAMtools.



	Floc	GAC 1	GAC 2	GAC 3	GAC 4	GAC 5	GAC 6	GAC 7	GAC 8
Betaproteobacteria	38.9	37.1	23.1	26.3	26.9	23.8	31.0	33.4	21.1
Gammaproteobacteria	11.6	15.3	15.0	27.7	27.7	19.1	23.1	23.3	16.3
Other Classes	14.7	25.9	26.0	16.4	14.8	22.4	14.4	11.3	28.4
Alphaproteobacteria	22.0	13.7	21.9	18.1	17.3	20.0	18.4	16.0	15.8
Unassigned *Proteobacteria*	11.2	4.2	8.2	5.0	7.5	9.7	7.4	8.6	7.5
Unassigned Taxa	1.6	3.8	5.9	6.6	5.8	5.0	5.6	7.3	10.8

Figure 5.3: Heat map of the percentage of reads mapping to contigs or bins assigned at the class level with a focus on the classes belonging to the phylum *Proteobacteria*. Taxonomy was assigned to the circularized contigs and bins using CAT and truncated to the class level. Reads were aligned to the contigs and bins using minimap2 and quantified using SAMtools.

Taxonomy was also assigned to 390 circularized contigs and high-quality bins across the 9 samples using the GTDB-Tk classify workflow in an attempt to get a higher resolution taxonomy for elements that should contain the marker genes used by GTDB-Tk for taxonomic assignment. While taxonomic assignment at the genus level was more successful than with CAT, there was still a overall lack of taxonomic resolution for these largely uncharacterized bacteria. Overall, 73.6% of the contigs and bins were assigned to some degree at the genus level, but the quality of the assignments was not necessarily consistent. For instance, the most common assignment at the genus level was OLB17, which was identified 18 times. This is an uncharacterized genus with the family *Pyrinomonadaceae*. Another example is the second-most common genus-level assignment of JAEUIA01, which is only characterized as belonging to the phylum *Planctomyce-*

tota. However, roughly 20% of the contigs and bins had taxonomic assignments to characterized genera including: *Rugosibacter, Ferruginibacter, Terricaulis, Accumulibacter, Manganitrophus, Macondimonas, Methylotenera, Sphingobium* and *Hyphomicrobium*.

To additionally assess the stability of the communities, contigs and bins that were ≥ 1Mb in size were clustered at a 95% sequence identity threshold, which is the rough estimate for the species boundary in bacteria. Overall, there was a high proportion of shared species across the samples (Figure 5.4). There were 10 clusters of contigs and bins that were shared between all 9 samples and 84 total clusters that were found to be present in 5 or more samples. While there were still a number of these contigs and bins that were unique to samples, the majority were shared between multiple samples.

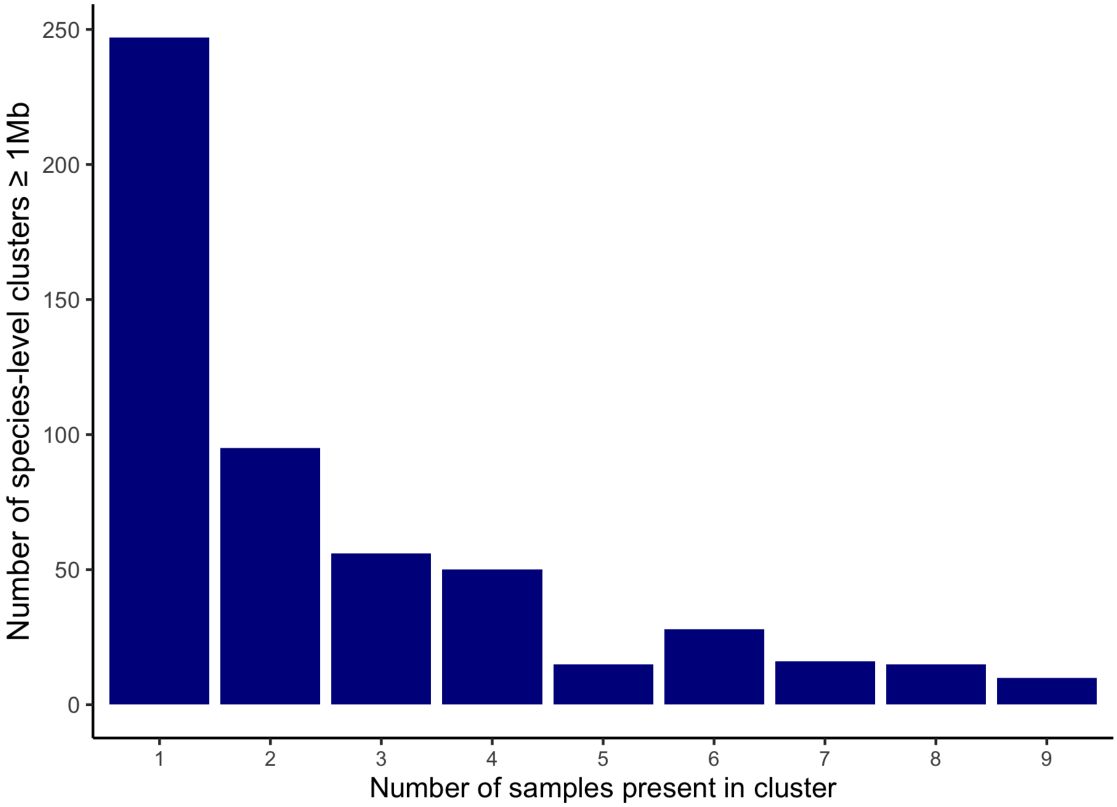


Figure 5.4: Plot of the clusters of species-level (95%) identity for contigs and bins greater than 1Mb in size and the proportion of the samples that they are found in. Those species present in 1 sample are unique to that sample and those found in 9 samples are species present in all samples.

5.3.3 Annotation of mobile elements and hydrocarbon metabolism genes

Featured among the assembled contigs and bins was a large number of contigs that were annotated as being plasmid sequences (Table 5.2). In addition, there were a number of contigs that contained pHMM matches for relaxase proteins, type IV coupling proteins, and type IV secretion system proteins, which is indicative of the contigs being conjugative elements. Of the predicted plasmid sequences, 13.7% of them were circular. For the conjugative systems, 22.4% of the sequences were annotated as circular by Flye.

Table 5.2: Summary statistics for the annotation of the circularized contigs, unbinned contigs, and secondarily-assembled bins. Plasmid-based contigs identified using PlasFlow and conjugative elements determined by matches to Pfam pHMMs of relaxases, type IV coupling proteins, and type IV secretion system proteins.

Sample ID	Plasmids	Conjugative elements	Percent with CANT-HYD annotations	Percent with AromaDeg annotations
Floc 1	213	41	29.9	20.3
GAC 1	501	21	22.6	17.9
GAC 2	879	100	24.1	18.0
GAC 3	1842	94	18.6	14.6
GAC 4	1908	71	12.8	9.2
GAC 5	676	73	23.5	17.3
GAC 6	1162	55	17.4	13.3
GAC 7	969	46	15.5	11.6
GAC 8	962	67	21.1	16.4

Genes responsible for the metabolism of aromatic hydrocarbon were observed to be highly abundant throughout the communities (Table 5.2), which is expected given the highly contaminated environment that these microbes grow in. All contigs and bins were annotated with two separate hydrocarbon databases, CANT-HYD and AromaDeg, to assess the commonality of these metabolic genes in the communities. For the AromaDeg database, a mean of 15.4% (SD 3.5) of contigs and bins contained at least one annotation for a gene within the database. A mean of 20.6% (SD 5.1) of contigs and bins contained an alignment to an entry within the CANT-HYD database.

5.4 Discussion

Mirroring the general algorithmic principles presented in Jorg [9], we were able to circularize a number of contigs from complex communities that colonize the wastewater treatment facilities of an oil refinery using long reads. The number of circularized contigs over 1Mb in size more than doubled by performing a secondary assembly using the reads that strongly aligned to the bins. GERENUQ was used to filter the reads and eliminate that cross map between species or chimeric reads that were erroneous sequenced by the Nanopore MinION [16]. Comparatively, contigs under 1Mb were not as successfully circularized by the secondary assembly compared to the primary assembly with only a 13% increase in the overall number of circularized contigs under 1Mb. This could be related to the phenomenon of mobile genetic elements, such as plasmids, being systematically excluded from metagenomic bins by algorithms such as MetaBAT2 [11, 12]. Because plasmids would be the target for circularization for contigs under 1Mb, their exclusion from the bins used for the secondary assembly explains the poor yield of the small, circular elements in the secondary assemblies. Indeed, roughly 24% of the plasmids predicted by PlasFlow are in the unbinned fraction of contigs. The systematic exclusions of plasmid sequences by binning algorithms highlights the need to identify them using tools such as PlasFlow and to circularize their sequences for easy recognition in a complex environment.

Taxonomic assignment of the contigs proved difficult given the novelty of the member species of these communities. As such only a small proportion of all contigs and bins could get a genus-level taxonomic assignment using a protein coding sequence-based tool in CAT. Even using the gold standard marker gene-based taxonomic assignment tool (GTDB TK) for the large-cicularized contigs and high-quality bins, many sequences could not be assigned with a high-resolution taxonomic assignment. However, for the contigs and bins that did have genus-level taxonomies assigned, the genera have previously been shown to possess metabolic capabilities that would enable growth in the waste water and filters found at an oil refinery. For example, nine bins were assigned the taxonomy of *Rugosibacter*, a genus that has been shown to be capable of using both monoaromatic and polyaromatic hydrocarbons for growth [30]. Other genera that were found in these samples such as *Sphingobium* [31, 32, 33], *Macondimonas* [34], and *Ferruginibacter* [35] have also been found in the literature to grow in hydrocarbon-contaminated wastewater. The potential of these bacteria to metabolize the hydrocarbons found in the environment is substantiated by the frequency of annotation for hydrocarbon degradation genes throughout the microbiome. CANT-HYD genes are found in over 20% of the assembled contigs and bins and AromaDeg annotations are found in over 15% (Table 5.2). It is clear that these complex and poorly characterized communities are involved in the metabolism and removal of hydrocarbon from the wastewater. Other members of the community may also participate in the bioremdiation of heavy metals from the environment.

In the GAC samples, there are species belonging to the genus *Manganitrophus*, whose member species have been shown to oxidize manganese as a source of energy and also contain genes related to the transport of a number of other heavy metals [36, 37]. *Ferruginibacter* have also been found to harbour genes for mercury transport and detoxification [38]. With the high concentrations of heavy metals in the oil refinery effluent, genes related to heavy metal transport and biotransformation are equally as important for bioremediation as genes related to hydrocarbon degradation.

The communities, which were sequenced over a period of over 2 years from the SunCor Energy Sarnia Canada facility, showed a high degree of consistency. Across all nine samples there was a consistent dominance of the phylum *Proteobacteria* (Figure 5.2). Some variation could be seen sample-to-sample in the relative abundances of the *Acidobacteria* and *Nitrospirota*, the latter of which is primarily composed of contigs and bins belonging to the manganese-oxidizing genus *Manganitrophus*. Further investigation into the association of the abundances of metals, such as manganese, in the effluent and the relative abundance *Manganitrophus* is warranted. Perhaps the best proxy for the stability of the complex communities is the shared proportion of contigs and bins over the size of 1Mb across samples (Figure 5.4). Using a sequence identity cutoff of 95%, which has been established as the rough boundary for species [19], 10 species clusters could be found in all samples. The majority of species clustered by dRep could be found in more than one of the samples, highlighting the shared features of these communities. Given the shared number of species in these communities, a pooled assembly might also increase the yield of high-quality or completed sequences at the cost of chimeric assemblies [39]. dRep clustering of the large contigs and bins also highlighted the extraordinary complexity of the communities with a total of 532 unique species-level clusters forming from the assemblies. The success of long read sequencing and the secondary assembly strategy in circularizing chromosomal and extra-chromosomal sequences from such complex communities speaks to their usefulness for metagenomic studies.

