Western Graduate&PostdoctoralStudies

Western University
Scholarship@Western

Electronic Thesis and Dissertation Repository

3-17-2022 2:00 PM

Applications of nanopore DNA sequencing for improved genome assembly

Daniel Giguere, *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

© Daniel Giguere 2022

Follow this and additional works at: https://ir.lib.uwo.ca/etd

Part of the Genomics Commons

Recommended Citation

Giguere, Daniel, "Applications of nanopore DNA sequencing for improved genome assembly" (2022). *Electronic Thesis and Dissertation Repository*. 8437.
https://ir.lib.uwo.ca/etd/8437

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlswadmin@uwo.ca.

Abstract

An organism's genome is the ultimate determinant of its functional potential. Understanding genomes is therefore essential to understand function, and a foundational knowledge of a genome is required transfer functions to and from microorganisms of interest. Sequencing DNA using nanopores is a recent advance that resolves technological limitations of previous technologies, enabling an improved understanding of genomes. For this thesis, I improved our understanding of microbial genomes by developing computational approaches to analyze long read sequencing data, setting the foundation for future synthetic biology work.

Long sequencing reads have enabled routine assembly of complete bacterial genomes by directly sequencing DNA extracted from bacterial communities. I showed that visualizing sequencing depth after filtering read alignments using a 95% query coverage cutoff (i.e., the entire read aligns to the genome) enabled the detection of mis-assemblies. I also showed it can be applied to detect recoverable alternate haplotypes containing important functional elements. Furthermore, I used this approach to demonstrate that a circular genome for a novel species of *Saccharibacteria*, enriched from a heavy-metal polluted Northern Albertan tailings pond, contains a recently acquired genomic island. I also determined this genomic island encodes heavy metal-resistance genes, suggesting that horizontal gene transfer from its host may be possible under selective pressure in *Saccharibacteria*.

Another track of my thesis focused on applying nanopore sequencing on a marine diatom, *Phaeodactylum tricornutum*, which has significant interest for synthetic biology applications like producing low-cost glycosylated proteins. This species does not have a complete genome assembly, despite a draft sequence being available since 2008. To determine the full structure of the genome, I used ultra-long sequencing reads to build a telomere-to-telomere genome assembly. I also developed a novel, assembly-free approach to determine the number of chromosomes from eukaryotes directly from nanopore sequencing reads as an orthogonal method to validate the assembly, which I term long-read karyocounting.

These studies provide complete genome assemblies for both novel bacterial species and a marine diatom who's genome structure had yet to be resolved. The approaches in this thesis also demonstrate that there is more information encoded in long read sequencing data than just the sum of assembled sequence.

Keywords: nanopore sequencing, genome assembly, network analysis, *Saccharibacteria*, *Phaeodactylum tricornutum*, synthetic biology

Lay summary

The code for life is written in every living organism's DNA as a unique combination of 4 chemical letters. This combination, called the DNA sequence, determines what the living being is capable of. Technology to characterize the sequence of DNA has improved dramatically since 2014 with the invention of "nanopore" DNA sequencing, where DNA is pulled through a tiny pore for characterization. The main improvement is that the full size of a piece of DNA can be characterized. For my thesis, I improved our understanding of DNA sequences for bacteria and algae by developing new ways to analyze nanopore data, setting the foundation for future research with these organisms.

Nanopore sequencing is improving how complete a DNA sequence can be. For example, while the first human DNA sequence was published in 2001, it was not actually completed until 2021. This new technology comes with new analysis challenges. I developed a filtering and visualization method using the sequences to find analysis errors. I also showed that this same technique can be used to uncover alternate versions of the DNA sequence when more than one exists. Furthermore, I used these visuals to show that a recently discovered bacterium from the Canadian oil sands contained a region of DNA that can move itself from one bacteria to another. This region contained a DNA sequence that is known to pump toxic metals out of its cell, suggesting the bacterium may be capable of acquiring new DNA regions to survive.

A separate track of my thesis focused on better understanding an algae with significant commercial interest because it can be used to make low-cost proteins like the SARS-CoV-2 proteins, required for rapid COVID-19 testing kits. Although a DNA sequence for this algae was published in 2008, it was not complete. In this thesis, I created the first complete DNA sequence for this algae. I also developed a separate analysis method to determine how many unique genome pieces (i.e., chromosomes) exist.

Overall, this thesis provides more complete DNA sequences for several new bacteria, and completes the DNA sequence for a commercially-valuable algae. The analysis methods I developed show that there is more information encoded in the DNA sequence than just the combination of the 4 different letters.

Co-Authorship Statement

For the work presented in Chapter 2, 3, 4, and 5, Daniel Giguere performed the research with exceptions noted below. Daniel Giguere conceived of and designed the experiments and analyzed the data, with input from Gregory Gloor. Daniel Giguere wrote all manuscripts with input from Gregory Gloor.

Chapter 2 - Gregory Gloor helped conceive of the filtering method and assisted in data interpretation. Alec Bahcheli helped implement the filtering algorithm as a python package.

Chapter 3 - Gregory Gloor helped with data analysis.

Chapter 4 - Gregory Gloor, David Edgell, Bogumil Karas, and Martin Flatley helped conceive the project. Samuel Slattery helped design of the DNA extraction protocol, and performed the DNA extraction. Rushali Pateli helped with analysis of methylation data. Tyler Browne helped with genome annotation. Gregory Gloor, and Bogumil Karas helped with data analysis.

Chapter 5 - Gregory Gloor, Lisa Gieg, and Julie Paulssen helped concieve the project. Julie Paulssen and Lisa Gieg maintained and provided the algal-bacterial cultures.

Dedication

This thesis is dedicated to all of my friends, mentors, and family who have helped me get to where I am today.

You miss 100% of the shots you don't take.

- Wayne Gretzky

Acknowledgments

First and foremost, thank you Dr. Greg Gloor for your patient support, the endless opportunities you enabled, and most importantly, your mentorship. I am grateful that you took a chance on agreeing to supervise me, and for the many more chances you took supporting my ideas. In addition to scientific research, you also taught me the importance of soft skills - leadership, integrity, and collaboration. The time working in your lab has been transformative, and for that I will be forever thankful.

I would also like to thank my thesis advisors, Dr. David Edgell and Dr. Art Poon for challenging me to continually improve. Thank you Dr. Edgell for allowing me to perform my lab work in your lab. While not officially an advisor, I would also like to thank Dr. Martin Flatley for playing as many roles as he did throughout my thesis, including opening my eyes to nanopore sequencing.

To all the members of the Edgell and Karas labs, thank you for your insightful discussions, and allowing me to share lab space with you. I would especially like to thank Thomas Hamilton for the uncountable and insightful discussions while "borrowing" your bench space. I am also very thankful for all the friends who agreed to stay on campus even longer to play sports. It's often said great friendships are made during graduate school, but I believe even stronger ones are made while losing at sports together.

I owe a lot of thanks to my lab-mates, past and present. Thank you Dr. Jean Macklaim for for getting me started on this journey and patiently answering my questions whenever you heard "Uhm, Jean?". Ben Joris - thank you for the unique discussions on what zero may or may not mean, the microbiology of poop, and your willingness to continuously collaborate. Thank you to all the students I had the opportunity to learn from - you have taught me more than you could ever know.

I would like to to thank my family for their continuous support throughout this journey, and for putting me on the path to get to where I am today. Finally, I would like to thank Heather, for your unwavering support. I am grateful that you made the difficult times better, and the good times great.

Contents

Certificate of Examination **ii**

Abstract **iii**

Lay summary **v**

Co-Authorship Statement **vi**

Dedication **vii**

Acknowledgments **viii**

List of Figures **xiv**

List of Tables **xv**

List of Appendices **xvi**

List of Abbreviations, Symbols, and Nomenclature **xvii**

1 General introduction **1**

1.1 DNA is the ultimate determinant for biological function 1

1.2 The development of DNA sequencing 2

1.3 Techniques for high-throughput DNA sequencing 2

1.3.1 Sequencing-by-synthesis 3

1.3.2 Amplicon sequencing 4

1.3.3 Nanopore sequencing 6
Library preparation 7
Nanopore sequencing enables observation of native DNA 8
Nanopore sequencing enables real time target enrichment 10
Quality of Nanopore sequencing 10
1.4 Methods for high-molecular weight DNA extraction 11
1.5 Genome and metagenome assembly 13
1.5.1 Overlap layout consensus for Sanger sequencing 13
1.5.2 New algorithms and data structures for high-throughput Illumina sequencing 13
1.5.3 New algorithms and data structures for error-prone long read sequencing . . 15
1.5.4 Polishing long-read assemblies 16
1.5.5 Algorithms and tools for metagenome assembly 17
1.5.6 Genome validation and quality control 19
1.5.7 Metagenome sequencing revised the tree of life 20
1.6 Scope and objectives of this thesis 21
1.7 References 23

2 Filtering long reads detects deletions in genome assemblies 32
2.1 Introduction 32
2.2 Methods 34
2.3 Results 35
2.4 Discussion 39
2.5 References 42

3 Long read karyocounting: an assembly-free method to estimate the number of chromosomes in eukaryotic genomes 45
3.1 Introduction 45
3.2 Methods 46

3.2.1 Data 46
3.2.2 Workflow 47
3.3 Results 48
3.3.1 Each component of aligned reads represents a single telomere 48
3.3.2 Chromosome estimate from number of clusters 50
3.3.3 Validation using assembled genome 50
3.4 Discussion 51
3.4.1 Consistency with known biology and genome assemblies 51
3.4.2 Applications 52
3.4.3 Limitations 52
3.5 Conclusions 53
3.6 References 53

4 Telomere-to-telomere genome assembly of *Phaeodactylum tricornutum* 57
4.1 Introduction 57
4.2 Methods 58
4.2.1 Growth 58
4.2.2 DNA extraction 59
4.2.3 Sequencing 60
4.2.4 Assembly 60
4.2.5 Telomere identification 60
4.2.6 Assembly 61
4.2.7 Polishing 62
4.2.8 Methylation 62
4.3 Results 62
4.3.1 Workflow 62
4.3.2 Tiling path of overlapping reads verify contiguity 64
4.3.3 Telomere-to-telomere assembly comprises previous scaffolds 64

4.3.4 Estimating number of chromosomes using ultra-long reads 65
4.3.5 Assembly quality 67
4.3.6 Filtered long-read coverage for Chromosome 19 is inconsistent with diploid state 67
4.3.7 5mC methylation and transposable elements 68
4.4 Discussion 70
4.5 Conclusions 71
4.6 Funding 71
4.7 References 72

5 Complete genome sequence of new species from the family *Saccharimonadaceae*, *UBA6175* sp., reveals acquisition of a genomic island 77
5.1 Introduction 77
5.2 Methods 79
5.2.1 DNA extraction 79
5.2.2 DNA sequencing 80
5.2.3 Genome assembly 81
5.2.4 Genome annotation and taxonomic prediction 82
5.3 Results 82
5.3.1 Genome quality 82
5.3.2 Phylogenomic analysis reveals *UBA6175* sp. belongs to Clade G1 85
5.3.3 *UBA6175* sp. has reduced functional capability 85
5.3.4 Region R1 may have been recently acquired 89
5.4 Discussion 89
5.5 Conclusions 91
5.6 References 91

6 General discussion 97

6.1 On the definition of a “complete” genome 97
6.2 Long reads enable deeper biological inference from genome assemblies 99
6.3 Fully understanding genomes is required for synthetic biology 101
6.4 Future work and final thoughts 102
6.5 References 103

A Supplemental information for Chapter 2 **107**

A.1 Supplemental figures 107

B Supplemental information for Chapter 4 **110**

B.1 Supplemental figures 110
B.2 Supplemental tables 113

Curriculum vitae **124**

List of Figures

1.1 Nanopore sequencing. 6
1.2 Two major methods for library preparation. 9

2.1 Example of a missed deletions. 36
2.2 Filtering Nanopore reads reveals a haplotype. 38
2.3 Filtering Nanopore reads for a fully assembled community identifies mis-assemblies and haplotypes. 40

3.1 Quality characteristics of ultra-long sequencing run generated by NanoPlot. 48
3.2 Summary of network graph analysis. 49

4.1 Workflow for telomere-to-telomere genome assembly. 63
4.2 Sequencing coverage and comparison to previous assembly. 66
4.3 Summary of genomic features for chromosome 3. 69

5.1 Complete genome of *UBA6175* sp. 84
5.2 Phylogenetic analysis of currently available Saccharibacteria (NCBI taxonomy) genomes and metagenomically-assembled genomes. 86
5.3 Metabolic model for *UBA6175* sp. 87

A.1 Performance of Gerenuq. 108
A.2 Long read coverage of an alternate haplotype. 109

B.1 Summary of genomic features for chromosome 19. 111
B.2 Histogram of telomere clusters. 112

List of Tables

1.1 Correspondence of q-score, base call accuracy and error rate. 11

3.1 Estimate of the number of chromosomes using the components. 51
3.2 Chromosome assignment of 92 clusters of telomere-containing long reads. 56

5.1 Summary results for metagenomically-assembled circular genomes. 81
5.2 Taxonomic predictions of assembled genomes by the Genome Taxonomy Database tool-kit. 83
5.3 Region R1 in the *UBA6175* sp. genome contains predicted proteins that are from Actinobacteriota. 88

B.1 Overlapping read path for long reads aligning to all chromosomes. 113
B.2 Quality value estimates. 123

List of Appendices

Appendix A: Supplemental information for Chapter 2 107
Appendix B: Supplemental information for Chapter 4 110

List of Abbreviations, Symbols, and Nomenclature

5mC	5-methyl cytosine
A	adenine
bp	base pair
BUSCO	Benchmarking Universal Single-Copy Orthologs
C	cytosine
CCAP	culture collection of algae and protozoa
contig	contiguous DNA sequence
CPR	Candidate Phyla Radiation
CRISPR	clustered regularly interspaced short palindromic repeats
CTAB	cetyl trimethyl ammonium bromide
DNA	deoxyribonucleic acid
dsDNA	double-stranded deoxyribonucleic acid
haplotype (prokaryote)	a block of sequence variants corresponding to a strain, sub-strain, or alternate version of a genome in a species
haplotype (eukaryote)	block(s) of sequence corresponding to sequence variants inherited together on a chromosome
g	standard gravity
G	guanine
GC	guanine and cytosine
kb	kilo base pair
KEGG	kyoto encyclopedia of genes and genomes
k-mer	substring of length *k* contained within a biological sequence
LTR	long terminal repeat
MAG	metagenomically assembled genome

MAWG	metagenomically assembled whole genome
metagenome	DNA sequencing of all DNA from a community or environment
nm	nanometer
N50	smallest sequencing read length of which 50% of total bases are found
PCR	polymerase chain reaction
Q	phred quality score
QV	quality value
RNA	ribonucleic acid
rRNA	ribosomal RNA
SNP	single nucleotide polymorphism
ssDNA	single-stranded deoxyribonucleic acid
SDS	sodium dodecyl sulfate
SRE	Short Read Eliminator
T	thymine
TCA	tricarboxylic acid
TE	tris EDTA
Tris-HCl	tris(hydroxymethyl)aminomethane hydrochloride
tRNA	transfer RNA
unitig	a high confidence contig

Chapter 1

General introduction

1.1 DNA is the ultimate determinant for biological function

A living organism's nucleic acid sequences are the ultimate determinant for its functional capability. This was discovered by determining that the causative material behind bacterial transformation is deoxyribonucleic acids (DNA) (1). Shortly after this discovery, the three-dimensional structure of DNA was solved (2). More recent technological advances enabled scientists to determine the sequence for the four canonical DNA bases that make up an organisms genetic code (adenine - A, cytosine - C, guanine - G, and thymine - T) (3), enabling researchers to decode the genetic sequence of many organisms. It is through decoding the sequence of DNA that it has been possible to understand how proteins, the molecular entities responsible for biological function, are ultimately encoded in biological systems (4, 5).

The overall goal of my thesis is to improve our understanding of the genetic code for several organisms by generating the best genome assemblies possible using a new DNA sequencing technology called nanopore sequencing, and developing new methods to ensure they are complete. To do this, I developed novel computational approaches using the noisy (i.e., low signal to noise ratio - therefore less accurate) ultra-long sequencing reads generated by the Oxford Nanopore Technologies MinION DNA sequencing platform. In particular, I: improved the detection of mis-assemblies and alternate haplotypes of metagenomically-assembled whole genomes in Chapter 2; developed an assembly-free method to estimate the number of eukaryotic nuclear chromosomes in Chapter 3; I applied these new approaches to resolve the first telomere-to-telomere genome assembly for the marine diatom *Phaeodactylum tricornutum* in Chapter 4; and I characterized a novel species of a recently proposed phylum, Patescibacteria, described in Chapter 5. The advances presented in this thesis further our understanding into these micro-organisms by better understanding their genetic code and genome structure, setting the foundation needed to enable future synthetic biology work in these organisms.

1.2 The development of DNA sequencing

High quality DNA sequencing began with Frederick Sanger, who designed a method to determine the sequence of nucleotides from DNA strands by replicating DNA *in vitro* using the enzyme DNA polymerase in combination with ^{32}P-labelled chain-terminating di-deoxynucleoside triphosphates, followed by visualization by electrophoresis (3). Improvements to this method, such as the incorporation of chain-terminating fluorescent dyes instead of radio-labelled nucleotides (6), and automation using capillary gel electrophoresis (7) enabled DNA sequencing to be scaled to thousands of bases per day. However, the human genome contains approximately 3.3 billion bases (8), and even large bacterial genomes can contain up to 15 million bases (9). High-throughput technologies were needed to generate enough data to assemble complete genomes inexpensively, so that we could understand the functional potential of any organisms.

1.3 Techniques for high-throughput DNA sequencing

While DNA sequencing began with automating Sanger sequencing for the human genome project (8, 10), the main driver to lowering the cost of DNA sequencing has been higher-throughput with several different platforms, including pyrosequencing (11) commercialized by 454 Life Science, and sequencing-by-synthesis (SBS) commercialized by Solexa, which was later acquired by Illumina (and will be referred to as Illumina sequencing) (12, 13). Other sequencing technologies were developed, such as ABI SOLiD sequencing (14), Ion Torrent sequencing (15), but these technologies quickly became obsolete as the massive throughput and lower cost of SBS technologies prevailed. While SBS techniques are well known, specific aspects of the methods cause limitations during genome assembly that are addressed in this thesis, and are therefore reviewed here.

1.3.1 Sequencing-by-synthesis

To sequence DNA on the Illumina sequencing platform, purified DNA is randomly sheared, and oligonucleotide sequencing adapters are ligated onto both ends of the double stranded DNA (dsDNA). The dsDNA is then denatured to single stranded DNA (ssDNA), and each of the adapters bind to complimentary fragments attached to the flow cells, resulting in a closed loop. Once bound, bridge amplification occurs, where unlabelled nucleotides and enzymes are added to the solution to build dsDNA bridges. Denaturation occurs again, resulting in a single-stranded template, and this process repeats until several million dense clusters of dsDNA are generated in each lane of the flow cell. The first sequencing cycle begins by adding four fluorescently-labelled terminators and primers, as well as other reagents like DNA polymerase. After fluorescent excitation with a laser, the emitted fluorescence is captured, which is later converted to one of four canonical bases. The fluorescent label is removed, leaving a 3'-hydroxy for the next nucleotide to be incorporated. This cycle repeats until the end of the read (12).

This method has two major limitations:

1. The sequenced read length produced is typically around 150 bp (ranging from 75-250 bp).
2. The signal observed is fluorescence produced from polymerase-chain reaction (PCR) amplified DNA. The signal observed is therefore not a direct observation of the native DNA.

These limitations ultimately mean that downstream analyses of this technology platform has great difficulty with assembling repetitive regions (discussed in Genome and Metagenome assembly), and information about the native DNA (e.g., DNA modifications such as 5-methylcytosine methylation) is lost before sequencing. In addition, PCR amplification is the source of a well-documented limitation of Illumina sequencing such as GC-bias during the PCR amplification steps (16–18). Poor amplification of high-GC regions can be caused by the presence of secondary structures and high melting points acting as permanent termination sites (19). This bias is especially problematic when sequencing DNA derived from bacterial communities (metagenomes) since the nucleic acid composition from various organisms may span a wide range of GC content. For example, the human pathogen *Clostridium difficile* has a GC content of approximately 28% (20), while a common gut commensal bacteria such as *Bifidobacterium longum* has a much higher GC content at approximately 60% (21). Some extremophiles such as *Deinococcus radiodurans* even have GC content as high as 67% (22). The large difference in GC content often causes amplification bias during the PCR step, which causes parts of an inidividual genome to not be amplified, resulting in partial data loss (23). This results in fragmented contig- or scaffold-level genome assemblies. Ultimately, this results in genome assemblies not being fully completed, limiting the biological insights that can be obtained.

1.3.2 Amplicon sequencing

Targeted amplicon sequencing of the 16S ribosomal RNA (rRNA) gene is an approach that is often used for culture-independent taxonomic surveys of bacterial communities because the 16S rRNA gene is the basis of molecular taxonomy (24). While it is known that there is some sequence variation in regions of this gene, it is hypothesized that relationships between all bacteria can be measured using it (25). Therefore, 16S rRNA targeted gene sequencing has been used to investigate bacterial isolates and communities for taxonomic classification studies (26). Once high-throughput DNA sequencing became available, “universal primers”, were designed to amplify the hyper-variable regions of the 16S rRNA gene from all bacteria in a community to obtain amplicons from all bacteria in a sample, which could later be analyzed and clustered by species (27, 28). While this technology successfully captured the 16S rRNA gene sequences of many bacteria that contained typical 16S rRNA gene sequences, there were still bacteria that were not captured by generating amplicons. It has been shown that there is a significant proportion of bacteria that contain 16S rRNA gene sequences that are too divergent to be amplified by these commonly used universal primers, and that these bacteria comprise more than 15% of all known bacterial species (29). This was determined after technological advances enabled whole genome assembly of these bacteria from metagenomes. Amplification would not occur for many of 16S rRNA genes because they contained unusual self-splicing introns, while other sequences were so divergent that commonly used universal primers (e.g., 515F and 806R (28)), would have failed to amplify. Therefore, the search space for targeted amplicon sequencing experiments is strongly biased towards bacteria with less divergent 16S rRNA sequences (which were often already well-studied bacteria). Targeted amplicon sequencing is thus not as suitable for discovery based experiments as is commonly believed, especially for bacterial communities where many bacteria may still be unknown, such as atypical or novel environmental samples.

16S rRNA gene sequencing has also been used to infer functional capabilities (30). A taxonomic lineage is commonly assigned by performing phylogenetic analysis of hyper-variable 16S rRNA gene regions, and function is inferred from the genes available in reference genomes of bacteria from that lineage. However, there is a critical limitation to this inference, namely that it does not consider mobile genetic elements or the pangenome. Mobile genetic elements, such as conjugative plasmids, provide a mechanism for horizontal gene transfer in the human microbiome (31, 32) and other naturally occurring environments (33). These conjugative plasmids can transfer operons with unique functions that would not appear in a publicly available reference genome for other species in the same genus. In addition, other elements like genomic islands have been shown to transfer genetic elements within bacterial communities (34). Therefore, functional inferences from taxonomic assignment by 16S rRNA gene sequencing may not capture the functional capability of a bacterium from a novel community.

These limitations are important to consider when designing an experiment since they will limit the amount of biological and functional information that can be obtained. Nanopore sequencing is a DNA sequencing technology that overcomes these limitations by providing an amplicon-free platform for whole metagenome sequencing.

1.3.3 Nanopore sequencing

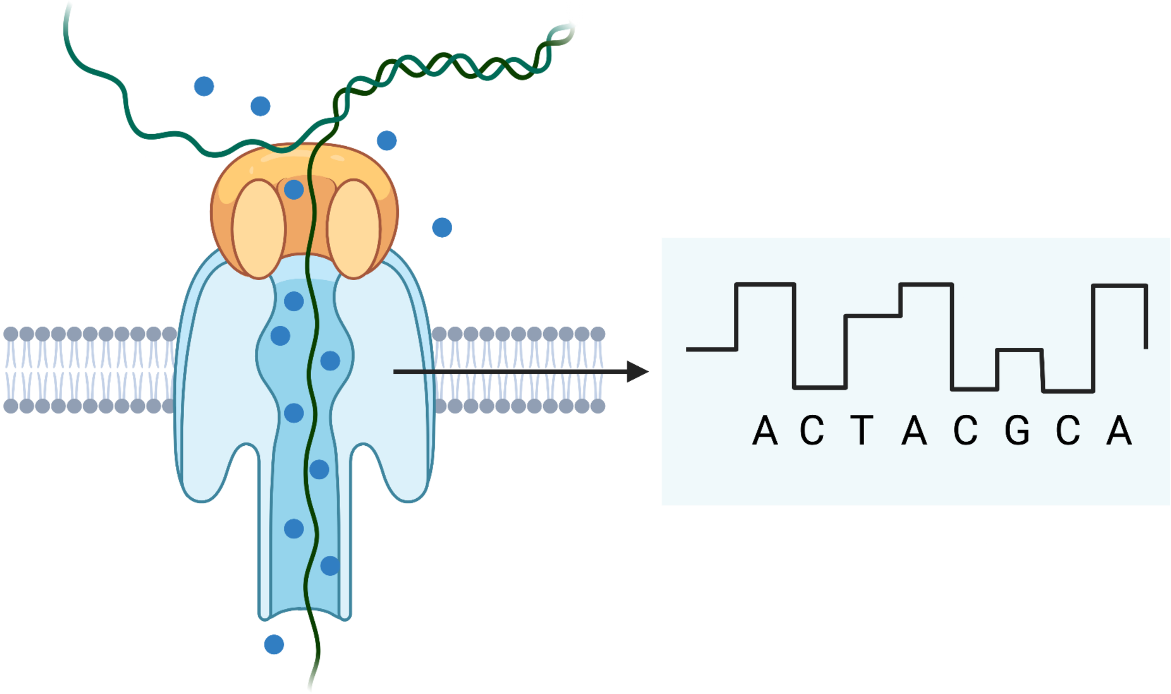


Figure 1.1: **Nanopore sequencing.** DNA is unwound by a motor protein, passed through a biological pore along with an ionic current, and the electrical signal is observed as bases pass through. This raw signal, recorded as "squiggles", is later converted to DNA sequence using machine learning algorithms. Figure made with Biorender.com

Nanopore sequencing is a rapidly-evolving technology that has improved significantly since the beginning of my thesis project. Improvements and important considerations are therefore reviewed below, with a highlight on changes that have occurred since my thesis project began.

This new technology for DNA sequencing uses a biological nanopore composed of an α-hemolysin engineered for DNA sequencing applications, which is embedded in a membrane (35). The first demonstration of nanopore sequencing worked by passing both an ionic current and a single-strand of DNA through a 2.6 nm diameter biological pore embedded on a lipid bilayer membrane (36). Voltage is applied across the membrane, and the current is observed for each pore as the DNA polymer passes through, partially blocking the channel. As negatively charged DNA molecules are passed through the pore, each base will produce a characteristic change in current that can be converted back into sequence algorithmically (37).