5.5 Conclusion

Secondary assembly of metagenomic bins is a successful computational strategy to maximize the number of complete, circular sequences generated from metagenomic assemblies with long reads. There is still much work to optimize the algorithm to improve the overall performance for general use cases. In Jorg, a similar approach for short-read assemblies, there are checks in place to evaluate in real-time whether the secondary assembly of the bin circularized the sequence, improved the assembly, or decreased the quality of the assembly. For many of the bins in this study, the secondary assembly did not improve general assembly quality statistics or circularize sequences, and the original bin would have been more suitable for analysis. Furthermore, some bins may have benefited from a tertiary assembly. For these cases, the extra quality checks and recursive assembly featured in Jorg may serve as an improvement over the presented workflow. However, the additional circularized genetic elements for these likely hydrocarbon-degrading communities will serve as quality scaffolds for future metatranscriptomic and community dynamic research to help elucidate their potential in bioremediation.

5.6 References

Bibliography

[1] Leho Tedersoo, Mads Albertsen, Sten Anslan, and Benjamin Callahan. Perspectives and Benefits of High-Throughput Long-Read Sequencing in Microbial Ecology. *Applied and Environmental Microbiology*, 87(17):e0062621, August 2021.

[2] Mikhail Kolmogorov, Derek M. Bickhart, Bahar Behsaz, Alexey Gurevich, Mikhail Rayko, Sung Bong Shin, Kristen Kuhn, Jeffrey Yuan, Evgeny Polevikov, Timothy P. L. Smith, and

Pavel A. Pevzner. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods*, 17(11):1103–1110, November 2020.

[3] Caitlin M. Singleton, Francesca Petriglieri, Jannie M. Kristensen, Rasmus H. Kirkegaard, Thomas Y. Michaelsen, Martin H. Andersen, Zivile Kondrotaite, Søren M. Karst, Morten S. Dueholm, Per H. Nielsen, and Mads Albertsen. Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing. *Nature Communications*, 12(1):2009, March 2021.

[4] Lei Liu, Yulin Wang, Yu Yang, Depeng Wang, Suk Hang Cheng, Chunmiao Zheng, and Tong Zhang. Charting the complexity of the activated sludge microbiome through a hybrid sequencing strategy. *Microbiome*, 9(1):205, December 2021.

[5] Eli L. Moss, Dylan G. Maghini, and Ami S. Bhatt. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nature Biotechnology*, 38(6):701–707, June 2020.

[6] Daniel J Giguere, Alexander T Bahcheli, Benjamin R Joris, Julie M Paulssen, Lisa M Gieg, Martin W Flatley, and Gregory B Gloor. Complete and validated genomes from a metagenome. preprint, Bioinformatics, April 2020.

[7] Derek M. Bickhart, Mikhail Kolmogorov, Elizabeth Tseng, Daniel M. Portik, Anton Korobeynikov, Ivan Tolstoganov, Gherman Uritskiy, Ivan Liachko, Shawn T. Sullivan, Sung Bong Shin, Alvah Zorea, Victòria Pascal Andreu, Kevin Panke-Buisse, Marnix H. Medema, Itzhak Mizrahi, Pavel A. Pevzner, and Timothy P. L. Smith. Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities. *Nature Biotechnology*, 40(5):711–719, May 2022.

[8] Roye Rozov, Aya Brown Kav, David Bogumil, Naama Shterzer, Eran Halperin, Itzhak Mizrahi, and Ron Shamir. Recycler: an algorithm for detecting plasmids from *de novo* assembly graphs. *Bioinformatics*, page btw651, December 2016.

[9] Lauren M. Lui, Torben N. Nielsen, and Adam P. Arkin. A method for achieving complete microbial genomes and improving bins from metagenomics data. *PLOS Computational Biology*, 17(5):e1008972, May 2021.

[10] Yi Yue, Hao Huang, Zhao Qi, Hui-Min Dou, Xin-Yi Liu, Tian-Fei Han, Yue Chen, Xiang-Jun Song, You-Hua Zhang, and Jian Tu. Evaluating metagenomics tools for genome binning with real metagenomic datasets and CAMI datasets. *BMC bioinformatics*, 21(1):334, July 2020.

[11] Finlay Maguire, Baofeng Jia, Kristen L. Gray, Wing Yin Venus Lau, Robert G. Beiko, and Fiona S. L. Brinkman. Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic Islands. *Microbial Genomics*, 6(10), October 2020.

[12] Benjamin R. Joris, Tyler S. Browne, Thomas A. Hamilton, David R. Edgell, and Gregory B. Gloor. Identification of type IV conjugative systems that are systematically excluded from metagenomic bins. preprint, In Review, March 2022.

[13] Wouter De Coster, Svenn D'Hert, Darrin T Schultz, Marc Cruts, and Christine Van Broeckhoven. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*, 34(15):2666–2669, August 2018.

[14] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, September 2018.

[15] Dongwan D. Kang, Feng Li, Edward Kirton, Ashleigh Thomas, Rob Egan, Hong An, and Zhong Wang. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, 7:e7359, 2019.

[16] abahcheli. GERENUQ, October 2021. original-date: 2020-10-31T15:13:33Z.

[17] Donovan H. Parks, Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, and Gene W. Tyson. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 25(7):1043–1055, July 2015.

[18] Matthew R. Olm, Christopher T. Brown, Brandon Brooks, and Jillian F. Banfield. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *The ISME journal*, 11(12):2864–2868, December 2017.

[19] Matthew R. Olm, Alexander Crits-Christoph, Spencer Diamond, Adi Lavy, Paula B.

Matheus Carnevali, and Jillian F. Banfield. Consistent Metagenome-Derived Metrics Verify and Delineate Bacterial Species Boundaries. *mSystems*, 5(1):e00731–19, February 2020.

[20] F. A. Bastiaan von Meijenfeldt, Ksenia Arkhipova, Diego D. Cambuy, Felipe H. Coutinho, and Bas E. Dutilh. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biology*, 20(1):217, December 2019.

[21] Pierre-Alain Chaumeil, Aaron J Mussig, Philip Hugenholtz, and Donovan H Parks. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, page btz848, November 2019.

[22] A. Murat Eren, Özcan C. Esen, Christopher Quince, Joseph H. Vineis, Hilary G. Morrison, Mitchell L. Sogin, and Tom O. Delmont. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*, 3:e1319, October 2015.

[23] Michael Y. Galperin, Yuri I. Wolf, Kira S. Makarova, Roberto Vera Alvarez, David Landsman, and Eugene V. Koonin. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Research*, 49(D1):D274–D281, January 2021.

[24] Pawel S. Krawczyk, Leszek Lipinski, and Andrzej Dziembowski. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Research*, 46(6):e35, April 2018.

[25] Sean R. Eddy. Accelerated Profile HMM Searches. *PLoS Computational Biology*, 7(10):e1002195, October 2011.

[26] Doug Hyatt, Gwo-Liang Chen, Philip F LoCascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1):119, December 2010.

[27] Márcia Duarte, Ruy Jauregui, Ramiro Vilchez-Vargas, Howard Junca, and Dietmar H. Pieper. AromaDeg, a novel database for phylogenomics of aerobic bacterial degradation of aromatics. *Database*, 2014, January 2014.

[28] Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment

using DIAMOND. *Nature Methods*, 12(1):59–60, January 2015.

[29] Varada Khot, Jackie Zorz, Daniel A. Gittins, Anirban Chakraborty, Emma Bell, María A. Bautista, Alexandre J. Paquette, Alyse K. Hawley, Breda Novotnik, Casey R. J. Hubert, Marc Strous, and Srijak Bhatnagar. CANT-HYD: A Curated Database of Phylogeny-Derived Hidden Markov Models for Annotation of Marker Genes Involved in Hydrocarbon Degradation. *Frontiers in Microbiology*, 12:764058, January 2022.

[30] Elizabeth M. Corteselli, Michael D. Aitken, and David R. Singleton. Rugosibacter aromaticivorans gen. nov., sp. nov., a bacterium within the family Rhodocyclaceae, isolated from contaminated soil, capable of degrading aromatic compounds. *International Journal of Systematic and Evolutionary Microbiology*, 67(2):311–318, February 2017.

[31] Jiro F. Mori and Robert A. Kanaly. Complete Genome Sequence of Sphingobium barthaii KK22, a High-Molecular-Weight Polycyclic Aromatic Hydrocarbon-Degrading Soil Bacterium. *Microbiology Resource Announcements*, 10(1):e01250–20, January 2021.

[32] Madhumita Roy and Tapan K. Dutta. Draft Whole-Genome Sequence of Sphingobium sp. Strain PNB, a Versatile Polycyclic Aromatic Hydrocarbon-Degrading Bacterium. *Microbiology Resource Announcements*, 10(48):e0092021, December 2021.

[33] Quanfeng Liang and Gareth Lloyd-Jones. Sphingobium scionense sp. nov., an aromatic hydrocarbon-degrading bacterium isolated from contaminated sawmill soil. *International Journal of Systematic and Evolutionary Microbiology*, 60(Pt 2):413–416, February 2010.

[34] Smruthi Karthikeyan, Luis M. Rodriguez-R, Patrick Heritier-Robbins, Minjae Kim, Will A. Overholt, John C. Gaby, Janet K. Hatt, Jim C. Spain, Ramon Rosselló-Móra, Markus Huettel, Joel E. Kostka, and Konstantinos T. Konstantinidis. ”Candidatus Macondimonas diazotrophica”, a novel gammaproteobacterial genus dominating crude-oil-contaminated coastal sediments. *The ISME journal*, 13(8):2129–2134, August 2019.

[35] Ryoich Tanaka, Katsuya Nouzaki, Ronald R. Navarro, Tomohiro Inaba, Tomo Aoyagi, Yuya Sato, Atsushi Ogata, Hiroshi Yanagishita, Tomoyuki Hori, and Hiroshi Habe. Activated sludge microbiome in a membrane bioreactor for treating Ramen noodle-soup wastewater.

The Journal of General and Applied Microbiology, 66(6):339–343, February 2021.

[36] Hang Yu and Jared R. Leadbetter. Bacterial chemolithoautotrophy via manganese oxidation. *Nature*, 583(7816):453–458, July 2020.

[37] Hang Yu, Grayson L. Chadwick, Usha F. Lingappa, and Jared R. Leadbetter. Comparative genomics on cultivated and uncultivated, freshwater and marine *Candidatus* Manganitrophaceae species implies their worldwide reach in manganese chemolithoautotrophy. preprint, Microbiology, November 2021.

[38] Shalini Porwal and Rajni Singh. Cloning of merA Gene from Methylotenera Mobilis for Mercury Biotransformation. *Indian Journal of Microbiology*, 56(4):504–507, December 2016.

[39] Jonathan D. Magasin and Dietlind L. Gerloff. Pooled assembly of marine metagenomic datasets: enriching annotation through chimerism. *Bioinformatics (Oxford, England)*, 31(3):311–317, February 2015.