Oxford Nanopore Technologies commercialized nanopore DNA sequencing, producing publicly accessible DNA sequencers since 2014 (38), however, the error rate was initially extremely high, estimated at approximately 38% (39). The introduction of the R9 pore in 2017 enabled single-pass read accuracy of approximately 85% (40), with further improvements up to 94% single-pass read accuracy with the R9.4 pore (41). While any individual read may have a 5-10% error rate using the R9.4 nanopores, consensus bacterial genome assemblies using these reads can achieve up to 99.99% accuracy (quality score of 40) because errors in basecalling are random, except for stretches of homopolymers (42, 43). Systematic insertions or deletions occur at homopolymers since there is no signal change as a homopolymer larger than 4-5 passes through the nanopore. The number of bases is dependent only on time. As of October 2021, the modal accuracy of raw sequencing reads (i.e., the accuracy of most sequencing reads) using the R9.4 pore was 98.3% according to Oxford Nanopore Technologies. An R10.3 nanopore with a dual-pore head (i.e., twice the sensing area) was released to help resolve homopolymers, and using the most recent basecalling algorithms, Oxford Nanopore Technologies advertises that fully completed 99.999% accurate genomes can be obtained. Homopolymers larger than 8-10 bases remain difficult to resolve with high accuracy. This accuracy corresponds with the quality of Illumina genome assemblies, and satisfies the definition of a “complete” genome in terms of consensus accuracy (44).

Library preparation

Two of the main library preparation methods (Figure 1.2) that Oxford Nanopore Technologies provides (ligation based and transposome based) differ significantly from Illumina sequencing. The choice in library preparation method is critical for downstream analysis, since the efficiency of each method for circular and linear DNA is different (45). For the ligation-based chemistry (46), sequencing adapters are ligated directly onto blunt ends of native dsDNA. Generally, this enables the length of sequenced fragments to be the length of the input DNA, however, in practice this is limited by hydrodynamic shear when preparing the sequencing library by pipetting (47). The end result is that there is a practical limit to the fragment length of DNA that can be obtained without taking special precautions. This protocol often results in a maximum read length N50 of 15-50 kilobases, depending on the sample type. However, a limitation to this is that small circular or supercoiled plasmids that are not sheared by pipetting will not have sequencing adapters attached, and will therefore not be sequenced.

The other library preparation method, the transposome-based chemistry, also called the rapid chemistry (48), involves randomly cleaving the DNA fragment and simultaneously attaching the sequencing adapter using a proprietary transposome complex. There are fewer steps involved in the library protocol for this, so it is possible to reduce hydrodynamic shear with this chemistry and produce longer read lengths. With this method, small plasmids can be sequenced since the transposome complex will cleave the plasmid randomly. Recovering small plasmids has been observed to occur more often using the rapid chemistry, than the ligation chemistry (45).

The choice of library preparation method is therefore important when sequencing different types of DNA. In the context of metagenome assembly, while the ligation kit optimizes for throughput and read length, it is biased against circular DNA elements where a free dsDNA end is not available for adapter ligation. Circular bacterial chromosomes are sheared into multiple linear fragments when pipetting, but smaller circular plasmids may not be sheared. The choice of library preparation method therefore may have an impact on the ability to assemble a full genome for an organism, and should be considered accordingly.

Nanopore sequencing enables observation of native DNA

A major difference with nanopore sequencing when compared to previous technologies is that the data obtained is a direct measurement of electrical signal of the native DNA, without amplification steps. There are two major advantages to this:

1. Signal information of modified DNA bases can be captured in addition to the canonical bases

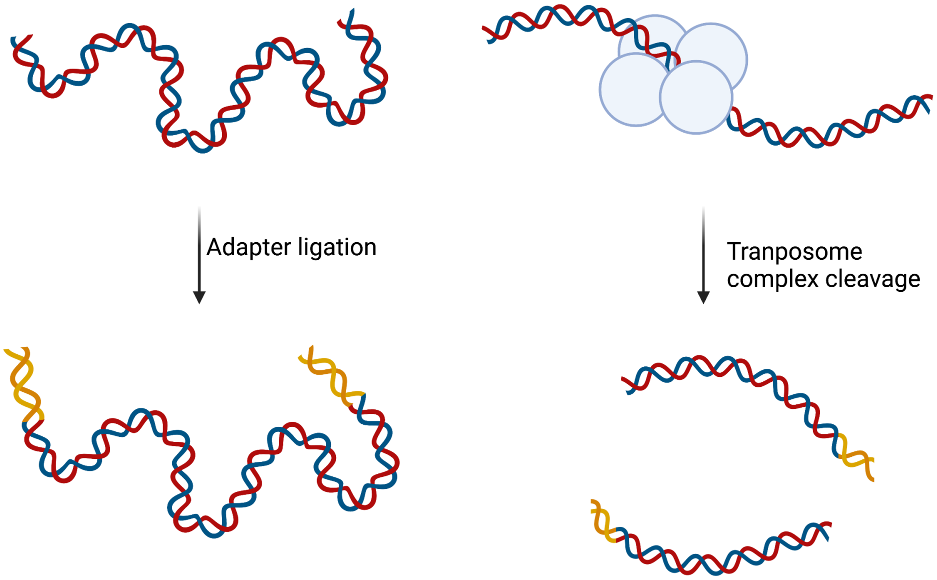


Figure 1.2: **Two major methods for library preparation**. Left, sequencing adapters with motor proteins are ligated using a T4 blunt end ligase after DNA repair. Right, a transposome complex simultaneously cleaves and attaches adapters, resulting in two dsDNA fragments, each with a sequencing adapter attached. Figure made with Biorender.com

since native DNA is sequenced.

2. The read length is theoretically dependent only on the length of the input DNA, rather than the length of an amplicon.

Since an electrical signal of the native DNA is observed, modified bases, such as the 5-methylcytosine and 5-hydroxymethylcytosine found in human DNA can be detected, in addition to and separately from the standard 4 canonical bases (49). This has been recently shown for the *E. coli* methylome (50) and the human methylome (51). More recent software development has enabled the detection of many types of methylation motifs in metagenomic data *de novo* by comparing the native DNA raw signal to whole-genome amplified signal (52) to identify methylated motifs. This information has even been used in real time to enhance decision-making on how aggressively to resect brain tumours during surgery since DNA methylation patterns are strongly correlated with prognosis (53). In addition to improving the contiguity (i.e., the size and number of the overlapping fragments representing the genome) of genome assemblies by providing long reads, nanopore sequencing also enables additional epigenetic information to be obtained in real time.

Nanopore sequencing enables real time target enrichment

Nanopore sequencing has now been used to perform real-time target enrichment by aligning bases against a reference in real time. The optimal translocation speed is currently approximately 400-450 bases per second (46). A 50 kb read would therefore take about two minutes to completely pass through the pore. If the first few hundred bases are analyzed and it is determined the fragment being sequenced is not a target, the voltage can be reversed to eject the strand of DNA from the pore in real time (54, 55). This can save significant sequencing capacity for on-target sequences only, and has already been able to enrich targeted human genome sequences to over 30X coverage (55).

Quality of Nanopore sequencing

The trade-off with nanopore sequencing has typically been the ability to obtain longer reads at the expense of read accuracy. Initially, the read accuracy was extremely poor, with an alignment accuracy less than 10% in 2014 (56), which represents a Phred score quality value (q-score) of less than 1. A q-score (Q) is logarithmically related to the probably of the base call error probabilities (P) (57). A q-score is defined as:

$$Q = -10 \times log10(P)$$

A table of Q-score, basecall accuracy and the corresponding error-rate (which usually ranges from a Q value of 1-50) is shown in Table 1.1.

Quality value	Base call accuracy	Error rate
1	20.56%	8 in 10
10	90%	1 in 10
20	99%	1 in 100
30	99.9%	1 in 1000
40	99.99%	1 in 10 000
50	99.999%	1 in 100 000

Table 1.1: **Correspondence of q-score, base call accuracy and error rate**. Illumina reads are typically Q30, while Oxford Nanopore reads are often around Q10, and with the latest chemistry advances, around Q20.

At the start of my thesis projects in 2018, modal Q-scores generated from nanopore sequencing were around 9, but are now around 15. The most recent chemistry enables modal Q-scores of 20 or higher (58). While quality used to be a trade-off for nanopore sequencing, both long and high quality reads can now be routinely obtained.

1.4 Methods for high-molecular weight DNA extraction

The read length achieved in nanopore sequencing is often limited by the input DNA fragment length. It is therefore essential to optimize DNA extraction to maintain the integrity of high-molecular weight DNA (59). Commercially available DNA extraction kits have been optimized for ease-of-use rather than maintaining high-molecular weight DNA since relatively short fragments (less than 1 kb) are needed for Illumina and Sanger sequencing.

Gentle cell lysis can often be achieved enzymatically in the presence of a detergent such as sodium dodecyl sulfate (SDS) and high salt for gram negative bacteria, even from tough environmental or soil samples (60). Cell lysis with proteinase K in the presence of detergent is often effective for gram negative bacteria since they do not contain a complex peptidoglycan layer (61), however, cell wall disruption is often difficult for gram-positive bacteria. Mechanical lysis using “bead-beating” is often employed, and this is typically effective for both gram positive and gram negative bacteria (62), but at the cost of high-molecular weight DNA, since the beads will mechanically shear the DNA. Depending on the composition of the cell wall for various microorganisms, it may be more difficult to effectively lyse certain cell types, especially in soil, leading to biases when considering the relative abundance of each organism (63). Other organisms, like the marine diatom *Phaeodactylum tricornutum*, have a cell wall composed of silica and polysaccharides (64, 65) that are not easily lysed by commonly used enzymes such as lysozyme. To extract high-molecular weight DNA in this diatom, the current best method is to mechanically grind cells in liquid nitrogen to expose the nucleus, followed by digestion of proteins in the nuclear envelope with proteinase K (66).

To capture genome assemblies from complex environmental metagenome samples where the exact composition is unknown, it is therefore critical to minimize DNA extraction bias towards gram negative bacteria by the addition of one or more lytic enzymes such as achromopeptidase (67), chitinase (68), lyticase (69), lysostaphin (70), lysozyme (71), and mutanolysin (72), which can increase yield from difficult-to-lyse bacteria. Although each DNA extraction technique will typically result in some bias (73), addition of several lytic enzymes can reduce the bias and improve the overall DNA yield obtained (74).

Furthermore, liquid handling also needs to be considered. During the extraction protocol, it is important to minimize hydrodynamic shearing, such as eliminating vortexing, mixing tubes slowly by inversion, and using wide-bore pipette tips when transferring DNA (75).

When DNA is partially sheared after DNA extraction, it is also possible to remove short fragments by selectively precipitating larger DNA fragment (76). Using a combination of high-salt and a buffer containing polyvinylpyrrolidone-360K or polyethylene glycol 8000 (77, 78), short fragments of DNA can be removed, which can help increase the average read length.

To obtain the best possible DNA sequencing data for nanopore sequencing, it is important to optimize DNA extraction protocols to maintain fragment length. This can be done by carefully considering cell lysis efficiency and biases when designing the protocol, and to ensure hydrodynamic shearing caused by liquid transfer and mixing is minimized.

1.5 Genome and metagenome assembly

Since DNA sequencing reads are less than the length of the genome being investigated (with the exception of very small viral genomes and plasmids), re-building the complete genome sequence from sub-sequences is required. Algorithms developed for reconstructing genomes are highly dependent on the sequencing technology used. The advantages and disadvantages of each algorithm and its associated technology is reviewed below to highlight where areas of improvement remain for genome and metagenome assemblies.

1.5.1 Overlap layout consensus for Sanger sequencing

Overlap layout consensus was one of the first approaches developed to rebuild contiguous sequence from sequencing reads (79, 80). This involves looking for sequence overlaps between each read, and stitching together overlapping reads to generate a contiguous DNA sequence (contig). This approach was commonly used for Sanger sequencing because of the extremely high quality reads and low sequencing coverage from this technology, however, it became computationally expensive as sequencing coverage increased and sequence length decreased. In addition, the relatively short read length of Sanger reads (less than 1 kb) made it impossible to resolve repetitive regions that are larger than the length of the read itself, such as a duplicated 16S rRNA gene.

1.5.2 New algorithms and data structures for high-throughput Illumina sequencing

Illumina sequencing created a new algorithmic problem - the reads were very short (initially 30 bases, now up to 250 bases long), and there was a very large amount of data to efficiently handle in a single dataset (millions to billions of individual reads per experiment). Overlap consensus layout assembly was no longer computationally tractable (81).

An alternate data structure, the De Bruijn graph, was introduced to reduce computational complexity (82). Instead of using each read as a vertex, with edges between vertices representing overlaps between reads, the De Bruijn Graph structure breaks down all reads in a dataset into a series of k-mers. Each k-mer is instead used as a vertex to create a Eulerian cycle (i.e., a trail of vertices that starts and ends at the same vertex) between vertices that can reconstruct the genome (83). The size of the assembly graph for this new data structure was therefore dependent on the genome size (i.e., number of unique k-mers in the dataset) instead of the number of reads, enabling much greater computational efficiency for re-building genomes where there is a deep sequencing depth. A genome assembler that adopted this algorithm was Velvet (84), which was intended for small genome sizes (bacterial or fungal). Additional algorithmic advances enabled full genome assembly for human-sized genomes with assemblers like SOAPDenovo (81).

However, there are four major assumptions for using De Bruijn graphs with high-throughput sequencing platforms noted previously (83), that are not necessarily true for high-throughput sequencing.

1. **It's possible to generate all k-mers in a genome**. This is not possible with the Illumina platform because of the polymerase-chain reaction steps - any region with extreme GC content will cause k-mers to be under represented.

2. **All k-mers are error free (i.e., the sequencing instrument is error free)**. The accuracy rate of Illumina is intended to achieve Q30, which corresponds to an error rate of 1 in 1000. Since millions of reads will be produced, reads with errors occur due to random chance.

3. **Each k-mer appears at most once in a genome**. Genomes may have gene duplicates, which would cause k-mers to appear more than once (e.g., duplicated 16S rRNA gene sequences, transposons, etc).

4. **The genome consists of a single circular chromosome**. This is often not the case for bacteria, since they may contain additional circular elements like plasmids.

All k-mers in a given genome may not be generated during Illumina sequencing due to GC bias, resulting in missing k-mers in the dataset. This is one reason why Illumina genome assemblies often result in contig or scaffold level assemblies. On the other hand, when k-mers are present more than once in a genome, it becomes impossible to determine the correct assembly graph, and this is another major reason why many Illumina-only bacterial genome assemblies remain as contigs. There can be duplicated k-mers that make it impossible to determine a unique solution to the assembly graph. Additionally, extrachromosomsal elements, such as plasmids may interfere with the set of reads obtained. For example, a high-copy plasmid may generate a large proportion of the sequencing reads, reducing the sequencing coverage of the genome being investigated.