Chapter 6

General Discussion

The resolution of the structure of DNA [1] and the establishment of the central dogma of biology [2] spurred a technological arms race to sequence genomes to unravel the mysteries of nature. First-generation sequencing technology permitted the curation of the first genomic sequence of the RNA bacteriophage MS2 in 1976 [3] and in 1977 the first DNA genome of the bacteriophage phi X174 [4], which is still used as a sequencing control to this day. Technological and computational advances continued throughout the era of first-generation sequencing, which allowed for larger and larger genomes to be assembled, eventually culminating in the assembly of the human genome by the International Human Genome Sequencing Consortium [5]. Second-generation sequencing technologies vastly increased the information capacity of genome sequencing, which enabled an increased complexity of sequencing experiments. One such avenue of research that was enabled by second-generation sequencing was shotgun metagenomic sequencing, which is capable of reconstructing the structure and metabolic capacity of complex bacterial communities [6, 7]. Recently, third-generation sequencing has been developed which enabled the telomere-to-telomere resolution of the human genome [8], completion of other eukaryotic genomes, and the circularization of bacterial genome assemblies from complex metagenomes [9, 10, 11]. In the interceding years between the generational leaps in sequencing technology, iterative improvements in algorithms and bioinformatic workflows optimize the information that can be obtained from the data.

6.1 Exclusion of conjugative elements from metagenomic analyses

In previous research, it has been shown that plasmids and genetic islands are excluded from metagenomic bins. Maguire and colleagues [12] found that only 38-44% of genomic island sequences and 1-29% of plasmid sequences were retained within MAGs and that nearly all plasmid-borne sequences for antimicrobial resistance were omitted from MAGs. In Chapter 2, I independently confirmed these findings in the subset of plasmids and genomic islands that are conjugative from 101 human gut microbiome samples (Figures 2.4 and 2.5). The rate of exclusion from MAGs for these conjugative elements falls within the range that Maguire and colleagues observed generally for plasmids, which suggests that conjugative elements are an area of weakness for metagenomic analyses that needs to be improved.

Conjugation mediated by type IV secretion systems is one of the ways that bacteria are able to exchange DNA with one another. This genetic information exchange is critical for bacterial adaptation to their environment [13], which makes their identification critical for understanding the dynamics of communities of bacteria. Antimicrobial resistance genes are commonly genetic cargo carried on conjugative plasmids or integrative and conjugative elements that are maintained by positive selection [14, 15]. Antimicrobial resistance is of major societal concern and being able to properly capture it in metagenomic sequencing analyses, particularly those involving human or animal sampling, is important for public health surveillance. For instance, reducing antibiotic use in livestock feed is a major focus to stop the spread of antimicrobial resistance due to the proliferation of antimicrobial-resistant bacteria in the gut of livestock [16]. However, if surveillance involves using metagenomic binning to find the bacterial genetic elements responsible, the investigator will miss the vast majority of the sequences of interest. Beyond antimicrobial resistance and virulence factors, conjugative elements are host to a wide array of other pathways such as bile salt detoxification, metal resistance, and polysaccharide usage [17]. Given this diversity of metabolic potential known to be carried on these elements, and likely a large unrecognized diversity of metabolism due to their exclusion from MAGs, conjugative elements could be a missing link in many microbiome studies and will be missed or misunderstood in a MAG-focused approach to analysis.

Chapter 2 demonstrates that MAGs generated using current, well-established techniques are not sufficient due to their exclusion of conjugative elements. Metagenomic bins were generated by following the protocol used to produce the Unified Human Gastrointestinal Genome (UHGG) collection [18, 19]. High-quality and well-accepted protocols have a blind spot for conjugative elements, and mobile genetic elements in general. To overcome these shortcomings, in chapter 2, I developed a framework for recovering these sequences from raw metagenomic assemblies. Using a database of profile hidden Markov models that contained models for relaxase, T4CPs, and T4SS protein, I was able to recover a fraction of the conjugative elements present in the assemblies. Predicted protein sequences from the assemblies were also aligned to the UniRef90 [20] database using DIAMOND [21], which yielded a greater number of identified type IV conjugative systems than using pHMMs at the cost of increased computational resources required. Implementation of this framework is simple in practice: identify the conjugative systems prior to binning and treat the conjugative systems not included in bins in the same manner as bins. Optimizations to this protocol could be in the form of increasing the diversity of pHMMs in the database of type IV conjugative proteins to increase the sensitivity or to parse the UniRef databases to only include the proteins of interest to reduce computation time. Unfortunately, these methods are not capable of discriminating between functional and non-functional conjugative elements. Checks for integrity of the synteny of the conjugative genes and the presence of oriT sequence may allow for greater confidence on the activity of the conjugative element. Additionally, transcriptomics could reveal which conjugative elements are being actively transcribed in a community. Implementation of third-generation sequencing technologies can also improve the analysis of conjugative elements. For example, because of the ability to circularize genetic elements with long-read sequencing data [9, 10, 11], the investigator can unambiguously place the genomic context (i.e. as a plasmid or integrative element). Additionally, using methylation data it is possible to bin plasmid and cognate chromosome together [22], which in this thesis was shown to be unobtainable with sequence composition and coverage binning methods. In theory, the plasmid could be associated with the cognate chromosome through taxonomy, but that is difficult to accomplish in a complex community.

6.2 Separation of geographic cohorts by conjugative elements

Chapter 3 builds on a question raised by chapter 2: are conjugative systems differential between populations? The human gut microbiome has been broadly associated with various human health concerns ranging from gastrointestinal to neurological [23]. The human gut microbiome has also shown broad differences between geographically-based cohorts [24]. If the differences in the broader microbiota are reflected in the composition of the conjugative elements, or perhaps equally interesting if the differences are not reflected, then it is imperative that conjugative systems are properly assessed in microbiome analyses. Metabolic pathways carried on these plasmids or integrative elements could indeed help to explain host phenotypes.

The data presented in chapter 3 demonstrates that conjugative elements identified by the methods developed in chapter 2 are differential between geographically-focused cohorts (Figure 3.3), reflecting what had been found in the broader microbiome. These finding again reinforce the need to include conjugative elements separately in metagenomic analyses. In chapter 2, one of the two cohorts used to develop the protocol for identifying conjugative elements, and proving their exclusion from MAGs, was a cohort of North American pre-term infants. Given that in chapter 2 I was able to identify 96 and 242 systems from this cohort using the pHMM and UniRef alignment methods, respectively, it is clear that there are indeed conjugative systems present in these samples. However, figure 3.2 demonstrates little-to-no signal in these samples when mapped to the conjugative elements that were identified from a human gut reference database, which is suggestive of a severe under-representation of these sequences in the database. Pre-term infants are given a constant dose of antibiotics from birth to stave off infection, and an under-representation of the conjugative plasmids in databases that could be carrying antimicrobial resistance genes is problematic for surveillance and patient care efforts. Conjugative plasmids with antimicrobial resistance genes carried by pathogenic *Proteobacteria* could pose a serious threat to the fragile health of pre-term infants and if their sequences are missing in the databases, then antimicrobial resistance screening may yield false negatives. Fortunately, *Proteobacteria*, and their resistance genes, are generally well-studied in isolate experiments and tools for antimicrobial resistance are not necessarily tied to metagenomic binning, so the impact of such a scenario is lessened. However, for the building of near-complete and non-redundant databases, the omission of conjugative plasmids could create a false sense of security for antimicrobial resistance. In figure 3.2 there is also a less diverse pattern of abundance for the West African and South American cohorts, which represent cohorts that are non-industrialized and likely have lower antibiotic use in the population. Whether this unique pattern of abundance of conjugative elements is a result of a lack of exposure to antibiotics or rather another bias in the database would require additional analysis. The human gut reference set of genomes was built off a comprehensive set of metagenomic samples, but there are few samples for the African and South American regions indexed on databases such as Data Repository For Human Gut Microbiota. Further research could illuminate the dynamics of conjugative systems and the prevalence of antimicrobial resistance in these populations and whether there are metabolic pathways on their conjugative systems that have not been observed in Western cohorts. For instance, with the divergent dietary tendencies between these cohorts and Western cohorts, there could be a greater diversity of polysaccharide-scavenging genes that have been previously observed as cargo on conjugative elements [17].

One weakness of the geographic cohort analysis in chapter 3 was the use of conjugative systems found in a database of MAGs as a way of measuring the abundances of conjugative systems. Being tied to the MAGs creates biases and gives an incomplete picture of the true abundances of conjugative elements due to the noted omission of the majority of conjugative elements found within these samples. However, assembly and annotation of 785 metagenomic samples was not feasible with the computational resources available, so the compromise was made to use the reference genome database. To remedy this shortcoming, reads were also mapped to the conjugative elements identified by pHMMs and MAGs whose taxonomy was assigned using CAT [25] from chapter 2. These data showed broad concordance between the proportions of reads mapping to each phyla, with some minute differences in composition and a noticeable increase in proportion of reads mapping to elements that could not have their taxonomy assigned (Figure 6.1). This is yet another reflection of the under-representation of these sequences in the databases as many have their cognate genome's taxonomy assigned, but were not published within the MAG to the database.

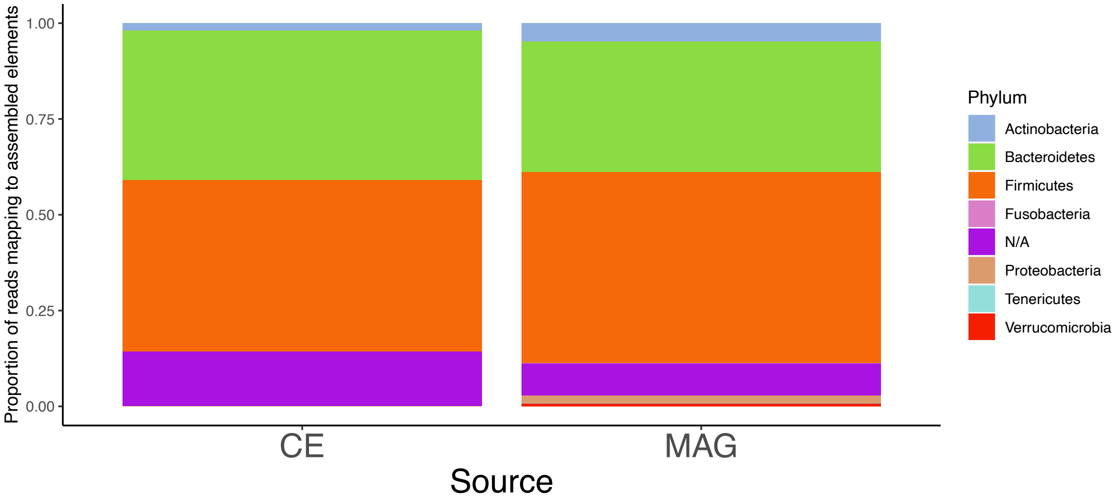


Figure 6.1: Proportions of reads mapping the the conjugative elements and MAGs within the general cohort samples assembled in chapter 2. Taxonomy of conjugative elements (CE) and metagenome-assembled genomes (MAG) we assigned using the program CAT.

6.3 The microbiome of spina bifida

In chapter 3, I also sought to apply the methods I developed in chapter 2 to a novel health-focused application. I looked to build on the earlier differences found between cohorts to prove the potential utility of including conjugative elements alongside a standard metagenomic bin analysis. For this, I utilized data from a collaboration looking at the composition of the human gut microbiome and its relation to spina bifida, which is a complex disease with a poorly understood pathogenesis. Some of the development of the disorder can be attributed to genetics, but much of the risk is unexplained by genetics alone [26]. Much of the variation in the pathogenesis has been also attributed to environmental and nutritional sources. For instance, maternal plasma levels of folate [27], vitamin

B12 [27, 28], methionine [29], choline [30], vitamin C [31], and zinc [32] are all associated with the development of spina bifida in the fetus. Additionally, maternal infection and inflammation are also risk factors for development. The bacteria that line the human gut are heavily involved in the synthesis of many vitamins and nutrients [33], so it spurred us to test the hypothesis that the composition of human gut microbiota may be associated with the development of spina bifida.