Ultimately, the major limitation to high-throughput short read sequencing remains the read length, which often leads to broken assemblies due to repetitive regions. These assumptions can be somewhat compensated for algorithmically by genome assemblers, but for short-read and high-throughput genome assemblies, the k-mer size is the limit of the repeat size that can be resolved.

1.5.3 New algorithms and data structures for error-prone long read sequencing

Algorithms for long-read sequencing have been designed with new heuristics because many of the underlying assumptions are incompatible with the new data type, that is, noisy very-long reads. New challenges arise from error-prone long read sequencing generated by the Oxford Nanopore MinION platform.

1. **Nanopore sequencing has a relatively high error rate**. The error rate for basecalling of nanopore sequenced DNA is much higher than previous sequencing platforms, with the typical modal read accuracy typically achieving Q10-Q15 average read quality (at the time of writing). The same sequencing data basecalled in 2017 and re-basecalled in 2021 using updated models shows a large improvement in read accuracy, and quality can vary depending on the basecalling model used.

2. **All k-mers can be sequenced, but homopolymer bases remain an issue**. Since the Oxford Nanopore platform sequences native DNA, there is minimal to no technical bias introduced

during the library preparation protocol caused by nucleotide frequency. However, since there is no change of signal as long homopolymer stretches pass through the pore, predicting the number of bases at homopolymer regions is dependent solely on time. When homopolymers are approximately 5 bases or longer, the accuracy of the basecalled regions is lower.

A computationally efficient implementation for long-read assembly of bacterial genomes based on overlap-consensus layout based assembly is available with minimap2 and miniasm (85), which perform the overlap and layout steps, respectively. Another algorithm that instead uses repeat-graphs for efficient genome assembly is available with Flye (86). Interestingly, a recent review on state-of-the-art prokaryotic genome assembly tools found that while several assemblers typically produce excellent results in specific circumstances (87), no single assembler was the best in all categories tested.

1.5.4 Polishing long-read assemblies

Genome assembly for long-read sequencing data typically creates a noisy draft assembly first, and then error-correction is performed in a process called 'polishing' (88, 89). The major assumption is that obtaining high sequencing coverage produces a much higher quality consensus sequence because the basecalling errors are random. This is often the case, however, systematic errors do exist in the case of homopolymer bases. In addition, because the native DNA is sequenced, any modifications to the DNA (such as methylation) may affect the quality of the basecall since the signal will differ from the trained model based on the canonical base structure. This can be an issue when sequencing organisms with unique DNA modifications that differ from what the basecalling models were trained with.

Several algorithms have been developed for polishing a draft long-read assembly. Pilon (90) can be used to create high quality consensus sequences, however, it requires Illumina reads to be available. The currently recommended polishing approach for nanopore-only sequencing combines one round of Racon (91) with one round of Medaka (provided by Oxford Nanopore Technologies). Other approaches to polishing a final genome assembly include taking multiple assemblies and determining the consensus with Trycycler (43). For eukaryotic genomes, haplotype aware polishing methods have recently been published (92).

Technological advances to long read sequencing technology have improved our ability to generate more complete genomes. While DNA sequencing for the human genome begun in 1990 with the human genome project, it was not until 2021 that one human genome was fully sequenced and assembled (93). In 2001, two reports were published where a draft of the human genome was obtained (8, 10), and this was significantly improved in 2003 by completing a significant majority of the euchromatic genome (94). However, it is important to note that due to limitations of short-read sequencing technology, many repeats (centromeres, telomeres, segmental duplications) could not be resolved with reads shorter than the repetitive region itself. The advances in long read sequencing, both accurate reads generated from Pacific Biosciences and ultra-long reads from Oxford Nanopore Technologies enabled the full completion of the genome, including the placement of repetitive regions. In addition, several new algorithmic approaches were developed for polishing to a final quality value above 70 (95). The ability to routinely sequence and fully assemble human genomes in the future will enable personalized medicine to significantly improve the health outcomes for disease.

1.5.5 Algorithms and tools for metagenome assembly

The first report describing the assembly of near-complete genomes from metagenomes was in 2004 (96). Two near-complete genomes were recovered in addition to three partial genomes using shotgun sequencing. Assembling bacterial genomes directly from bacterial communities (i.e., metagenome assembly) presents even more challenges. Advances in sequencing throughput have enabled the capture of high sequencing coverage in many bacterial communities, including from projects like The Human Microbiome Project (97) and the TARA ocean metagenome project (98). However, these communities are extremely complex, often containing hundreds of species with varying nucleotide frequencies and sequencing coverage. In addition, there are genes and other genomic fragments that may be highly conserved in a community, where most or all of the bacteria may contain a highly similar copy of the same gene (e.g., conserved regions of the 16S rRNA gene (98, 99)), causing short-read genome assemblies to break at these regions.

metaSpades (100) is an assembler developed using Spades (101) as a base, with advances to address some of the difference between genomes and metagenomes. More efficient algorithms such as a succinct De Bruijn graph have also been implemented (102). After contiguous sequences are generated, determining which organism they are derived from presents a challenge. To generate a collection of contigs that likely originate from the same organism, several "binning" algorithms have been developed. Concoct was one of the first algorithms proposed that uses nucleotide composition and sequencing coverage to group contigs into "bins" that each represent a conceptual single genome (103). MetaBat2 uses tetranucleotide frequency and other algorithms to bin contigs together (104). Further yet, the DAS tool was developed to combine the output from existing genome binning methods and use the strengths of each algorithm to aggregate bins (105). To visualize these bins, Anvi'o has enabled aggregating many analyses into a single visualization platform (106). While automated binning algorithms are often very effective, Anvi'o enables manual curation of bins to remove spurious artifacts and analyze partial genome assemblies manually. The key is that a bin is a collection of small contigs that are predicted to be derived from the same genome.

For long reads, metagenome assemblers with novel algorithms have been developed specifically for error-prone reads such as with metaFlye (107), Canu (108), and Raven (109). It was shown that a mock community can be sequenced very deeply and all individual bacteria can be fully assembled, directly from metagenome data (110). While mock communities are great for case studies, the complexity does not represent the complexity of a naturally occurring community in the human microbiome, or environmental samples. New studies have demonstrated that complete genomes can be assembled directly from human stool samples (111). With new tools, it is now possible to generate hundreds of complete, circularized genomes from a single sample (112). Further to this, we provided a proof of principle showing that the majority of a community could be assembled and validated (113) in Chapter 5.

1.5.6 Genome validation and quality control

An important question that is difficult to ask on a per-genome basis when tens or hundreds of genomes are generated is “how good is the assembly quality of this genome”? Mis-assemblies do occur, and validation of contiguity and sequence is an important part of the process. While this was likely less of an issue with Sanger sequencing due to extremely high basecalling quality (Q50), more technical errors were introduced in Illumina reads (Q30), and even more are introduced in nanopore reads (Q10). Due to the trade-off between read length and raw read quality, it is important to consider this question since, to the best of our knowledge, genome and metagenome assemblers for nanopore-only assemblies do not provide estimates of the assembly quality for each contig or genome produced. Interpretation of the output is left to the researcher, but in the case of datasets with many genomes, it is often not performed. This is important to consider since it has been shown that many long-read assemblers suffer from inaccurate circularization of the bacterial genome, often leading to missing sequence in the output (87).

Several genome quality tools have been developed, each answering the question of assembly quality in different ways. REAPR ensures paired-end Illumina reads are correctly aligned throughout an assembled bacterial genome (114). QUAST can check genome assembly quality by comparing to a reference genome, and also can provide descriptive statistics such as the number of contigs, their size, the sequencing coverage of each contig for *de novo* assembled genomes (115). MetaQuest can perform this function for metagenomes (116). However, the main function of both QUAST and MetaQUAST is comparing to an already assembled reference, which cannot provide a quality estimate for a unique *de novo* assembled genome. In addition, these tools also assume that the reference being used is correct and accurately represents the assembled genome.

For genome bins generated from metagenomes, it has been proposed to use the expected number of single-copy core genes to estimate how “complete” or “redundant” a genome bin is. CheckM has been specifically developed for prokaryotic genomes (117), and the software tool BUSCO (Benchmarking Universal Single-Copy Orthologs) is available for eukaryotes (118). A limitation of these tools, however, is that if there is a novel genome identified, the results from these tools can falsely suggest a poor assembly. For example, it has been noted that bacterial members of the recently proposed Candidate Phyla Radiation often contain fewer of what is considered the core set of “essential” bacterial genes (29), and this has lead to a proposal to modify the current tree of life (119). These genome bins would therefore often appear to lack several single-copy core genes, resulting in an apparent “poor quality” assembly, even though the assembly may have been good quality. For BUSCO predictions, it is often required to choose the appropriate collection of single-copy core genes based on the taxonomic rank of the species being investigated. Related to this, another approach that estimates the percentage of truncated open-reading frames caused by poor nanopore polishing has been proposed (120).

Merqury is a recently proposed tool that estimates the quality value (QV) of an assembled genome estimating the number of k-mers present in the genome assembly that are not present in a set of mapped Illumina reads (121). This is a step towards estimating the quality of genome assemblies in a reference free manner, although it does require Illumina reads.

For nanopore-generated genomes, especially those assembled from metagenomes, there are few tools available to check the quality of genomes after assembly. One validation method recently proposed has been to ensure that there is at least one read that aligns across the genome, with a minimum alignment length of the average read size (111). However, a single read is not sufficient evidence to ensure there are no mis-assemblies when sequencing depth is often at least 100 fold.

Importantly, none of these quality metrics evaluate whether a genome assembly is fully contiguous or biologically complete for *de novo* assemblies. They report only the characteristics of assembled contigs, without biological inference. There is therefore an important gap, discussed in this thesis, which is novel approaches for ensuring genome assembles for bacterial genomes, metagenomically-assembled genomes, and eukaryotic genomes are complete and contiguous.

1.5.7 Metagenome sequencing revised the tree of life

Thanks to the advances in high-throughput sequencing and analysis algorithms, the tree of life was revised to account for the more than 15% of bacterial sequences recently obtained through genome-resolved metagenomics that diverged from the previous tree. In 2015, 8 complete and 789 draft genomes were reconstructed from publicly available data to infer a new bacterial lineage, Candidate Phyla Radiation (CPR) (29). Many of these bacteria are obligate epibionts, living directly on the cells of other hosts. As a result, this phylum has been better characterized due to improvements to genome-resolved metagenomic sequencing. This work ultimately lead to a proposed revision of the tree of life, to include a new superphylum CPR as a completely separate clade of bacteria (119). Many of the bacteria belonging to this phylum are unculturable, and therefore have not been thoroughly investigated. I assemble and explore one such bacterium in Chapter 5.

1.6 Scope and objectives of this thesis

At the beginning of my thesis, there were very few examples in the literature of completing bacterial genomes directly from metagenomes. There were also relatively few publications using the Oxford Nanopore MinION platform for DNA sequencing. The beginning of my thesis project was a collection of hypothesis-generating projects where I could apply nanopore sequencing to learn the most recent technological advances in the DNA sequencing and genome assembly fields. While nanopore sequencing has been available through early access programs since 2014 from Oxford Nanopore Technologies, a version that could achieve a tolerable accuracy of 90% per read was not made available until October 2016. During the beginning of my PhD thesis work in 2017, I began applying this technology to determine its capabilities, short falls, and to determine areas where data analysis could be improved. At the time, there were few studies using this technology, and few algorithms had been developed to process the data until later into my thesis project.

DNA extraction protocols intended for Illumina sequencing weren't designed to maintain the integrity of the DNA, however for nanopore sequencing, I optimized a protocol for efficient, high-molecular weight DNA extraction from initial exploratory projects. One such project was characterizing the microbiome of an activated charcoal filter at a wastewater treatment facility at an oil refinery. Initially, this environment was especially difficult to obtain high-molecular weight

DNA since the charcoal adsorbs high concentrations of metals, toxic hydrocarbons like naphthenic acids and asphaltenes, and other hydrocarbons that interfered with commercial kits and spin columns. Successfully developing a high-molecular weight extraction protocol for this difficult environment enabled me to apply the techniques to other projects, including metagenomic sequencing of a 1-adamantanecarboxylic acid-degrading community generated in Chapter 2 and 5, and high-molecular weight DNA extraction of *Phaeodactlyum tricornutum* in Chapter 3 and 4, with sequencing read N50s surpassing 25 kb and 35 kb, respectively. While not presented as a separate chapter, optimizing DNA extraction protocols for each sample was necessary to achieve the completed genome assemblies described in this thesis. Obtaining Very long sequencing reads larger than 50 kilobases was essential for successfully completing a telomere-to-telomere genome in Chapter 4, completing the metagenomically-assembled whole genome in Chapter 5, understanding limitations of current quality control methods in Chapter 2, and developing algorithms to estimate the number of eukaryotic chromosomes in Chapter 3.

An important issue in the field of genome assembly right now, both from bacterial metagenome and eukaryotic genome assemblies derived from nanopore data, is being able to estimate the quality of *de novo* assembled genomes in terms of per-base accuracy, contiguity, and structural completeness. Estimating the quality of genome assemblies, and even the number of chromosomes in an organism remains a challenge without an already complete reference genome. In Chapter 2, I showed that mis-assemblies, such as deletions go undetected without visual inspection of each genome assembled from a metagenome, and that this is critically important because large alternative bacterial haplotypes or multiple strains can exist in bacteria with fluid genomes, such as when a mobile genetic elements is inserted into only a subset of the population. In Chapter 3, I show that it is possible to estimate the number of chromosomes using only long-read sequence data for novel eukaryotes that are difficult or impossible to karyotype due to the structure of their cell walls. This methodological advance was instrumental for completing the telomere-to-telomere genome of *Phaeodactylum tricornutum* in Chapter 4, which could not be resolved with similar data as recently as 2021 (122). Finally, I applied all of these principles to describe a novel species of

Saccharimonadaceae in Chapter 5 that contains a novel genomic island, potentially from its host bacterium.

Overall, this thesis presents advances to genome analysis and interpretation using state-of-the-art technology that enables more accurate genome assemblies. These studies have resulted in new approaches for evaluating genome assemblies for both prokaryotes and eukaryotes, and examples of their applications are shown in this thesis. Improving DNA sequencing and assembly methods enables a more complete genomic understanding of newly discovered organisms, and an improved understanding of organisms that have potential industrial uses in the field of synthetic biology.

1.7 References

[1] Avery, O. T.; MacLeod, C. M.; McCarty, M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types : Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *Journal of Experimental Medicine* **1944**, *79*, 137–158.

[2] Watson, J. D.; Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **1953**, *171*, 737–738.

[3] Sanger, F.; Nicklen, S.; Coulson, A. R. DNA Sequencing with Chain-Terminating Inhibitors. *Proceedings of the National Academy of Sciences* **1977**, *74*, 5463–5467.

[4] Nirenberg, M. W.; Matthaei, J. H. The Dependence of Cell-Free Protein Synthesis in E. Coli upon Naturally Occurring or Synthetic Polyribonucleotides. *Proceedings of the National Academy of Sciences* **1961**, *47*, 1588–1602.

[5] Crick, F. H. C.; Barnett, L.; Brenner, S.; Watts-Tobin, R. J. General Nature of the Genetic Code for Proteins. *Nature* **1961**, *192*, 1227–1232.

[6] Prober, J. M.; Trainor, G. L.; Dam, R. J.; Hobbs, F. W.; Robertson, C. W.; Zagursky, R. J.; Cocuzza, A. J.; Jensen, M. A.; Baumeister, K. A System for Rapid DNA Sequencing with Fluorescent Chain-Terminating Dideoxynucleotides. *Science* **1987**, *238*, 336–341.

[7] Swerdlow, H.; Gesteland, R. Capillary Gel Electrophoresis for Rapid, High Resolution DNA Sequencing. *Nucleic Acids Research* **1990**, *18*, 1415–1419.

[8] Venter, J. C. et al. The Sequence of the Human Genome. *Science* **2001**, *291*, 1304–1351.

[9] Han, K.; Li, Z.-f.; Peng, R.; Zhu, L.-p.; Zhou, T.; Wang, L.-g.; Li, S.-g.; Zhang, X.-b.; Hu, W.; Wu, Z.-h.; Qin, N.; Li, Y.-z. Extraordinary Expansion of a *Sorangium Cellulosum* Genome from an Alkaline Milieu. *Scientific Reports* **2013**, *3*, 2101.

[10] Lander, E. S. et al. Initial Sequencing and Analysis of the Human Genome. *Nature* **2001**, *409*, 860–921.

[11] Margulies, M. et al. Genome Sequencing in Microfabricated High-Density Picolitre Reactors. *Nature* **2005**, *437*, 376–380.

[12] Bentley, D. R. et al. Accurate Whole Human Genome Sequencing Using Reversible Terminator Chemistry. *Nature* **2008**, *456*, 53–59.

[13] Mardis, E. R. A Decade's Perspective on DNA Sequencing Technology. *Nature* **2011**, *470*, 198–203.

[14] Pandey, V.; Nutter, R. C.; Prediger, E. *Next Generation Genome Sequencing*; John Wiley & Sons, Ltd, 2008; Chapter 3, pp 29–42.

[15] Rothberg, J. M. et al. An Integrated Semiconductor Device Enabling Non-Optical Genome Sequencing. *Nature* **2011**, *475*, 348–352.

[16] Dohm, J. C.; Lottaz, C.; Borodina, T.; Himmelbauer, H. Substantial Biases in Ultra-Short Read Data Sets from High-Throughput DNA Sequencing. *Nucleic Acids Research* **2008**, *36*.

[17] Chen, Y.-C.; Liu, T.; Yu, C.-H.; Chiang, T.-Y.; Hwang, C.-C. Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly. *PLOS ONE* **2013**, *8*, e62856.

[18] Sato, M. P.; Ogura, Y.; Nakamura, K.; Nishida, R.; Gotoh, Y.; Hayashi, M.; Hisatsune, J.; Sugai, M.; Takehiko, I.; Hayashi, T. Comparison of the Sequencing Bias of Currently Available Library Preparation Kits for Illumina Sequencing of Bacterial Genomes and Metagenomes. *DNA Research* **2019**, *26*, 391–398.

[19] McDowell, D. G.; Burns, N. A.; Parkes, H. C. Localised Sequence Regions Possessing High Melting Temperatures Prevent the Amplification of a DNA Mimic in Competitive PCR. *Nucleic Acids Research* **1998**, *26*, 3340–3347.

[20] Sebaihia, M. et al. The Multidrug-Resistant Human Pathogen Clostridium Difficile Has a Highly Mobile, Mosaic Genome. *Nature Genetics* **2006**, *38*, 779–786.

[21] Wei, Y.-X.; Zhang, Z.-Y.; Liu, C.; Zhu, Y.-Z.; Zhu, Y.-Q.; Zheng, H.; Zhao, G.-P.; Wang, S.; Guo, X.-K. Complete Genome Sequence of Bifidobacterium Longum JDM301. *Journal of Bacteriology* **2010**, *192*, 4076–4077.

[22] White, O. et al. Genome Sequence of the Radioresistant Bacterium *Deinococcus Radiodurans* R1. *Science (New York, N.Y.)* **1999**, *286*, 1571–1577.

[23] Browne, P. D.; Nielsen, T. K.; Kot, W.; Aggerholm, A.; Gilbert, M. T. P.; Puetz, L.; Rasmussen, M.; Zervas, A.; Hansen, L. H. GC Bias Affects Genomic and Metagenomic Reconstructions, Underrepresenting GC-Poor Organisms. *GigaScience* **2020**, *9*.

[24] Woese, C. R. Bacterial Evolution. *Microbiological Reviews* **1987**, *51*, 221–271.

[25] Clarridge, J. E. Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases. *Clinical Microbiology Reviews* **2004**, *17*, 840–862.

[26] Mizrahi-Man, O.; Davenport, E. R.; Gilad, Y. Taxonomic Classification of Bacterial 16S rRNA Genes Using Short Sequencing Reads: Evaluation of Effective Study Designs. *PLoS ONE* **2013**, *8*, e53608.

[27] Hao, X.; Jiang, R.; Chen, T. Clustering 16S rRNA for OTU Prediction: A Method of Unsupervised Bayesian Clustering. *Bioinformatics (Oxford, England)* **2011**, *27*, 611–618.

[28] Caporaso, J. G.; Lauber, C. L.; Walters, W. A.; Berg-Lyons, D.; Huntley, J.; Fierer, N.; Owens, S. M.; Betley, J.; Fraser, L.; Bauer, M.; Gormley, N.; Gilbert, J. A.; Smith, G.; Knight, R. Ultra-High-Throughput Microbial Community Analysis on the Illumina HiSeq and MiSeq Platforms. *The ISME journal* **2012**, *6*, 1621–1624.

[29] Brown, C. T.; Hug, L. A.; Thomas, B. C.; Sharon, I.; Castelle, C. J.; Singh, A.; Wilkins, M. J.; Wrighton, K. C.; Williams, K. H.; Banfield, J. F. Unusual Biology across a Group Comprising More than 15% of Domain Bacteria. *Nature* **2015**, *523*, 208–211.

[30] Langille, M. G. I.; Zaneveld, J.; Caporaso, J. G.; McDonald, D.; Knights, D.; Reyes, J. A.; Clemente, J. C.; Burkepile, D. E.; Vega Thurber, R. L.; Knight, R.; Beiko, R. G.; Huttenhower, C. Predictive Functional Profiling of Microbial Communities Using 16S rRNA Marker Gene Sequences. *Nature biotechnology* **2013**, *31*, 814–821.

[31] Ronda, C.; Chen, S. P.; Cabral, V.; Yaung, S. J.; Wang, H. H. Metagenomic Engineering of the Mammalian Gut Microbiome *in situ*. *Nature Methods* **2019**, *16*, 167–170.

[32] Aviv, G.; Rahav, G.; Gal-Mor, O. Horizontal Transfer of the Salmonella Enterica Serovar Infantis Resistance and Virulence Plasmid pESI to the Gut Microbiota of Warm-Blooded Hosts. *mBio* **2016**, *7*, e01395–16.

[33] Dahlberg, C.; Bergstrom, M.; Hermansson, M. *In Situ* Detection of High Levels of Horizontal Plasmid Transfer in Marine Bacterial Communities. *Applied and Environmental Microbiology* **1998**, *64*, 2670–2675.

[34] Juhas, M.; van der Meer, J. R.; Gaillard, M.; Harding, R. M.; Hood, D. W.; Crook, D. W. Genomic Islands: Tools of Bacterial Horizontal Gene Transfer and Evolution. *FEMS Microbiology Reviews* **2009**, *33*, 376–393.

[35] Maglia, G.; Restrepo, M. R.; Mikhailova, E.; Bayley, H. Enhanced Translocation of Single DNA Molecules through α-Hemolysin Nanopores by Manipulation of Internal Charge. *Proceedings of the National Academy of Sciences* **2008**, *105*, 19720–19725.

[36] Kasianowicz, J. J.; Brandin, E.; Branton, D.; Deamer, D. W. Characterization of Individual Polynucleotide Molecules Using a Membrane Channel. *Proceedings of the National Academy of Sciences* **1996**, *93*, 13770–13773.

[37] Deamer, D. W.; Branton, D. Characterization of Nucleic Acids by Nanopore Analysis. *Accounts of Chemical Research* **2002**, *35*, 817–825.

[38] Ashton, P. M.; Nair, S.; Dallman, T.; Rubino, S.; Rabsch, W.; Mwaigwisya, S.; Wain, J.; O'Grady, J. MinION Nanopore Sequencing Identifies the Position and Structure of a Bacterial Antibiotic Resistance Island. *Nature Biotechnology* **2015**, *33*, 296–300.

[39] Laver, T.; Harrison, J.; O'Neill, P. A.; Moore, K.; Farbos, A.; Paszkiewicz, K.; Studholme, D. J. Assessing the Performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification* **2015**, *3*, 1–8.

[40] Jain, M. et al. MinION Analysis and Reference Consortium: Phase 2 Data Release and Analysis of R9.0 Chemistry. 2017.

[41] Tyler, A. D.; Mataseje, L.; Urfano, C. J.; Schmidt, L.; Antonation, K. S.; Mulvey, M. R.; Corbett, C. R. Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications. *Scientific Reports* **2018**, *8*, 10931.

[42] Wick, R. R.; Judd, L. M.; Holt, K. E. Performance of Neural Network Basecalling Tools for Oxford Nanopore Sequencing. *Genome Biology* **2019**, *20*, 129.

[43] Wick, R. R.; Judd, L. M.; Cerdeira, L. T.; Hawkey, J.; Méric, G.; Vezina, B.; Wyres, K. L.; Holt, K. E. Trycycler: Consensus Long-Read Assemblies for Bacterial Genomes. 2021.

[44] Bowers, R. M. et al. Minimum Information about a Single Amplified Genome (MISAG) and a Metagenome-Assembled Genome (MIMAG) of Bacteria and Archaea. *Nature Biotechnology* **2017**, *35*, 725–731.

[45] Wick, R. R.; Judd, L. M.; Wyres, K. L.; Holt, K. E. . Recovery of Small Plasmid Sequences via Oxford Nanopore Sequencing. *Microbial Genomics 7*, 000631.

[46] Ligation Sequencing Kit. https://store.nanoporetech.com/ligation-sequencing-kit110.html.

[47] Bowman, R. D.; Davidson, N. Hydrodynamic Shear Breakage of DNA. *Biopolymers* **1972**, *11*, 2601–2624.

[48] Rapid Sequencing Kit. https://store.nanoporetech.com/rapid-sequencing-kit.html.

[49] Laszlo, A. H.; Derrington, I. M.; Brinkerhoff, H.; Langford, K. W.; Nova, I. C.; Samson, J. M.; Bartlett, J. J.; Pavlenok, M.; Gundlach, J. H. Detection and Mapping of 5-Methylcytosine and 5-Hydroxymethylcytosine with Nanopore MspA. *Proceedings of the National Academy of Sciences* **2013**, *110*, 18904–18909.

[50] Rand, A. C.; Jain, M.; Eizenga, J. M.; Musselman-Brown, A.; Olsen, H. E.; Akeson, M.; Paten, B. Mapping DNA Methylation with High-Throughput Nanopore Sequencing. *Nature Methods* **2017**, *14*, 411–413.

[51] Simpson, J. T.; Workman, R. E.; Zuzarte, P. C.; David, M.; Dursi, L. J.; Timp, W. Detecting DNA Cytosine Methylation Using Nanopore Sequencing. *Nature Methods* **2017**, *14*, 407–410.

[52] Tourancheau, A.; Mead, E. A.; Zhang, X.-S.; Fang, G. Discovering Multiple Types of DNA Methylation from Bacteria and Microbiome Using Nanopore Sequencing. *Nature Methods* **2021**, *18*, 491–498.

[53] Djirackor, L.; Halldorsson, S.; Niehusmann, P.; Leske, H.; Capper, D.; Kuschel, L. P.; Pahnke, J.; Due-Tønnessen, B. J.; Langmoen, I. A.; Sandberg, C. J.; Euskirchen, P.; Vik-Mo, E. O. Intraoperative DNA Methylation Classification of Brain Tumors Impacts Neurosurgical Strategy. *Neuro-oncology Advances* **2021**, *3*, vdab149.

[54] Payne, A.; Holmes, N.; Clarke, T.; Munro, R.; Debebe, B. J.; Loose, M. Readfish Enables Targeted Nanopore Sequencing of Gigabase-Sized Genomes. *Nature Biotechnology* **2021**, *39*, 442–450.