Indeed, the data presented in chapter 3 suggest that there is an association between the human gut microbiome and the pathogenesis of spina bifida. From the analysis of the binned genomes, *Campylobacter hominis* had the strongest enrichment in the mothers who gave birth to infants with spina bifida. To date, this is the first association of *Campylobacter hominis* with disease as a result of its colonization of the human intestinal tract. Other species of the genus *Campylobacter* are known pathogens of the intestinal tract of humans, so it is possible that the presence of these bacteria are inducing a proinflammatory state. Inflammation in the human gut is known to increase the ‘leakiness’ of the gut and impair nutrient uptake [34]. For a disease that is heavily associated with a lower serum concentration of many key vitamins and nutrients, this impairment of nutrient uptake could help explain the development of spina bifida following nutrient supplementation during pregnancy. In addition to *Campylobacter* there was also an enrichment of a MAG belonging to the genus *Peptoniphilus* and enrichments of genes belonging to MAGs in the genera *Ruminococcus* and *Porphyromonas*. *Peptoniphilus* has been associated with blood infections [35] and *Ruminococcus* and *Porphyromonas* are associated with the pro-inflammatory gut microbiome of individuals with obesity. The enrichment of these taxa are all indicative that gut inflammation is a important risk factor to consider as part of the pathogenesis of spina bifida in expectant mothers.

To tie in conjugative systems into this analysis, conjugative elements were predicted from the raw assemblies using the pHMM approach proposed in chapter 2. The strongest enrichment in the mothers who gave birth to infants with spina bifida was once again an element belonging to the species *Campylobacter hominis*. However, this conjugative element had a much lower effect size than the MAG of *Campylobacter hominis* did. Perhaps the conjugative element is not involved in the pathogenesis of the disease and rather just commonly carried by the bacteria that contribute to the disease. The conjugative element does not appear to have any cargo that would be related to the development of spina bifida with the vast majority of genes being related to conjugation and mobilization of the element. However, this conjugative plasmid or integrative and conjugative element could be a vector to express genes of interest within *Campylobacter hominis*. Recently, it has been shown that conjugative plasmids can be constructed *de novo* using one of the sequences identified from the human gut reference database set of bacterial MAGs (Thomas A. Hamilton, personal communication). This plasmid was able to conjugate with greater efficiency into its host species than to other species. Using a CRISPR killing array previously developed and tested [36], it could be possible to selectively kill *Campylobacter hominis in vivo* in the intestinal tract of expectant mothers to help prevent, or lessen, the inflammation that may be associated with the pathogenesis of spina bifida.

The involvement of *Campylobacter hominis* in the pathogenesis of spina bifida needs to be confirmed with a larger-scale study that optimally utilized metatranscriptomics. The lower sample size prevented the corrected p-values from falling below the accepted level of significance (despite a robust effect size), so a large sample could act as confirmation. Transcriptomics would add functional information for the pathogenesis of *Campylobacter hominis* in spina bifida that could not be obtained by metagenomics alone. For instance, which genes are being actively transcribed and does the transcriptional activity of certain genes correlate with intestinal inflammation.

6.4 Complete genomes and conjugative plasmids of strains from manganese-oxidizing genus

A recently described genus of bacteria expanded what was known to be metabolically possible in nature by utilizing the oxidation of manganese to fuel the reverse tricarboxylic acid cycle [37]. Oxidation of manganese as a source of energy had been theorized to be possible in the past though not previously observed in nature. The first genome of the genus *Manganitrophus* was published in 2020 for the species *Candidatus Manganitrophus noduliformans*, which was enriched from tap water [37]. In 2021, a preprint was published with two new strains belonging to the species *Candidatus Manganitrophus morganii*, which were isolated from a rock surface in South Africa and a rusted iron pipe in California [38]. In chapter 4, I uncovered the sequences from five novel strains from another dissimilar environment. These novel strains are a member of an extremely complex biofilm that grows on the charcoal filters in a wastewater treatment facility for the Suncor oil refinery in Sarnia, Canada. The presence of these species belonging to the genus *Manganitrophus* in four dissimilar environments raises the question of 'how prevalent are these bacteria globally?' As well, how many other contaminant metals are bacteria able to oxidize for energy beyond known metals such as iron and now manganese?

Third-generation sequencing of the communities with the Oxford Nanopore MinION allowed for circularization of the first *Candidatus Manganitrophus* strain from the first sample, but failed with the following samples. As shown in the full community analysis in chapter 5, these are extremely complex communities, which pose a challenge for metagenomic assemblies. The complexity increases the probabilities that the assembly will not be completed due to repetitive sequences found in multiple species within the community. I therefore needed to employ a more targeted approach than a metagenomic assembly for the final 4 samples to yield high-quality or circularized genomes. As a solution, I utilized a reference-guided assembly, which has show to improve assembly metrics even for complex heterozygotic eukaryotic assemblies [39]. I mapped the reads to the circularized assembly of GAC1 and filtered the mapped reads with GERENUQ [40] to retain only reads that mapped end-to-end with the genome. Filtering of the reads was to ensure that the reads that were likely to cause the assembly to fail or be contaminated (e.g. cross-mapping reads from other species or chimeric reads produced by errors in the Nanopore sequencing) were omitted. Assembly with this refined set of reads allowed for the circularized sequences of two additional genomes to be assembled. For the other two samples, there was an apparent strain heterogeneity that prevented the assembly from being high-quality and completed. In these samples, there are potentially two or more strains of *Manganitrophus* existing simultaneously in the samples.

I also applied the conjugative element identification protocol from chapter 2 to the raw metagenomic assembly of sample GAC1. One of the circularized elements that were sub-chromosomal in size had a full array of conjugative proteins and was also annotated as belonging to the same taxa as the genome of GAC1. As previously discussed, there needs to be additional evidence to demonstrate that indeed this large, conjugative plasmid does belong to *Candidatus Manganitrophus* sp. GAC1 within this complex community. The initial evidence of the same assigned taxonomy and a similar mapping coverage across the sequences are good initial indicators, but not sufficient with how many bacteria grow on these biofilms. I aligned the assembled conjugative plasmid to the published genome of *Candidatus Manganitrophus noduliformans*, which was sequenced as an isolate, and found that this novel plasmid aligns to multiple genomic fragments (Figure A.1). These sequences were not previously identified as being a plasmid or conjugative system in the original publication. In addition, this plasmid aligned to the published genome of *Candidatus Manganitrophus morganii SA1*. By utilizing third-generation sequencing and applying the methods developed in chapter 2, I resolved the full genomic context of these species by showing that there is a conserved conjugative mega-plasmid in multiple members of the genus. This plasmid carries genes for the transport and binding of copper and cadmium, among other cations, which is are important for its ability to survive in an environment that is heavily contaminated with heavy metals.

In the two poorer quality assemblies in chapter 4, strain resolution was possible, but difficult. Manual separation of the substrains in sample GAC5 allowed for the assembly of one of the strains in the sample. I aligned and filtered the reads to both substrains and managed to produce one assembly that had high completion, low redundancy, and no strain heterogeneity. Recently, it has been shown that full strain resolution is possible with extraordinarily deep sequencing of a single sample, but would be too expensive for regular applications [10]. Deeper sequencing of the communities may have allowed for resolution of the strains in samples GAC3 and GAC5 and provided more confidence in the genome sequences that were assembled. Because the genomes were only sequenced using a Nanopore MinION, the overall sequence quality will be lower than a genome polished using highly-accurate Illumina sequencing reads.

Annotations of these novel strains revealed that metabolic pathways for polyaromatic hydrocarbons are conserved throughout the genus. Additionally, I found that there is a greater diversity of genes related to heavy metal transport and binding in the GAC strains than there are in *Candidatus Manganitrophus morganii* or *Candidatus Manganitrophus noduliformans*. These unique genes are a reflection of the adaptation of these strains to their environment, which would be more highly contaminated with heavy metals than the other two species. Transport and binding of these heavy metals may help enable their survival in such a toxic environment. There is also evidence that ammonia oxidation in the phylum *Nitrospirota* is inhibited by environmental heavy metal contamination [41], so the transportation and binding machinery in these strains may be adaptations to help rescue similar metabolic capabilities, such as hydrocarbon metabolism.

6.5 Secondary assembly to complete additional closed genomes

Knowing the value of closed, circular bacterial genomes for metagenomic analyses, particularly for mobile genetic elements, I developed a framework that improved the yield of circular genetic elements and applied it to complex bacterial communities. I took inspiration from the successful reference-based assemblies in chapter 4 and Jorg [42], a similar framework that is designed for short-read data, and applied it to long-read sequencing data generated by the Oxford Nanopore Technologies MinION platform. By performing a secondary assembly on metagenomic bins generated by MetaBAT2 [43], there was more than a doubling of the number of contigs 1Mb or greater that were circular. For contigs under 1Mb, the increase in circular elements was much more modest with only about a 15% increase, which is to be expected given the data shown in Chapters 2 and 3. Because the secondary assembly was performed on bins generated by MetaBAT2, the exclusion of conjugative systems and other mobile genetic elements [12] from would reduce the probability of such elements being successfully circularized in a secondary assembly. Performing secondary assemblies on the unbinned fraction of contigs may have yielded additional circular contigs, but I omitted it from the analysis due to high computational burden that secondary assemblies require. The primary measure of circularization of contigs is the reporting of the structure of the assembly by Flye. However, I also generated read tiling paths for each contig in the assembly (reads over 5kb that overlap by at least 500 bp and a read that maps to the start and end of the fasta entry) to check for contig contiguity and circularity. By this metric, there were an additional 45 contigs under 1Mb per sample in the unbinned that were considered circular, but not flagged as circular by Flye, which indeed suggests that secondarily assembling those contigs would have been fruitful.