[55] Kovaka, S.; Fan, Y.; Ni, B.; Timp, W.; Schatz, M. C. Targeted Nanopore Sequencing by Real-Time Mapping of Raw Electrical Signal with UNCALLED. *Nature Biotechnology* **2021**, *39*, 431–441.

[56] Mikheyev, A. S.; Tin, M. M. Y. A First Look at the Oxford Nanopore MinION Sequencer. *Molecular Ecology Resources* **2014**, *14*, 1097–1102.

[57] Ewing, B.; Green, P. Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. *Genome Research* **1998**, *8*, 186–194.

[58] Ligation Sequencing Kit (Q20+). https://store.nanoporetech.com/ligation-sequencing-kit-112.html.

[59] Quick, J.; Loman, N. J. *Nanopore Sequencing: An Introduction*; World Scientific, 2019.

[60] Zhou, J.; Bruns, M. A.; Tiedje, J. M. DNA Recovery from Soils of Diverse Composition. *Applied and Environmental Microbiology* **1996**, *62*, 316–322.

[61] Grimberg, J.; Maguire, S.; Belluscio, L. A Simple Method for the Preparation of Plasmid and Chromosomal E.Coli DNA. *Nucleic Acids Research* **1989**, *17*, 8893.

[62] Videnska, P.; Smerkova, K.; Zwinsova, B.; Popovici, V.; Micenkova, L.; Sedlar, K.; Budinska, E. Stool Sampling and DNA Isolation Kits Affect DNA Quality and Bacterial Composition Following 16S rRNA Gene Sequencing Using MiSeq Illumina Platform. *Scientific Reports* **2019**, *9*, 13837.

[63] Frostegård, Å.; Courtois, S.; Ramisse, V.; Clerc, S.; Bernillon, D.; Le Gall, F.; Jeannin, P.; Nesme, X.; Simonet, P. Quantification of Bias Related to the Extraction of DNA Directly from Soils. *Applied and Environmental Microbiology* **1999**, *65*, 5409–5420.

[64] Hecky, R. E.; Mopper, K.; Kilham, P.; Degens, E. T. The Amino Acid and Sugar Composition of Diatom Cell-Walls. *Marine Biology* **1973**, *19*, 323–331.

[65] Le Costaouëc, T.; Unamunzaga, C.; Mantecon, L.; Helbert, W. New Structural Insights into the Cell-Wall Polysaccharide of the Diatom *Phaeodactylum tricornutum*. *Algal Research* **2017**, *26*, 172–179.

[66] Bowler, C. et al. The Phaeodactylum Genome Reveals the Evolutionary History of Diatom Genomes. *Nature* **2008**, *456*, 239–244.

[67] Ezaki, T.; Suzuki, S. Achromopeptidase for Lysis of Anaerobic Gram-Positive Cocci. *Journal of Clinical Microbiology* **1982**, *16*, 844–846.

[68] Hiramatsu, S.; Ishihara, M.; Fujie, M.; Usami, S.; Yamada, T. Expression of a Chitinase Gene and Lysis of the Host Cell Wall during *Chlorella* Virus CVK2 Infection. *Virology* **1999**, *260*, 308–315.

[69] Scott, J. H.; Schekman, R. Lyticase: Endoglucanase and Protease Activities That Act Together in Yeast Cell Lysis. *Journal of Bacteriology* **1980**, *142*, 414–423.

[70] Bastos, M. d. C. d. F.; Coutinho, B. G.; Coelho, M. L. V. Lysostaphin: A Staphylococcal Bacteriolysin with Potential Clinical Applications. *Pharmaceuticals (Basel, Switzerland)* **2010**, *3*, 1139–1161.

[71] Sharon, N. The Chemical Structure of Lysozyme Substrates and Their Cleavage by the Enzyme. *Proceedings of the Royal Society of London. Series B, Biological Sciences* **1967**, *167*, 402–415.

[72] Yokogawa, K.; Kawata, S.; Nishimura, S.; Ikeda, Y.; Yoshimura, Y. Mutanolysin, Bacteriolytic Agent for Cariogenic Streptococci: Partial Purification and Properties. *Antimicrobial Agents and Chemotherapy* **1974**, *6*, 156–165.

[73] Delmont, T. O.; Robe, P.; Cecillon, S.; Clark, I. M.; Constancias, F.; Simonet, P.; Hirsch, P. R.; Vogel, T. M. Accessing the Soil Metagenome for Studies of Microbial Diversity. *Applied and Environmental Microbiology* **2011**, *77*, 1315–1324.

[74] Tighe, S. et al. Genomic Methods and Microbiological Technologies for Profiling Novel and Extreme Environments for the Extreme Microbiome Project (XMP). *Journal of Biomolecular Techniques : JBT* **2017**, *28*, 31–39.

[75] Oppert, B.; Stoss, S.; Monk, A.; Smith, T. Optimized Extraction of Insect Genomic DNA for Long-Read Sequencing. *Methods and Protocols* **2019**, *2*, 89.

[76] Workman, High Molecular Weight DNA Extraction from Recalcitrant Plant Species for Third Generation Sequencing. https://www.researchsquare.com, 2018.

[77] Lis, J. T.; Schleif, R. Size Fractionation of Double-Stranded DNA by Precipitation with Polyethylene Glycol. *Nucleic Acids Research* **1975**, *2*, 383–389.

[78] Tyson, J. Size Selective Precipitation of DNA Using PEG & Salt V1.

[79] Staden, R. A Strategy of DNA Sequencing Employing Computer Programs. *Nucleic Acids Research* **1979**, *6*, 2601–2610.

[80] Flicek, P.; Birney, E. Sense from Sequence Reads: Methods for Alignment and Assembly. *Nature Methods* **2009**, *6*, S6–S12.

[81] Li, R.; Zhu, H.; Ruan, J.; Qian, W.; Fang, X.; Shi, Z.; Li, Y.; Li, S.; Shan, G.; Kristiansen, K.; Li, S.; Yang, H.; Wang, J.; Wang, J. De Novo Assembly of Human Genomes with Massively Parallel Short Read Sequencing. *Genome Research* **2010**, *20*, 265–272.

[82] Idury, R. M.; Waterman, M. S. A New Algorithm for DNA Sequence Assembly. *Journal of Computational Biology* **1995**, *2*, 291–306.

[83] Compeau, P. E. C.; Pevzner, P. A.; Tesler, G. Why Are de Bruijn Graphs Useful for Genome Assembly? *Nature biotechnology* **2011**, *29*, 987–991.

[84] Zerbino, D. R.; Birney, E. Velvet: Algorithms for de Novo Short Read Assembly Using de Bruijn Graphs. *Genome Research* **2008**, *18*, 821–829.

[85] Li, H. Minimap and Miniasm: Fast Mapping and de Novo Assembly for Noisy Long Sequences. *Bioinformatics* **2016**, *32*, 2103–2110.

[86] Kolmogorov, M.; Yuan, J.; Lin, Y.; Pevzner, P. A. Assembly of Long, Error-Prone Reads Using Repeat Graphs. *Nature Biotechnology* **2019**, *37*, 540–546.

[87] Wick, R. R.; Holt, K. E. Benchmarking of Long-Read Assemblers for Prokaryote Whole Genome Sequencing. *F1000Research* **2021**, *8*, 2138.

[88] Chin, C.-S.; Alexander, D. H.; Marks, P.; Klammer, A. A.; Drake, J.; Heiner, C.; Clum, A.; Copeland, A.; Huddleston, J.; Eichler, E. E.; Turner, S. W.; Korlach, J. Nonhybrid, Finished Microbial Genome Assemblies from Long-Read SMRT Sequencing Data. *Nature Methods* **2013**, *10*, 563–569.

[89] Loman, N. J.; Quick, J.; Simpson, J. T. A Complete Bacterial Genome Assembled de Novo Using Only Nanopore Sequencing Data. *Nature Methods* **2015**, *12*, 733–735.

[90] Walker, B. J.; Abeel, T.; Shea, T.; Priest, M.; Abouelliel, A.; Sakthikumar, S.; Cuomo, C. A.; Zeng, Q.; Wortman, J.; Young, S. K.; Earl, A. M. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PloS One* **2014**, *9*, e112963.

[91] Vaser, R.; Sović, I.; Nagarajan, N.; Šikić, M. Fast and Accurate de Novo Genome Assembly from Long Uncorrected Reads. *Genome Research* **2017**, *27*, 737–746.

[92] Shafin, K.; Pesout, T.; Chang, P.-C.; Nattestad, M.; Kolesnikov, A.; Goel, S.; Baid, G.; Kolmogorov, M.; Eizenga, J. M.; Miga, K. H.; Carnevali, P.; Jain, M.; Carroll, A.; Paten, B. Haplotype-Aware Variant Calling with PEPPER-Margin-DeepVariant Enables High Accuracy in Nanopore Long-Reads. *Nature Methods* **2021**, *18*, 1322–1332.

[93] Nurk, S. et al. The Complete Sequence of a Human Genome. 2021.

[94] International Human Genome Sequencing Consortium, Finishing the Euchromatic Sequence of the Human Genome. *Nature* **2004**, *431*, 931–945.

[95] Cartney, A. M. M. et al. Chasing Perfection: Validation and Polishing Strategies for Telomere-to-Telomere Genome Assemblies. 2021.

[96] Tyson, G. W.; Chapman, J.; Hugenholtz, P.; Allen, E. E.; Ram, R. J.; Richardson, P. M.; Solovyev, V. V.; Rubin, E. M.; Rokhsar, D. S.; Banfield, J. F. Community Structure and Metabolism through Reconstruction of Microbial Genomes from the Environment. *Nature* **2004**, *428*, 37–43.

[97] Parks, D. H.; Rinke, C.; Chuvochina, M.; Chaumeil, P.-A.; Woodcroft, B. J.; Evans, P. N.; Hugenholtz, P.; Tyson, G. W. Recovery of Nearly 8,000 Metagenome-Assembled Genomes Substantially Expands the Tree of Life. *Nature Microbiology* **2017**, *2*, 1533–1542.

[98] Delmont, T. O.; Quince, C.; Shaiber, A.; Esen, Ö. C.; Lee, S. T.; Rappé, M. S.; McLellan, S. L.; Lücker, S.; Eren, A. M. Nitrogen-Fixing Populations of Planctomycetes and Proteobacteria Are Abundant in Surface Ocean Metagenomes. *Nature Microbiology* **2018**, *3*, 804–813.

[99] Rheims, H.; Rainey, F. A.; Stackebrandt, E. A Molecular Approach to Search for Diversity among Bacteria in the Environment. *Journal of Industrial Microbiology* **1996**, *17*, 159–169.

[100] Nurk, S.; Meleshko, D.; Korobeynikov, A.; Pevzner, P. A. metaSPAdes: A New Versatile Metagenomic Assembler. *Genome Research* **2017**, *27*, 824–834.

[101] Bankevich, A. et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology* **2012**, *19*, 455–477.

[102] Li, D.; Liu, C.-M.; Luo, R.; Sadakane, K.; Lam, T.-W. MEGAHIT: An Ultra-Fast Single-Node Solution for Large and Complex Metagenomics Assembly via Succinct de Bruijn Graph. *Bioinformatics* **2015**, *31*, 1674–1676.

[103] Alneberg, J.; Bjarnason, B. S.; de Bruijn, I.; Schirmer, M.; Quick, J.; Ijaz, U. Z.; Lahti, L.; Loman, N. J.; Andersson, A. F.; Quince, C. Binning Metagenomic Contigs by Coverage and Composition. *Nature Methods* **2014**, *11*, 1144–1146.

[104] Kang, D. D.; Li, F.; Kirton, E.; Thomas, A.; Egan, R.; An, H.; Wang, Z. MetaBAT 2: An Adaptive Binning Algorithm for Robust and Efficient Genome Reconstruction from Metagenome Assemblies. *PeerJ* **2019**, *7*, e7359.

[105] Sieber, C. M. K.; Probst, A. J.; Sharrar, A.; Thomas, B. C.; Hess, M.; Tringe, S. G.; Banfield, J. F. Recovery of Genomes from Metagenomes via a Dereplication, Aggregation and Scoring Strategy. *Nature Microbiology* **2018**, *3*, 836–843.

[106] Eren, A. M.; Esen, Ö. C.; Quince, C.; Vineis, J. H.; Morrison, H. G.; Sogin, M. L.; Delmont, T. O. Anvi'o: An Advanced Analysis and Visualization Platform for 'omics Data. *PeerJ* **2015**, *3*, e1319.

[107] Kolmogorov, M.; Bickhart, D. M.; Behsaz, B.; Gurevich, A.; Rayko, M.; Shin, S. B.; Kuhn, K.; Yuan, J.; Polevikov, E.; Smith, T. P. L.; Pevzner, P. A. metaFlye: Scalable Long-Read Metagenome Assembly Using Repeat Graphs. *Nature Methods* **2020**, *17*, 1103–1110.

[108] Koren, S.; Walenz, B. P.; Berlin, K.; Miller, J. R.; Bergman, N. H.; Phillippy, A. M. Canu: Scalable and Accurate Long-Read Assembly via Adaptive k-Mer Weighting and Repeat Separation. *Genome Research* **2017**, *27*, 722–736.

[109] Vaser, R.; Šikić, M. Time- and Memory-Efficient Genome Assembly with Raven. *Nature Computational Science* **2021**, *1*, 332–336.

[110] Nicholls, S. M.; Quick, J. C.; Tang, S.; Loman, N. J. Ultra-Deep, Long-Read Nanopore Sequencing of Mock Microbial Community Standards. *GigaScience* **2019**, *8*.

[111] Moss, E. L.; Maghini, D. G.; Bhatt, A. S. Complete, Closed Bacterial Genomes from Microbiomes Using Nanopore Sequencing. *Nature Biotechnology* **2020**, *38*, 701–707.

[112] Feng, X.; Cheng, H.; Portik, D.; Li, H. Metagenome Assembly of High-Fidelity Long Reads with Hifiasm-Meta. *arXiv:2110.08457 [q-bio]* **2021**,

[113] 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* **2020**, 2020.04.08.032540.

[114] Hunt, M.; Kikuchi, T.; Sanders, M.; Newbold, C.; Berriman, M.; Otto, T. D. REAPR: A Universal Tool for Genome Assembly Evaluation. *Genome Biology* **2013**, *14*, R47.

[115] Gurevich, A.; Saveliev, V.; Vyahhi, N.; Tesler, G. QUAST: Quality Assessment Tool for Genome Assemblies. *Bioinformatics (Oxford, England)* **2013**, *29*, 1072–1075.

[116] Mikheenko, A.; Saveliev, V.; Gurevich, A. MetaQUAST: Evaluation of Metagenome Assemblies. *Bioinformatics* **2016**, *32*, 1088–1090.

[117] Parks, D. H.; Imelfort, M.; Skennerton, C. T.; Hugenholtz, P.; Tyson, G. W. CheckM: Assessing the Quality of Microbial Genomes Recovered from Isolates, Single Cells, and Metagenomes. *Genome Research* **2015**, *25*, 1043–1055.

[118] Simão, F. A.; Waterhouse, R. M.; Ioannidis, P.; Kriventseva, E. V.; Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs. *Bioinformatics (Oxford, England)* **2015**, *31*, 3210–3212.

[119] Hug, L. A. et al. A New View of the Tree of Life. *Nature Microbiology* **2016**, *1*, 1–6.

[120] Watson, M.; Warr, A. Errors in Long-Read Assemblies Can Critically Affect Protein Prediction. *Nature Biotechnology* **2019**, *37*, 124–126.

[121] Rhie, A.; Walenz, B. P.; Koren, S.; Phillippy, A. M. Merqury: Reference-Free Quality, Completeness, and Phasing Assessment for Genome Assemblies. *Genome Biology* **2020**, *21*, 245.

[122] Filloramo, G. V.; Curtis, B. A.; Blanche, E.; Archibald, J. M. Re-Examination of Two Diatom Reference Genomes Using Long-Read Sequencing. *BMC Genomics* **2021**, *22*, 379.

Chapter 2

Filtering long reads detects deletions in genome assemblies

2.1 Introduction

New long read sequencing platforms such as the Oxford Nanopore minION and PacBio HiFi platforms facilitate bacterial genome assembly. Both platforms can provide read lengths that are longer than typical repetitive regions in bacterial genomes, enabling accurate and complete isolate bacterial genome assembly. Recent advances to the Oxford Nanopore platform, such as enhanced basecalling accuracy and increases in both read N50 (a measure of read length) and throughput, have vastly improved the ability to generate accurate genome assemblies. Furthermore, longer read N50s and new metagenome assembly algorithms (1–3) now permit the completion of metagenomically-assembled whole genomes (MAWGs) directly from bacterial communities. This has been validated in mock bacterial communities (4) and applied to novel bacterial communities (5, 6). Depending on the complexity, composition of the community, and data quality, it is now possible to circularize the most abundant species, and even to fully assemble and close most bacterial genomes from a single metagenomic nanopore sequencing experiment. Furthermore, there may be two or more populations of single species of alternate haplotypes that co-exist in a metagenome due to the addition or removal of mobile genetic elements.

However, the ability to generate multiple complete genomes directly from metagenomic sequencing experiments should come with the responsibility to ensure each genome is high quality. The gold standard to evaluate genome completeness and contiguity is to ensure consistent read coverage using paired-end Illumina reads and a tool such as REAPR (7). This validation procedure is not possible when Nanopore-only sequencing is used. Other common genome quality assessment tools like CheckM (8) and BUSCO (9) look for the presence of marker genes. However, this approach can only provide information about the expected gene content for a complete genome, not the contiguity of the assembly itself. This approach also has only recently accounted for genomes of unexpected low marker gene content, such as the smaller genomes from the Candidate Phyla Radiation (10). Recently developed tools such as Merqury use a k-mer based approach when Illumina reads are available to estimate genome assembly quality (11), however obtaining complimentary Illumina data may not always be possible.

At a minimum, a complete and accurate nanopore-assembled MAWG should have tiled and consistent long-read coverage with no gaps present. Recently developed workflows, such as Lathe (5), identifies misassemblies by finding regions spanned by one or zero long reads in windows smaller than the average read length. Another proposed validation method is to visualize each contig or completed genome with a coverage plot of filtered reads by query coverage. Such an approach has been used for validating contiguity in hybrid-assembled contigs (12) and has been proposed to evaluate *de novo* assembled genomes from metagenomes in a reference-free manner (6). In addition to visualizing coverage, ensuring that only correct alignments are retained can influence downstream processes such as affecting the functional understanding of the bacteria in question. For Illumina reads, other tools have been developed to filter reads by removing any reads with soft or hard clips like SamClip (13). However, this approach is not suitable for long error-prone reads since was designed for short reads only, and it will falsely exclude many true alignments because many long read alignments for nanopore typically contain soft-clipping (i.e., bases are trimmed from 5’ and 3’ end of alignment) due to lower per-base accuracy of nanopore reads. Filtering long reads by query coverage has been used previously (6, 12) and has been implemented in recent isolate bacterial genome assembly workflows like Trycycler (14). However, investigation of the consequences of mis-mapped long reads in completed genomes generated from metagenomic data has yet to be described.

In this work we propose that each long-read MAWG should be individually evaluated using a coverage-based approach. We show that filtering Nanopore reads by query coverage and length is essential to identify bacterial strains or haplotypes and mis-assembled genomes. We demonstrate that applying filtering reduces the number of mis-aligned reads for genomes extracted from whole metagenome datasets. Performing this extra step for each metagenomically-assembled whole genome will help ensure that only high-quality MAWGs are deposited into public databases. Filtering is enabled by a fast, easy-to-use, and publicly available tool called Gerenuq developed in this work for common alignment formats.

Note: During the time that data was collected and analyzed for this project in 2019, state-of-the-art genome assemblers for long reads were not able to resolve haplotypes in bacterial metagenomes. It was not until October and November 2020 that tools were available for resolving haplotypes, were published and publicly available (1, 15). The initial results for this approach were developed in 2019 and posted as a pre-print in April 2020 (6). This chapter represents an approach that was necessary for me to develop to understand data I generated in 2019, but would now otherwise be reported directly from bacteria metagenome assembly algorithms that are haplotype-aware that were published in late 2020, such as metaFlye v2.7 (1). This approach still represents an alternative method that can be used to investigate genomes.

2.2 Methods

Sequencing reads were obtained from a previous sequencing run (fully described in Chapter 5, and raw data is available from the European Nucleotide Archive project PRJEB36155). Genomes that I assembled from a previous metagenome study were used for this analysis (6). To summarize the previous work, the workflow was as follows: to investigate the composition of a 1-adamantanecarboxylic acid degrading microbial community (16) high-molecular weight DNA was extracted from and the Short Read Eliminator Kit (Circulomics) protocol was applied before sequencing on both a Oxford Nanopore MinION R9.4.1 flow cell, and Illumina NextSeq 550 mid-output. For Nanopore sequencing, a read N50 of approximately 24 kb was achieved. Metagenomic assembly was performed using metaFlye v2.6 (1) and polished using Racon (17) and Pilon (18).

A previously metagenomically-assembled whole genome was used (6) for develop this approach *Blastomonas*. *Blastomonas* was arbitrarily chosen as a the example genome due to sufficient sequencing depth. A 100 kb deletion was manually introduced into the genome to demonstrate the increased alignment quality after filtering. These reads were mapped against the genomes using minimap2 using the parameters -aLQx map-ont -t 40 and filtered using Gerenuq. Read depth was calculated in 1000 base windows using mosdepth (19). Plots were generated using the R package circlize (20).

Gerenuq (v0.2.6) can be installed via conda (conda install -c conda-forge -c bioconda -c abahcheli gerenuq), pip (pip install gerenuq) or Github (git clone https://github.com/abahcheli/gerenuq). Gerenuq can filter bam, sam and paf files according to default or user defined parameters from a command line tool.

2.3 Results

After mapping long reads against the *Blastomonas* genome with default minimap2 settings, reads were filtered using Gerenuq with the parameters -m 0.95 and -l 5000; this means that 95 percent of the read must be aligned against the draft genome and the minimum read length is 5 kb. A summary measuring the speed of the script is shown in Supplemental Figure A.1. We found that using 4 or more threads results in an acceptable trade off between time required and computing power, with the maximum speed at approximately 200 mega-bp per second.

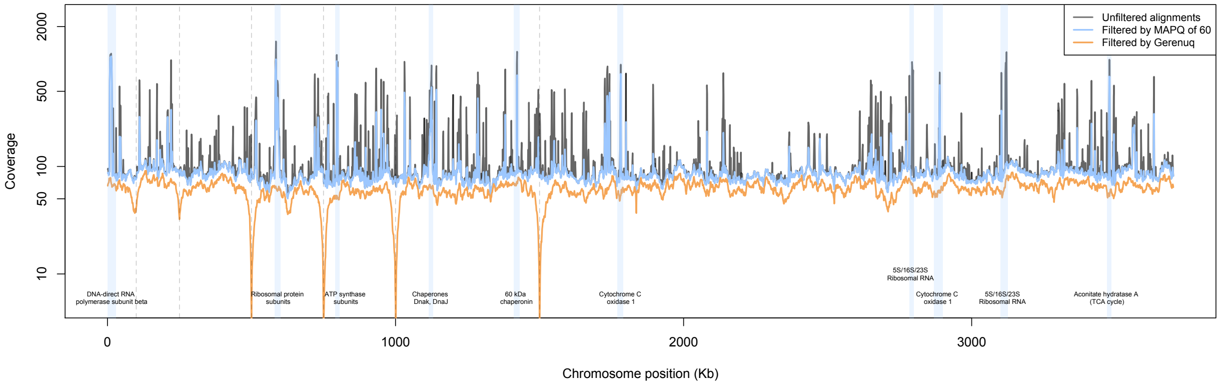


Figure 2.1: **Example of a missed deletions**. 0.5 kb, 1 kb, 5 kb, 10 kb, 20 kb, 100 kb of sequence were arbitrarily deleted from a previously metagenomically-assembled whole genome (dashed grey lines from left to right: 0.5, 1, 5, 10, 20, 100 kilobases). All metagenomic reads were mapped against only this genome using minimap2 with the parameters -aLQx map-ont -t 40. Coverage was then calculated in 1000 base windows for unfiltered, filtered by a MAPQ of 60, and by Gerenuq (minimum 5000 base, minimum 95 percent query coverage). Prokka annotations for regions with higher than 500X coverage after filtering by MAPQ are shown in light blue boxes, in addition to 16S rRNA genes

Deletions down to 500 bp remain undetected without filtering reads

Long read coverage can be used to evaluate the contiguity of the genome by ensuring consistent tiling coverage with no gaps, part of the recently proposed genome assembly reporting criteria for complete genomes (21). When assembling genomes from isolates, consistent read coverage can often be observed without any filtering since few, if any, repetitive elements from outside the genome are present. However, for genomes assembled from metagenomes, common genomic regions can result in multiple alignments per read and incorrect alignments can be reported even if the alignment score threshold is high (such as a MAPQ of 60). Figure 2.1 shows the coverage of the *Blastomonas* genome with unfiltered (grey), filtered by a MAPQ of 60 using samtools (22) (blue), and filtered by Gerenuq (orange) using a minimum length of 5000 kb reads and 95% query coverage. The increases in coverage are regions with a significant number of incorrectly aligned reads; these are derived from conserved sequences in many bacteria from the metagenome (highlighted in light blue boxes). The median read coverage for the unfiltered reads is approximately

85X, however, over 400 coverage windows have a coverage greater than 150X with several up to 1700X. Filtering using the highest possible MAPQ reported by default minimap2 settings (MAPQ of 60) helps reduce some incorrect alignments. However, even after this filtering there are still many regions with significant coverage spikes over 1000X, indicating retention of mis-aligned reads. After filtering with Gerenuq using a minimum read length of 5000 bp and a query coverage of 95 percent, none of the coverage windows showed coverage greater than 2X the median, suggesting few, if any, mis-mapped alignments are retained.

Importantly, filtering by query coverage revealed drops in coverage inconsistent with a complete genome where we manually introduced deletions of various sizes (vertical grey dashes). This demonstrates that deletions down to 500 base pairs can be detected visually, although this depends on how strict the query coverage cutoff is. For example, at a 1000 base deletion, you may still expect a read of 15 kb to pass filtering since 14000 matches divided by 15000 bases is a theoretical 93% query coverage. This explains why some coverage remains at the 0.5 and 1 kb deletions. The alignments that are unfiltered and filtered by MAPQ coverage appear completely consistent at the deleted region, which would have resulted in missing these deletions in the assembly.