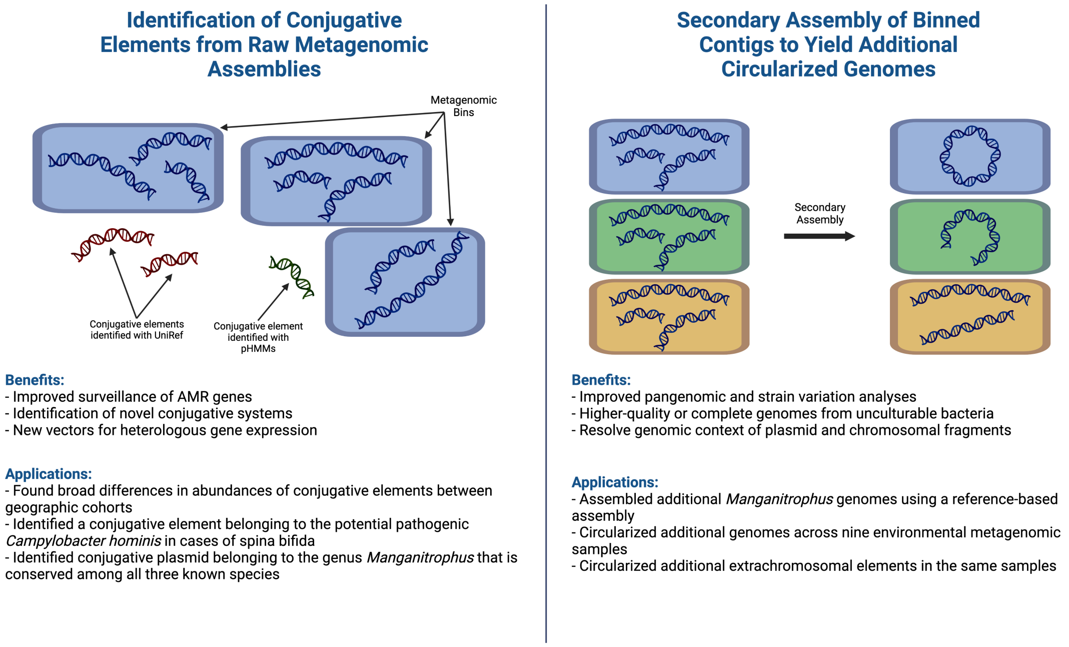
As a framework to generate additional complete bacterial genetic sequences from a complex environment, secondarily assembling metagenomic bins was successful. However, there is still much work to be done to optimize the performance of such a framework. As mentioned, the computational resources and wall clock time required to run is prohibitive for regular use. For each sample, the full secondary assembly run time was roughly one week while running with 30 threads on a server with 128GB of memory. The primary assembly portion of the pipeline only takes a matter of hours, so the vast majority of the run time is dedicated entirely to the secondary assembly process. If the research questions are not benefited by having complete chromosomal and extra-chromosomal sequences, then the secondary assembly framework is an computationally expensive and unnecessary process. However, if the algorithm can be tuned to minimize the run time, it may become more reasonable for general metagenomic applications. In addition, quality checks to determine whether the bin prior to reassembly is of higher quality and whether extra round of reassembly beyond one would improve assembly quality would be other avenues to improve the framework. As it stands, the current framework yielded minimal to no improvements on general assembly metrics such as N50, completion, and contamination. Adding in extra checks for quality could make the framework more useful for generally improving assembly quality and not just a tool to obtain more complete genomes.

6.6 The microbiota of an oil refinery wastewater treatment facility

As a proof of principle of the secondary assembly framework, it was applied to nine samples that were derived from the wastewater treatment facility of an oil refinery wastewater treatment facility to characterize their composition. Eight samples were from the biofilms that form on granular activated charcoal filters and the final sample was of the flocculent of a collection basin, which were collected over the time period of over two years. Despite the large number of complete chomosomal sequences, there was a lack of specific taxonomies assigned within the communities. Many taxonomies pointed to largely uncharacterized clades mostly within the phylum *Proteobacteria*. Some of the taxa that were assigned within the communities were *Rugosibacter, Ferruginibacter, Terricaulis, Accumulibacter, Manganitrophus, Macondimonas, Methylotenera, Sphingobium* and *Hyphomicrobium*. *Manganitrophus* species within these communities and their potential for oxidation of manganese as an energy source, bioremediation of other heavy metals, and the bioremediation of hydrocarbons was explored in Chapter 4. A number of the other genera listed have been shown in previous literature to metabolize hydrocarbons and sequester and biotransform heavy metals [44, 45, 46, 47, 48, 49]. A current limitation is that the taxonomy of many of the assembled MAGs could not be assigned due to the presumed novelty of the species in the community. Isolation experiments are required to know the true functional capabilities of the individual species and to properly assign a taxonomy. The composition of the communities is relatively stable over time with the majority of species-level clusters being found in more than one samples. As such, the completed genomic sequences assembled in Chapter 5 represent a set of quality scaffolds to align to in future community dynamics and metatranscriptomic studies of the involvement that member species have in the metabolism of hydrocarbons and heavy metals.

6.7 Future directions

This thesis has advanced the capability to comprehensively analyze both long- and short-read metagenomic datasets (Figure 6.2). I have applied the techniques to a human health-focused cohort with mothers who gave birth to infants with spina bifida, a broad comparison of geographically-based cohorts, and to communities of bacteria growing on charcoal filters and in flocculent. However, there are still many unexplored applications of the techniques that I developed and room to improve the usability and sensitivity of the tools.


Metagenomic Bins
Conjugative elements identified with UniRef
Conjugative element identified with pHMMs
Secondary Assembly

Identification of Conjugative Elements from Raw Metagenomic Assemblies

Benefits:
- Improved surveillance of AMR genes
- Identification of novel conjugative systems
- New vectors for heterologous gene expression

Applications:
- Found broad differences in abundances of conjugative elements between geographic cohorts
- Identified a conjugative element belonging to the potential pathogenic *Campylobacter hominis* in cases of spina bifida
- Identified conjugative plasmid belonging to the genus *Manganitrophus* that is conserved among all three known species

Secondary Assembly of Binned Contigs to Yield Additional Circularized Genomes

Benefits:
- Improved pangenomic and strain variation analyses
- Higher-quality or complete genomes from unculturable bacteria
- Resolve genomic context of plasmid and chromosomal fragments

Applications:
- Assembled additional *Manganitrophus* genomes using a reference-based assembly
- Circularized additional genomes across nine environmental metagenomic samples
- Circularized additional extrachromosomal elements in the same samples

Figure 6.2: Overview of methods developed within this thesis. On the left is a diagrammatic depiction of the method for identifying conjugative elements from a raw metagenomic assembly that highlights their exclusion from metagenomic bins. On the right is a schematic of the improved contiguity and circularization of assemblies following a secondary assembly of binned contigs. Created with BioRender

The scale of human metagenomic studies exploring the connections between human health and disease is vast, but as I previously mentioned, conjugative systems are a substantial blind spot. The creation of the UHGG has proven to be a worthwhile endeavour [18, 19], and the same principles that led to its creation would be sufficient rationale to build a similar database for conjugative systems. Prior to such an effort, there would first have to be some improvements to the computational methods proposed in this thesis. On the Data Repository For Human Gut Microbiota, there are over 26,000 human gut metagenomic samples listed as being publicly available to analyze. On this scale, improvements to the speed of the process to annotate the raw assembly need to made to conserve computational resources. Curating a database solely of conjugative proteins from the UniRef90 database should reduce the run time of annotating the assemblies, while maintaining the sensitivity advantage that it demonstrated over the pHMM approach (Figure 2.3). Alternatively, the already computationally efficient pHMM approach could have its databases expanded to the point where the gap in sensitivity would be outweighed the reduced burden on computational resources. With the computational limitations resolved, the construction of a comprehensive conjugative element database could serve many purposes. For instance, it could be used for applications like differential abundance analyses for human health conditions or serve as reference for the synthesis or conjugative systems.

With the number of samples publicly available for re-analysis, a case could be made for a large-scale and high-quality meta-analysis of the association of conjugative elements and human health outcomes. The data presented in chapter 3 hint at these associations being relevant in cohorts beyond mothers who give birth to infants with spina bifida. Conjugative elements and their cargo could, in some cases, be relevant to the pathogenesis of human diseases, and a meta-analysis would have the ability to identify such patterns from the data. Care would need to be taken to ensure that the functions carried as cargo are functionally relevant to the research question as differences in abundances of conjugative elements could oftentimes simply be a reflection of the abundances of the bacteria that carry them. As mentioned in the case for spina bifida, conjugative elements can serve as a vector to carry CRISPR systems that can modulate the microbiome by selectively killing bacteria [36, 50] or to express other genes of interest [51]. With these applications of conjugative elements, the conjugative elements do not need to be directly involved with pathogenesis to be of interest in a meta-analysis. Conjugative systems that belong to the species mediating the negative health outcomes would also be of interest, as with the *Campylobacter hominis*. Synthesis of these conjugative elements *de novo* and loaded with a CRISPR system would be selective antibiotic that could 'fix' the microbiota with little overall disruption.

De novo assembly of conjugative plasmids from metagenomic samples will be difficult with the current data because assembly with short-read data, which the vast majority of samples are, will not yield complete conjugative element sequences. If we hope to *de novo* build additional conjugative plasmids, knowing that it is being built off of the full sequence found in nature will eliminate the possibility of building a plasmid that is missing critical elements needed for its propagation and transfer. Additionally, human gut samples will need to be sequenced and assembled using third-generation sequencing technologies in tandem with optimal bioinformatic workflows to yield a greater number of complete genomes and plasmids for this application. Furthermore, there may be instances where the conjugative element has been rendered non-function by recombination, so additional checks may be required to ensure the core complement of genes needed for conjugation are intact and not just present. Improvements to tools for detecting oriT sequences should also be pursued as an oriT is necessary for conjugation, but at present these tools fall short at detecting these difficult-to-predict sequences.

Additionally, complete genomes and plasmids will allow for a greater understanding of community dynamics and composition. As with the methods for the identification of conjugative elements, there are improvements to be made with the secondary assembly workflow outlined in chapter 5. A run time of over a week per sample is far too long to be used in regular analyses, so progress would need to be made with the efficiency the algorithm and how it works with all the programs it interfaces with. Reducing the computation time needed for mapping should be possible by removing reads that strongly align to a bin from future iterations of mapping in the loop. However, the bulk of the computation time is spent on assembly, which would require much engineering on multithreading of the assemblies to optimize the wall-clock time for this step. Additionally, with the methylation data available, which can associate plasmid with chromosomal sequences [22], with some third-generation sequencing data, studies should now be able to look at the dynamics of conjugative elements in different environments. In theory, the researcher will be able to monitor their spread throughout the community and how they vary in abundance when compared to their cognate genomes. Modelling of the dynamics of conjugative systems in the human gut could be monitored in real time using a chemostat to mimic the human gut environment [52]. Phenomena, such as the spread of antimicrobial resistance or CRISPR-carrying conjugative plasmids, could be modeled and help inform *in vivo* experiments or public health monitoring efforts.

For the applications of the methods that I explored in this thesis, there is still much to be discovered. For spina bifida, confirmational studies with larger sample sizes will need to confirm the roles of *Campylobacter* and *Peptoniphilus* in the induction of intestinal inflammation and reduced nutrient uptake in the expectant mothers. As well, transcriptomic, proteomic, and metabolomic experiments should also be performed in parallel to identify which genes, proteins, and metabolic products of the gut microbiome are significantly affecting the pathogenesis of spina bifida.

Isolation and study of the *Manganitrophus* species from chapter 4 would be beneficial to better understand the organism. While I speculate that it retains the capacity to oxidize manganese, this would need to be tested following isolation. Using a manganese-laden media to isolate could allow for a symbiont to be identified as well as demonstrated in the original *Manganitrophus noduliformans* publications [37, 38]. Further exploration of the heavy metal resistance and polyaromatic hydrocarbon metabolism is also warranted for this organism because it was not explored in the other two publications of the genus.

Metatranscriptomics and metaproteomics of carbon-degrading communities growing on the carbon filters and in the flocculent would allow the ability to understand how the communities interface with their environment. Metagenomic analyses only can provide information on what the communities might be capable of metabolically, but not with certainty. Observing which genes are being highly expressed when concentrations of hydrocarbons or heavy metals are higher, or in conditions of ‘upset’ within the reactor system, would highlight what pathways and operons are most important for the survival and proliferation of bacteria in these complex communities. Quantification of the detoxification capabilities of these bacteria in culture could also help clarify the potential of these communities, or certain important members of the communities, in bioremediation efforts.

Heavy metal and hydrocarbon contamination is a global concern, and a better understanding of these bacteria that can thrive within these environments could prove to be a missing piece of the puzzle for solving the issues we face. To help supplement this, additional communities at the wastewater treatment facilities, such as the flocculent which was sequenced once in chapter 5, should be further studied. The filtrate that passes through the GAC filters is a combination of the water that comes with the oil in the pipe, water that is in contact with the oil during refining and water that is the result of standard waste practices at the refinery. As a result, the source of the oil may have an effect on the composition of the communities and deserves further investigation.

The broad goal of the thesis was to continue the step-wise improvement of bioinformatic analyses for microbiome studies. By outlining and applying protocols for the identification of conjugative elements and the circularization of additional chromosomal and extra-chromosomal elements from third-generation metagenomic sequencing experiments, I believe I achieved this goal. Continued application of these protocols, and progress in the improvement of them, will further optimize the ability to analyze the data generated from microbiome studies.

6.8 References

Bibliography

[1] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, April 1953.

[2] F. H. Crick. On protein synthesis. *Symposia of the Society for Experimental Biology*, 12:138–163, 1958.

[3] W. Fiers, R. Contreras, F. Duerinck, G. Haegeman, D. Iserentant, J. Merregaert, W. Min Jou, F. Molemans, A. Raeymaekers, A. Van den Berghe, G. Volckaert, and M. Ysebaert. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature*, 260(5551):500–507, April 1976.

[4] F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, C. A. Fiddes, C. A. Hutchison, P. M. Slocombe, and M. Smith. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, 265(5596):687–695, February 1977.

[5] International Human Genome Sequencing Consortium, Whitehead Institute for Biomedical Research, Center for Genome Research:, Eric S. Lander, Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, Roel Funke, Diane Gage, Katrina Harris, Andrew Heaford, John Howland, Lisa Kann, Jessica Lehoczky, Rosie LeVine, Paul McEwan, Kevin McKernan, James Meldrim, Jill P. Mesirov, Cher Miranda, William Morris, Jerome Naylor, Christina Raymond, Mark Rosetti, Ralph Santos, Andrew Sheridan, Carrie Sougnez, Nicole Stange-Thomann, Nikola Stojanovic, Aravind Subramanian, Dudley Wyman, The Sanger Centre:, Jane Rogers, John Sulston, Rachael Ainscough, Stephan Beck, David Bentley, John Burton, Christopher Clee, Nigel Carter, Alan Coulson, Rebecca Deadman, Panos Deloukas, Andrew Dunham, Ian Dunham, Richard Durbin, Lisa French, Darren Grafham, Simon Gregory, Tim Hubbard, Sean Humphray, Adrienne Hunt, Matthew Jones, Christine Lloyd, Amanda McMurray, Lucy Matthews, Simon Mercer, Sarah Milne, James C. Mullikin, Andrew Mungall, Robert Plumb, Mark Ross, Ratna Shownkeen, Sarah Sims, Washington University Genome Sequencing Center, Robert H. Waterston, Richard K. Wilson, LaDeana W. Hillier, John D. McPherson, Marco A. Marra, Elaine R. Mardis, Lucinda A. Fulton, Asif T. Chinwalla, Kymberlie H. Pepin, Warren R. Gish, Stephanie L. Chissoe, Michael C. Wendl, Kim D. Delehaunty, Tracie L. Miner, Andrew Delehaunty, Jason B. Kramer, Lisa L. Cook, Robert S. Fulton, Douglas L. Johnson, Patrick J. Minx, Sandra W. Clifton, US DOE Joint Genome Institute:, Trevor Hawkins, Elbert Branscomb, Paul Predki, Paul Richardson, Sarah Wenning, Tom Slezak, Norman Doggett, Jan-Fang Cheng, Anne Olsen, Susan Lucas, Christopher Elkin, Edward Uberbacher, Marvin Frazier, Baylor College of Medicine Human Genome Sequencing Center:, Richard A. Gibbs, Donna M. Muzny, Steven E. Scherer, John B. Bouck, Erica J. Sodergren, Kim C. Worley, Catherine M. Rives, James H. Gorrell, Michael L.

Metzker, Susan L. Naylor, Raju S. Kucherlapati, David L. Nelson, George M. Weinstock, RIKEN Genomic Sciences Center:, Yoshiyuki Sakaki, Asao Fujiyama, Masahira Hattori, Tetsushi Yada, Atsushi Toyoda, Takehiko Itoh, Chiharu Kawagoe, Hidemi Watanabe, Yasushi Totoki, Todd Taylor, Genoscope and CNRS UMR-8030:, Jean Weissenbach, Roland Heilig, William Saurin, Francois Artiguenave, Philippe Brottier, Thomas Bruls, Eric Pelletier, Catherine Robert, Patrick Wincker, Department of Genome Analysis, Institute of Molecular Biotechnology:, André Rosenthal, Matthias Platzer, Gerald Nyakatura, Stefan Taudien, Andreas Rump, GTC Sequencing Center:, Douglas R. Smith, Lynn Doucette-Stamm, Marc Rubenfield, Keith Weinstock, Hong Mei Lee, JoAnn Dubois, Beijing Genomics Institute/Human Genome Center:, Huanming Yang, Jun Yu, Jian Wang, Guyang Huang, Jun Gu, Multimegabase Sequencing Center, The Institute for Systems Biology:, Leroy Hood, Lee Rowen, Anup Madan, Shizen Qin, Stanford Genome Technology Center:, Ronald W. Davis, Nancy A. Federspiel, A. Pia Abola, Michael J. Proctor, University of Oklahoma's Advanced Center for Genome Technology:, Bruce A. Roe, Feng Chen, Huaqin Pan, Max Planck Institute for Molecular Genetics:, Juliane Ramser, Hans Lehrach, Richard Reinhardt, Cold Spring Harbor Laboratory, Lita Annenberg Hazen Genome Center:, W. Richard McCombie, Melissa de la Bastide, Neilay Dedhia, GBF—German Research Centre for Biotechnology:, Helmut Blöcker, Klaus Hornischer, Gabriele Nordsiek, *Genome Analysis Group (listed in alphabetical order, also includes individuals listed under other headings):, Richa Agarwala, L. Aravind, Jeffrey A. Bailey, Alex Bateman, Serafim Batzoglou, Ewan Birney, Peer Bork, Daniel G. Brown, Christopher B. Burge, Lorenzo Cerutti, Hsiu-Chuan Chen, Deanna Church, Michele Clamp, Richard R. Copley, Tobias Doerks, Sean R. Eddy, Evan E. Eichler, Terrence S. Furey, James Galagan, James G. R. Gilbert, Cyrus Harmon, Yoshihide Hayashizaki, David Haussler, Henning Hermjakob, Karsten Hokamp, Wonhee Jang, L. Steven Johnson, Thomas A. Jones, Simon Kasif, Arek Kaspryzk, Scot Kennedy, W. James Kent, Paul Kitts, Eugene V. Koonin, Ian Korf, David Kulp, Doron Lancet, Todd M. Lowe, Aoife McLysaght, Tarjei Mikkelsen, John V. Moran, Nicola Mulder, Victor J. Pollara, Chris P.

Ponting, Greg Schuler, Jörg Schultz, Guy Slater, Arian F. A. Smit, Elia Stupka, Joseph Szustakowki, Danielle Thierry-Mieg, Jean Thierry-Mieg, Lukas Wagner, John Wallis, Raymond Wheeler, Alan Williams, Yuri I. Wolf, Kenneth H. Wolfe, Shiaw-Pyng Yang, Ru-Fang Yeh, Scientific management: National Human Genome Research Institute, US National Institutes of Health:, Francis Collins, Mark S. Guyer, Jane Peterson, Adam Felsenfeld, Kris A. Wetterstrand, Stanford Human Genome Center:, Richard M. Myers, Jeremy Schmutz, Mark Dickson, Jane Grimwood, David R. Cox, University of Washington Genome Center:, Maynard V. Olson, Rajinder Kaul, Christopher Raymond, Department of Molecular Biology, Keio University School of Medicine:, Nobuyoshi Shimizu, Kazuhiko Kawasaki, Shinsei Minoshima, University of Texas Southwestern Medical Center at Dallas:, Glen A. Evans, Maria Athanasiou, Roger Schultz, Office of Science, US Department of Energy:, Aristides Patrinos, The Wellcome Trust:, and Michael J. Morgan. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001.

[6] J. Craig Venter, Karin Remington, John F. Heidelberg, Aaron L. Halpern, Doug Rusch, Jonathan A. Eisen, Dongying Wu, Ian Paulsen, Karen E. Nelson, William Nelson, Derrick E. Fouts, Samuel Levy, Anthony H. Knap, Michael W. Lomas, Ken Nealson, Owen White, Jeremy Peterson, Jeff Hoffman, Rachel Parsons, Holly Baden-Tillson, Cynthia Pfannkoch, Yu-Hui Rogers, and Hamilton O. Smith. Environmental genome shotgun sequencing of the Sargasso Sea. *Science (New York, N.Y.)*, 304(5667):66–74, April 2004.

[7] Gene W. Tyson, Jarrod Chapman, Philip Hugenholtz, Eric E. Allen, Rachna J. Ram, Paul M. Richardson, Victor V. Solovyev, Edward M. Rubin, Daniel S. Rokhsar, and Jillian F. Banfield. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43, March 2004.

[8] Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, Sergey Aganezov, Savannah J. Hoyt, Mark Diekhans, Glennis A. Logsdon, Michael Alonge, Stylianos E. Antonarakis, Matthew Borchers, Gerard G. Bouffard, Shelise Y. Brooks, Gina V.

Caldas, Haoyu Cheng, Chen-Shan Chin, William Chow, Leonardo G. de Lima, Philip C. Dishuck, Richard Durbin, Tatiana Dvorkina, Ian T. Fiddes, Giulio Formenti, Robert S. Fulton, Arkarachai Fungtammasan, Erik Garrison, Patrick G.S. Grady, Tina A. Graves-Lindsay, Ira M. Hall, Nancy F. Hansen, Gabrielle A. Hartley, Marina Haukness, Kerstin Howe, Michael W. Hunkapiller, Chirag Jain, Miten Jain, Erich D. Jarvis, Peter Kerpedjiev, Melanie Kirsche, Mikhail Kolmogorov, Jonas Korlach, Milinn Kremitzki, Heng Li, Valerie V. Maduro, Tobias Marschall, Ann M. McCartney, Jennifer McDaniel, Danny E. Miller, James C. Mullikin, Eugene W. Myers, Nathan D. Olson, Benedict Paten, Paul Peluso, Pavel A. Pevzner, David Porubsky, Tamara Potapova, Evgeny I. Rogaev, Jeffrey A. Rosenfeld, Steven L. Salzberg, Valerie A. Schneider, Fritz J. Sedlazeck, Kishwar Shafin, Colin J. Shew, Alaina Shumate, Yumi Sims, Arian F. A. Smit, Daniela C. Soto, Ivan Sović, Jessica M. Storer, Aaron Streets, Beth A. Sullivan, Françoise Thibaud-Nissen, James Torrance, Justin Wagner, Brian P. Walenz, Aaron Wenger, Jonathan M. D. Wood, Chunlin Xiao, Stephanie M. Yan, Alice C. Young, Samantha Zarate, Urvashi Surti, Rajiv C. McCoy, Megan Y. Dennis, Ivan A. Alexandrov, Jennifer L. Gerton, Rachel J. O'Neill, Winston Timp, Justin M. Zook, Michael C. Schatz, Evan E. Eichler, Karen H. Miga, and Adam M. Phillippy. The complete sequence of a human genome. preprint, Genomics, May 2021.

[9] Daniel J Giguere, Alexander T Bahcheli, Benjamin R Joris, Julie M Paulssen, Lisa M Gieg, Martin W Flatley, and Gregory B Gloor. Complete and validated genomes from a metagenome. preprint, Bioinformatics, April 2020.

[10] Derek M. Bickhart, Mikhail Kolmogorov, Elizabeth Tseng, Daniel M. Portik, Anton Korobeynikov, Ivan Tolstoganov, Gherman Uritskiy, Ivan Liachko, Shawn T. Sullivan, Sung Bong Shin, Alvah Zorea, Victòria Pascal Andreu, Kevin Panke-Buisse, Marnix H. Medema, Itzhak Mizrahi, Pavel A. Pevzner, and Timothy P. L. Smith. Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities. *Nature Biotechnology*, January 2022.

[11] Eli L. Moss, Dylan G. Maghini, and Ami S. Bhatt. Complete, closed bacterial genomes from

microbiomes using nanopore sequencing. *Nature Biotechnology*, 38(6):701–707, June 2020.

[12] Finlay Maguire, Baofeng Jia, Kristen L. Gray, Wing Yin Venus Lau, Robert G. Beiko, and Fiona S. L. Brinkman. Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic Islands. *Microbial Genomics*, 6(10), October 2020.

[13] Chris Smillie, M. Pilar Garcillán-Barcia, M. Victoria Francia, Eduardo P. C. Rocha, and Fernando de la Cruz. Mobility of plasmids. *Microbiology and molecular biology reviews: MMBR*, 74(3):434–452, September 2010.

[14] Anuradha Ravi, Lorena Valdés-Varela, Miguel Gueimonde, and Knut Rudi. Transmission and persistence of IncF conjugative plasmids in the gut microbiota of full-term infants. *FEMS microbiology ecology*, 94(1), January 2018.

[15] Heidi Gumpert, Jessica Z. Kubicek-Sutherland, Andreas Porse, Nahid Karami, Christian Munck, Marius Linkevicius, Ingegerd Adlerberth, Agnes E. Wold, Dan I. Andersson, and Morten O. A. Sommer. Transfer and Persistence of a Multi-Drug Resistance Plasmid in situ of the Infant Gut Microbiota in the Absence of Antibiotic Treatment. *Frontiers in Microbiology*, 8:1852, 2017.

[16] Scott A. McEwen and Peter J. Collignon. Antimicrobial Resistance: a One Health Perspective. *Microbiology Spectrum*, 6(2), March 2018.

[17] Xiaofang Jiang, Andrew Brantley Hall, Ramnik J. Xavier, and Eric J. Alm. Comprehensive analysis of chromosomal mobile genetic elements in the gut microbiome reveals phylum-level niche-adaptive gene pools. *PloS One*, 14(12):e0223680, 2019.

[18] Alexandre Almeida, Alex L. Mitchell, Miguel Boland, Samuel C. Forster, Gregory B. Gloor, Aleksandra Tarkowska, Trevor D. Lawley, and Robert D. Finn. A new genomic blueprint of the human gut microbiota. *Nature*, 568(7753):499–504, April 2019.

[19] Alexandre Almeida, Stephen Nayfach, Miguel Boland, Francesco Strozzi, Martin Beracochea, Zhou Jason Shi, Katherine S. Pollard, Ekaterina Sakharova, Donovan H. Parks, Philip Hugenholtz, Nicola Segata, Nikos C. Kyrpides, and Robert D. Finn. A unified cat-

alog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology*, 39(1):105–114, January 2021.

[20] Baris E Suzek, Yuqi Wang, Hongzhan Huang, Peter B McGarvey, Cathy H Wu, and UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–32, March 2015.

[21] Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*, 12(1):59–60, January 2015.

[22] Alan Tourancheau, Edward A. Mead, Xue-Song Zhang, and Gang Fang. Discovering multiple types of DNA methylation from bacteria and microbiome using nanopore sequencing. *Nature Methods*, 18(5):491–498, May 2021.

[23] Eman Zakaria Gomaa. Human gut microbiota/microbiome in health and diseases: a review. *Antonie Van Leeuwenhoek*, 113(12):2019–2040, December 2020.

[24] Edoardo Pasolli, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica Armanini, Francesco Beghini, Paolo Manghi, Adrian Tett, Paolo Ghensi, Maria Carmen Collado, Benjamin L Rice, Casey DuLong, Xochitl C Morgan, Christopher D Golden, Christopher Quince, Curtis Huttenhower, and Nicola Segata. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell*, 176(3):649–662.e20, January 2019.

[25] F. A. Bastiaan von Meijenfeldt, Ksenia Arkhipova, Diego D. Cambuy, Felipe H. Coutinho, and Bas E. Dutilh. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biology*, 20(1):217, December 2019.

[26] Andrew J. Copp, N. Scott Adzick, Lyn S. Chitty, Jack M. Fletcher, Grayson N. Holmbeck, and Gary M. Shaw. Spina bifida. *Nature Reviews. Disease Primers*, 1:15007, April 2015.

[27] P. N. Kirke, A. M. Molloy, L. E. Daly, H. Burke, D. G. Weir, and J. M. Scott. Maternal plasma folate and vitamin B12 are independent risk factors for neural tube defects. *The Quarterly Journal of Medicine*, 86(11):703–708, November 1993.

[28] J. G. Ray and H. J. Blom. Vitamin B12 insufficiency and the risk of fetal neural tube defects.

QJM: monthly journal of the Association of Physicians, 96(4):289–295, April 2003.

[29] G. M. Shaw, E. M. Velie, and D. M. Schaffer. Is dietary intake of methionine associated with a reduction in risk for neural tube defect-affected pregnancies? *Teratology*, 56(5):295–299, November 1997.

[30] Gary M. Shaw, Richard H. Finnell, Henk J. Blom, Suzan L. Carmichael, Stein Emil Vollset, Wei Yang, and Per M. Ueland. Choline and risk of neural tube defects in a folate-fortified population. *Epidemiology (Cambridge, Mass.)*, 20(5):714–719, September 2009.

[31] C. J. Schorah, J. Wild, R. Hartley, S. Sheppard, and R. W. Smithells. The effect of periconceptional supplementation on blood vitamin concentrations in women at recurrence risk for neural tube defect. *The British Journal of Nutrition*, 49(2):203–211, March 1983.

[32] E. M. Velie, G. Block, G. M. Shaw, S. J. Samuels, D. M. Schaffer, and M. Kulldorff. Maternal supplemental and dietary zinc intake and the occurrence of neural tube defects in California. *American Journal of Epidemiology*, 150(6):605–616, September 1999.

[33] Ian Rowland, Glenn Gibson, Almut Heinken, Karen Scott, Jonathan Swann, Ines Thiele, and Kieran Tuohy. Gut microbiota functions: metabolism of nutrients and other food components. *European Journal of Nutrition*, 57(1):1–24, February 2018.

[34] Ravinder Nagpal, Tiffany M. Newman, Shaohua Wang, Shalini Jain, James F. Lovato, and Hariom Yadav. Obesity-Linked Gut Microbiome Dysbiosis Associated with Derangements in Gut Permeability and Intestinal Cellular Homeostasis Independent of Diet. *Journal of Diabetes Research*, 2018:3462092, 2018.

[35] G. M. Shaw, K. Todoroff, E. M. Velie, and E. J. Lammer. Maternal illness, including fever and medication use as risk factors for neural tube defects. *Teratology*, 57(1):1–7, January 1998.

[36] Thomas A Hamilton, Gregory M Pellegrino, Jasmine A Therrien, Dalton T Ham, Peter C Bartlett, Bogumil J Karas, Gregory B Gloor, and David R Edgell. Efficient inter-species conjugative transfer of a CRISPR nuclease for targeted bacterial killing. *Nat Commun*, 10(1):4544, October 2019.

[37] Hang Yu and Jared R. Leadbetter. Bacterial chemolithoautotrophy via manganese oxidation. *Nature*, 583(7816):453–458, July 2020.

[38] Hang Yu, Grayson L. Chadwick, Usha F. Lingappa, and Jared R. Leadbetter. Comparative genomics on cultivated and uncultivated, freshwater and marine *Candidatus* Manganitrophaceae species implies their worldwide reach in manganese chemolithoautotrophy. preprint, Microbiology, November 2021.

[39] Heidi E. L. Lischer and Kentaro K. Shimizu. Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC bioinformatics*, 18(1):474, November 2017.

[40] abahcheli. GERENUQ, October 2021. original-date: 2020-10-31T15:13:33Z.

[41] Cheng Yu, Xi Tang, Lu-Shan Li, Xi-Lin Chai, Ruiyang Xiao, Di Wu, Chong-Jian Tang, and Li-Yuan Chai. The long-term effects of hexavalent chromium on anaerobic ammonium oxidation process: Performance inhibition, hexavalent chromium reduction and unexpected nitrite oxidation. *Bioresource Technology*, 283:138–147, July 2019.

[42] Lauren M. Lui, Torben N. Nielsen, and Adam P. Arkin. A method for achieving complete microbial genomes and improving bins from metagenomics data. *PLOS Computational Biology*, 17(5):e1008972, May 2021.

[43] Dongwan D. Kang, Feng Li, Edward Kirton, Ashleigh Thomas, Rob Egan, Hong An, and Zhong Wang. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, 7:e7359, 2019.

[44] Elizabeth M. Corteselli, Michael D. Aitken, and David R. Singleton. Rugosibacter aromaticivorans gen. nov., sp. nov., a bacterium within the family Rhodocyclaceae, isolated from contaminated soil, capable of degrading aromatic compounds. *International Journal of Systematic and Evolutionary Microbiology*, 67(2):311–318, February 2017.

[45] Jiro F. Mori and Robert A. Kanaly. Complete Genome Sequence of Sphingobium barthaii KK22, a High-Molecular-Weight Polycyclic Aromatic Hydrocarbon-Degrading Soil Bacterium. *Microbiology Resource Announcements*, 10(1):e01250–20, January 2021.

[46] Madhumita Roy and Tapan K. Dutta. Draft Whole-Genome Sequence of Sphingobium sp. Strain PNB, a Versatile Polycyclic Aromatic Hydrocarbon-Degrading Bacterium. *Microbiology Resource Announcements*, 10(48):e0092021, December 2021.

[47] Quanfeng Liang and Gareth Lloyd-Jones. Sphingobium scionense sp. nov., an aromatic hydrocarbon-degrading bacterium isolated from contaminated sawmill soil. *International Journal of Systematic and Evolutionary Microbiology*, 60(Pt 2):413–416, February 2010.

[48] Smruthi Karthikeyan, Luis M. Rodriguez-R, Patrick Heritier-Robbins, Minjae Kim, Will A. Overholt, John C. Gaby, Janet K. Hatt, Jim C. Spain, Ramon Rosselló-Móra, Markus Huettel, Joel E. Kostka, and Konstantinos T. Konstantinidis. ”Candidatus Macondimonas diazotrophica”, a novel gammaproteobacterial genus dominating crude-oil-contaminated coastal sediments. *The ISME journal*, 13(8):2129–2134, August 2019.

[49] Ryoich Tanaka, Katsuya Nouzaki, Ronald R. Navarro, Tomohiro Inaba, Tomo Aoyagi, Yuya Sato, Atsushi Ogata, Hiroshi Yanagishita, Tomoyuki Hori, and Hiroshi Habe. Activated sludge microbiome in a membrane bioreactor for treating Ramen noodle-soup wastewater. *The Journal of General and Applied Microbiology*, 66(6):339–343, February 2021.

[50] Kevin Neil, Nancy Allard, Frédéric Grenier, Vincent Burrus, and Sébastien Rodrigue. Highly efficient gene transfer in the mouse gut microbiota is enabled by the Incl2 conjugative plasmid TP114. *Commun Biol*, 3(1):523, September 2020.

[51] Paul O. Sheridan, Jennifer C. Martin, Nigel P. Minton, Harry J. Flint, Paul W. O’Toole, and Karen P. Scott. Heterologous gene expression in the human gut bacteria Eubacterium rectale and Roseburia inulinivorans by means of conjugative plasmids. *Anaerobe*, 59:131–140, October 2019.

[52] Julie A. K. McDonald, Kathleen Schroeter, Susana Fuentes, Ineke Heikamp-Dejong, Cezar M. Khursigara, Willem M. de Vos, and Emma Allen-Vercoe. Evaluation of microbial community reproducibility, stability and composition in a human distal gut chemostat model. *Journal of Microbiological Methods*, 95(2):167–174, November 2013.

Appendix A

Supplemental figures for Chapter 4

A.1 Supplemental figures

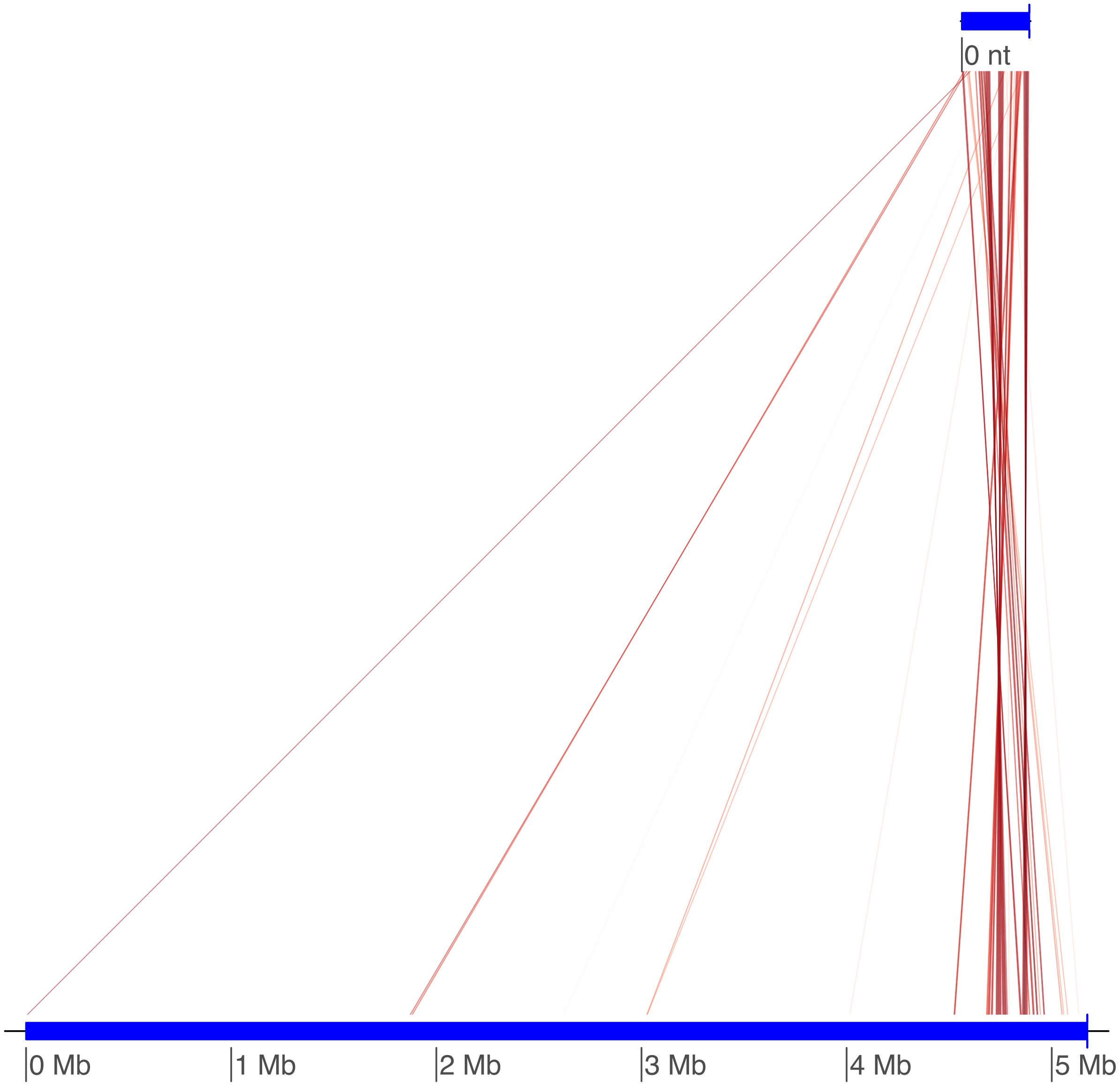


Figure A.1: Alignment graphic produced by FastANI demonstrating the alignment of the conjugative plasmid identified in sample GAC1 (top) to the published *Manganitrophus noduliformans* genome (bottom). The published *Manganitrophus noduliformans* genome is split into 22 fragments and the GAC1 plasmid aligns to four.

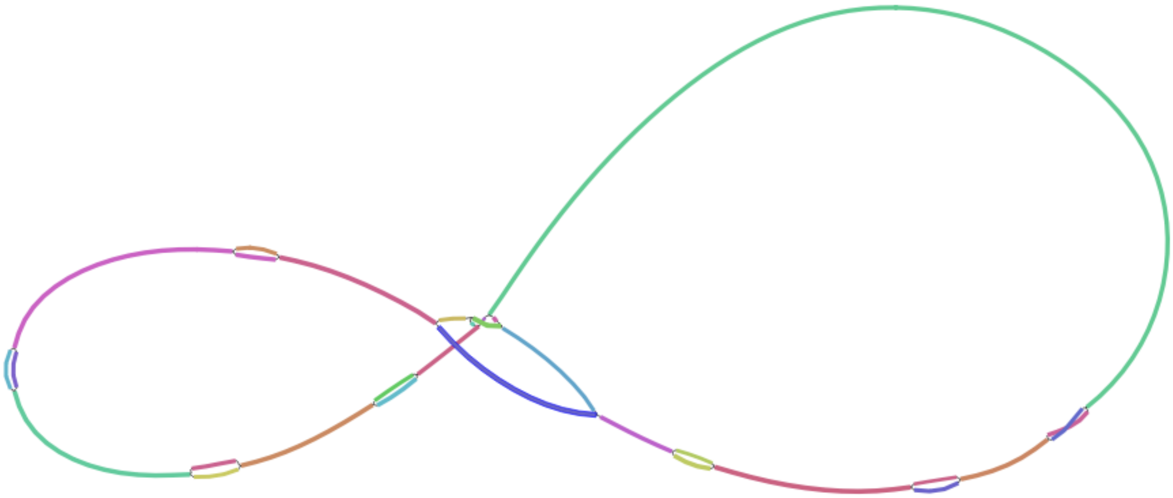


Figure A.2: Assembly graph of GAC3 following reference-based assembly. Colouring of edges is arbitrary.

Appendix B

Supplemental table for Chapter 3

Table B.1: ALDEx2 output for spina bifida microbiome Pfam annotations with effect sizes of an absolute value greater than 0.8. Effect sizes with a negative value have a greater relative abundance of reads aligning to their open reading frames in the mother who gave birth to infant with spina bifida. For positive effect size values, it suggest an relative enrichment in the control samples.

Pfam ID	Pfam name	Benjamini-Hochberg corrected p-value	Effect size
PF05057	Putative serine esterase (DUF676)	0.174765003	-1.155696146
PF05538	Campylobacter major outer membrane protein	0.432899815	-0.949042263
PF16730	DnaG-primase C-terminal, helicase-binding domain	0.397986629	-0.91006937
PF18644	Phage integrase SAM-like domain	0.46082165	-0.893867793
PF14790	Tetrahydrodipicolinate N-succinyltransferase N-terminal	0.480360128	-0.879814532
PF18527	STT3/PglB C-terminal beta-barrel domain	0.421510122	-0.878682462
PF02516	Oligosaccharyl transferase STT3 subunit	0.429277359	-0.868898852
PF11874	Domain of unknown function (DUF3394)	0.442497833	-0.866719687
PF02433	Cytochrome C oxidase, mono-heme subunit/FixO	0.437774994	-0.862846335
PF08376	Nitrate and nitrite sensing	0.444199746	-0.861641416
PF07655	Secretin N-terminal domain	0.445006513	-0.855857978
PF15436	Plasminogen-binding protein pgbA N-terminal	0.500849144	-0.854846005
PF18472	HP1451 C-terminal domain	0.469412759	-0.832981966
PF04366	Las17-binding protein actin regulator	0.450683565	-0.830455458
PF05573	NosL	0.467580738	-0.817239347
PF04028	Domain of unknown function (DUF374)	0.380245105	-0.816951698
PF10108	Predicted 3’-5’ exonuclease related to the exonuclease domain of PolB	0.490982418	-0.808548941
PF13563	2’-5’ RNA ligase superfamily	0.348532907	0.867253049

Curriculum Vitae

Benjamin R. Joris

Department of Biochemistry

University of Western Ontario

London, Ontario N6A 5C1

Education:

PhD Candidate, Biochemistry, 2018-current

Western University, London, Ontario

Master of Science, Biochemistry, 2017-2018

Western University, London, Ontario

Bachelor of Medical Sciences, Honors Specialization in Interdisciplinary Medical Sciences, 2013-2017

Western University, London, Ontario

Publications submitted for peer review:

Joris B.R., Browne T.S., Hamilton T.A., Edgell D.E., Gloor G.B. Identification of type IV conjugative systems that are systematically excluded from metagenomic bins. Microbiome, https://doi.org/10.21203/rs.3.rs-1428512/v1 (Submitted March 9th, 2022).

Peer-reviewed publications:

Joris B.R., Gloor G.B. Unaccounted risk of cardiovascular disease: the role of the microbiome in lipid metabolism. Current opinion in lipidology 30, 2 (2019). doi: 10.1097/MOL.0000000000000582

Non peer-reviewed pre-prints:

Giguere, D.J, Bahcheli, A.T., **Joris, B.R.**, Paulssen, J.M, Gieg, L.M., Flatley, M.W., Gloor G.B. Complete and validated genomes from a metagenome. biorXiv, https://doi.org/10.1101/2020.04.08.032540 (April 9 2020).

Presentations:

London Health Research Day, Schulich School of Medicine, virtual, May 2021 (abstract/poster)

Great Lakes Bioinformatics Conference, International Society for Computational Biology, virtual, May, 2021 (abstract/poster)

DNA Genotek/Diversigen, virtual, November, 2020 (invited lecture)

Exploring Human Host-Microbiome Interactions in Health and Disease, Wellcome Genome Campus, September, 2020 (abstract/poster)

Awards and Honors:

Award	Value	Level	Period Held
Dean's Research Scholarship	$28,000	Institutional	2020/09-2020/09
Ontario Gradudate Scholarship	$15,000	Provincial	2019/09-2020/09
Western Graduate Research Scholarship	$32,333	Institutional	2018/09-2022/08

Experience:

Extra-curricular

University Consulting Group, Consultant, 2021-2022

Biochemistry Graduate Student Association, finance representative, 2020-2021

Society of Graduate Students Finance Committee, committee member, 2020-2021

Teaching Assistantships

Biochemistry 9546R, Western University 2021

Biochemistry 9545Q, Western University 2021

Biochemistry 9546R, Western University 2020

Biochemistry 9545Q, Western University 2020

Biochemistry 9546R, Western University 2019

Biochemistry 9545Q, Western University 2019

Biochemistry 2280A, Western University 2019

Biochemistry 3382A, Western University 2018