Haplotypes can be detected and resolved using filtered ultra-long reads

Although drops in filtered long-read coverage typically indicate deletions or mis-assemblies, it is also possible that two or more populations of a single species or alternate haplotypes co-exist in a metagenome. In this case, evaluating whether reads overlap can indicate whether the coverage drop is due to a mis-assembly or a true alternate bacterial haplotype in the population. This is shown in Figure 2.2 using the *Parvibaculum* genome. After filtering Nanopore reads, a drop in read coverage was observed that is characteristic of a deletion in the assembly. However, it was found that a tiling path was supported by the reads and no gap existed in the assembly. By taking reads that partially mapped where this drop in coverage occurred, it was found that an additional 35 kb region was supported by the majority of reads (Supplemental Figure A.2). Filtered long Nanopore reads span the entire region for both versions of this genome, providing evidence that

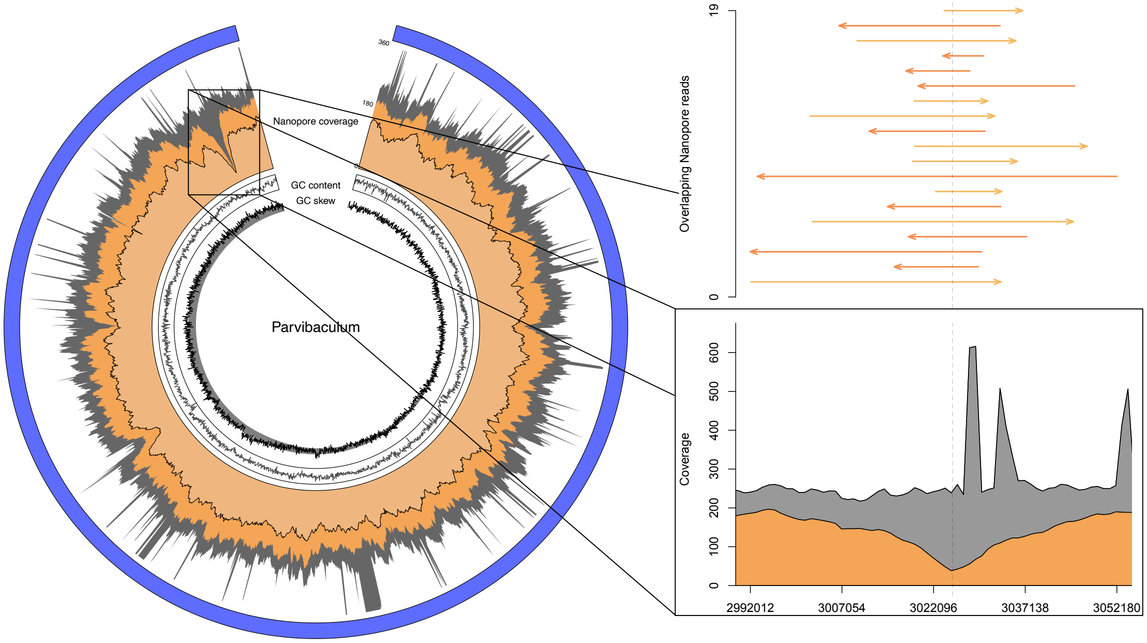


Figure 2.2: **Filtering Nanopore reads reveals a haplotype**. Left; circlize plot of coverage for reads filtered by MAPQ = 60 (grey), filtered by 90% coverage and 5000 read length (dark orange), filtered by 90% coverage and 15000 base read length (light orange). Top right; overlapping reads that span the entirety of the low coverage region. Bottom right; close up view of coverage plot. Dotted vertical line indicates the lowest coverage point.

both haplotypes truly exist in the population. Not filtering by query coverage resulted in missing this potentially biologically relevant region altogether. An additional 32 coding sequences were recovered from this region using prokka (23). Importantly, a ferrodoxin protein was found in this region, demonstrating that key proteins related to important functions such as sulphate assimilation may be recovered by extracting alternative haplotype sequences.

Mis-assembled genome detected by filtering reads only

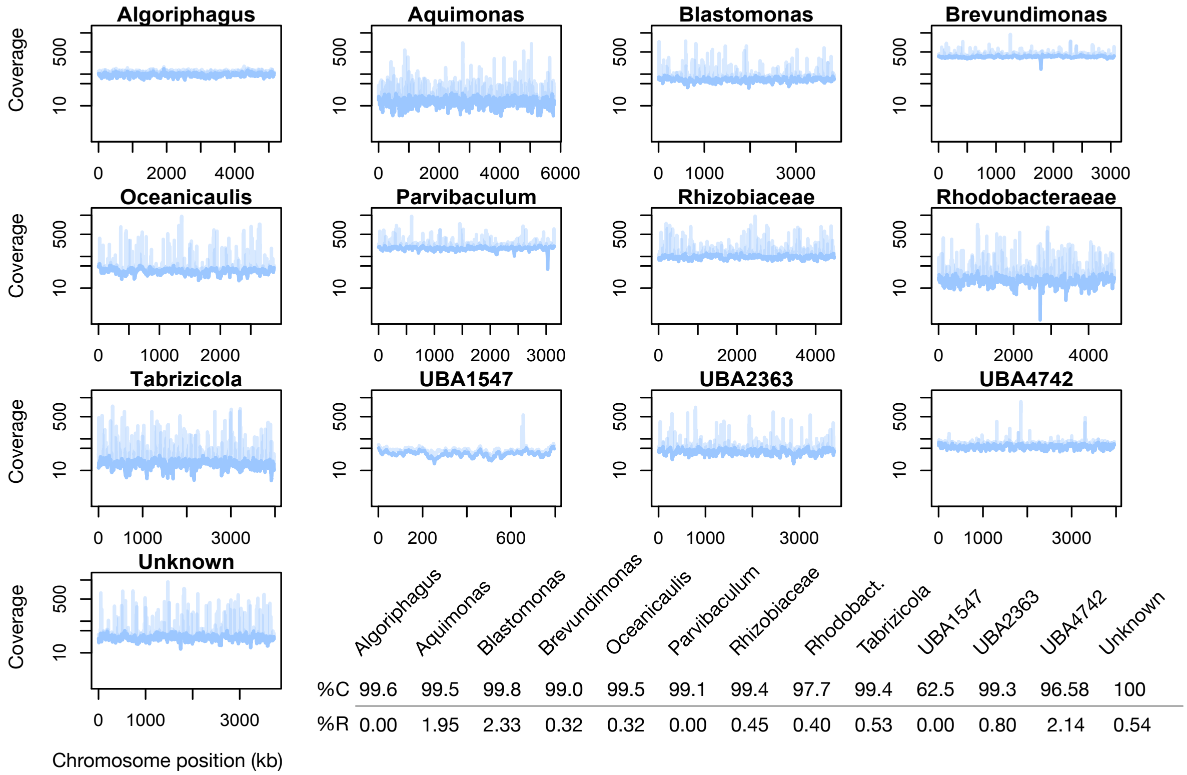
We applied this method to all 13 MAWGs assembled from a previous study (6), across a variety of read depths, shown in Figure 2.3. While many of the genomes have similar GC content around 60-

70 percent, it appears that unique genomes (e.g., low GC content in a high GC community) have fewer mis-mapped reads. In addition, the *Rhodobacteraceae* genome was reported as circularized from the assembly output even though there is a region where filtered coverage drops to zero. This indicates that there is a mis-assembly in the genome, and that it should be further refined before being considered complete.

While completeness and contamination estimates by CheckM (8) can provide evidence that a draft genome may contain the full set of expected genes in a genome, filtering long-read coverage was required to determine that contiguity is incomplete in the *Parvibaculum*, *Brevundimonas* and *Rhodobacteraceae* genomes. Figure 2.3 demonstrates the existence of alternative haplotypes in *Parvibaculum* and *Brevundimonas*, and an assembly error in *Rhodobacteraceae*. Using Gerenuq, we were able to confirm the lowest coverage genome (*Aquimonas*) is contiguous at an average of 13X average coverage, while detecting an assembly error at 17X coverage.

2.4 Discussion

With the rapid improvements of long read sequencing technology, complete genomes can now be assembled directly from metagenomic data thanks to improved basecalling accuracy, read N50, throughput, and assembly algorithms. Complete genome assemblies from both isolates and metagenomes will be more commonly generated and be of higher quality in the future as high-molecular weight DNA extraction and sequencing protocols are improved, and as the available computational tools improve. While the output from assembly algorithms are often correct, validation of each individual genome is necessary to ensure a contiguous assembly is present before submitting to public databases.



	Algoriphagus	Aquimonas	Blastomonas	Brevundimonas	Oceanicaulis	Parvibaculum	Rhizobiaceae	Rhodobact.	Tabrizicola	UBA1547	UBA2363	UBA4742	Unknown
%C	99.6	99.5	99.8	99.0	99.5	99.1	99.4	97.7	99.4	62.5	99.3	96.58	100
%R	0.00	1.95	2.33	0.32	0.32	0.00	0.45	0.40	0.53	0.00	0.80	2.14	0.54

Figure 2.3: **Filtering Nanopore reads for a fully assembled community identifies mis-assemblies and haplotypes**. All metagenomic reads were mapped against each individual genome, and filtered by MAPQ = 60 (light blue) or Gerenuq (dark blue - length 5000 and query coverage of 90%). Coverage was calculated in 1000 base windows using mosdepth. Percent completion and contamination (or redundancy) are shown in bottom right portion for each genome calculated by CheckM (%C and %R, respectively).

Filtering long reads is essential to evaluate contiguity of metagenomically-derived whole genomes

When individual genomes are assembled and extracted from long-read metagenomic data, it is important to filter the reads by query coverage to ensure there are no mis-assemblies or deletions. This method is a reference-free way to evaluating contiguity of these genomes, and also enables the detection of large indels and alternate haplotypes. Performing this extra step will help identify genomes that are not fully circular (i.e., represented by more than one contig), and can potentially lead to additional functional information of the bacteria. We deleted multiple regions from a MAWG and showed that filtering alignments by alignment score is not sufficient to detect even a 100 kb deletion in the assembly. Thus, visualizing potential assembly errors within a MAWG is possible using long-reads only when they are filtered by query coverage.

Limitations

Recently developed assemblers, including the assembler metaFlye (1), Canu (2), appear to perform quite well with a read N50 of greater than 10 kb for metagenomic datasets. However, high coverage is typically required since a fraction of reads is removed when filtering by read length and query coverage. In this example, median coverage was reduced from about 85X to about 60X after filtering. Having high coverage is especially important to perform this filtering on genomes that are relatively low abundance in the population.

Due to how minimap2 aligns reads to a fasta file, any reads that overlaps a circular genome at the beginning and end of a fasta file will be reported as two separate alignments. To ensure drops in coverage don't occur due to this for circular elements, the query coverage is calculated using the query end - query start instead of the query length in regions near the start and end of the file. We also note that this is only observed for circular genomic elements like genomes or circular plasmids.

Conclusions

Gerenuq will enable researchers to improve their metagenomically-assembled whole genomes in two ways. First, by ensuring they are indeed contiguous (even when at low coverage), which will reduce the probability of introducing erroneous assemblies into public databases. Second, by enabling researchers to extract potentially relevant functional information from alternate bacterial haplotypes. This may be especially relevant when considering mobile functional genetic elements.

2.5 References

Bibliography

[1] Kolmogorov, M.; Bickhart, D. M.; Behsaz, B.; Gurevich, A.; Rayko, M.; Shin, S. B.; Kuhn, K.; Yuan, J.; Polevikov, E.; Smith, T. P. L.; Pevzner, P. A. metaFlye: Scalable Long-Read Metagenome Assembly Using Repeat Graphs. *Nature Methods* **2020**, *17*, 1103–1110.

[2] Koren, S.; Walenz, B. P.; Berlin, K.; Miller, J. R.; Bergman, N. H.; Phillippy, A. M. Canu: Scalable and Accurate Long-Read Assembly via Adaptive k-Mer Weighting and Repeat Separation. *Genome Research* **2017**, *27*, 722–736.

[3] Li, H. Minimap and Miniasm: Fast Mapping and de Novo Assembly for Noisy Long Sequences. *Bioinformatics* **2016**, *32*, 2103–2110.

[4] Nicholls, S. M.; Quick, J. C.; Tang, S.; Loman, N. J. Ultra-Deep, Long-Read Nanopore Sequencing of Mock Microbial Community Standards. *GigaScience* **2019**, *8*.

[5] Moss, E. L.; Maghini, D. G.; Bhatt, A. S. Complete, Closed Bacterial Genomes from Microbiomes Using Nanopore Sequencing. *Nature Biotechnology* **2020**, *38*, 701–707.

[6] 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* **2020**, 2020.04.08.032540.

[7] Hunt, M.; Kikuchi, T.; Sanders, M.; Newbold, C.; Berriman, M.; Otto, T. D. REAPR: A Universal Tool for Genome Assembly Evaluation. *Genome Biology* **2013**, *14*, R47.

[8] Parks, D. H.; Imelfort, M.; Skennerton, C. T.; Hugenholtz, P.; Tyson, G. W. CheckM: Assessing the Quality of Microbial Genomes Recovered from Isolates, Single Cells, and Metagenomes. *Genome Research* **2015**, *25*, 1043–1055.

[9] Simão, F. A.; Waterhouse, R. M.; Ioannidis, P.; Kriventseva, E. V.; Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs. *Bioinformatics (Oxford, England)* **2015**, *31*, 3210–3212.

[10] Brown, C. T.; Hug, L. A.; Thomas, B. C.; Sharon, I.; Castelle, C. J.; Singh, A.; Wilkins, M. J.; Wrighton, K. C.; Williams, K. H.; Banfield, J. F. Unusual Biology across a Group Comprising More than 15% of Domain Bacteria. *Nature* **2015**, *523*, 208–211.

[11] Rhie, A.; Walenz, B. P.; Koren, S.; Phillippy, A. M. Merqury: Reference-Free Quality, Completeness, and Phasing Assessment for Genome Assemblies. *Genome Biology* **2020**, *21*, 245.

[12] Caceres, E. F.; Lewis, W. H.; Homa, F.; Martin, T.; Schramm, A.; Kjeldsen, K. U.; Ettema, T. J. G. Near-Complete Lokiarchaeota Genomes from Complex Environmental Samples Using Long and Short Read Metagenomic Analyses. 2019.

[13] Seemann, T. Samclip. 2021.

[14] Wick, R. R.; Judd, L. M.; Cerdeira, L. T.; Hawkey, J.; Méric, G.; Vezina, B.; Wyres, K. L.; Holt, K. E. Trycycler: Consensus Long-Read Assemblies for Bacterial Genomes. 2021.

[15] Nicholls, S. M.; Aubrey, W.; De Grave, K.; Schietgat, L.; Creevey, C. J.; Clare, A. On the Complexity of Haplotyping a Microbial Community. *Bioinformatics (Oxford, England)* **2020**, btaa977.

[16] Paulssen, J. M.; Gieg, L. M. Biodegradation of 1-Adamantanecarboxylic Acid by Algal-Bacterial Microbial Communities Derived from Oil Sands Tailings Ponds. *Algal Research* **2019**, *41*, 101528.

[17] Vaser, R.; Sović, I.; Nagarajan, N.; Šikić, M. Fast and Accurate de Novo Genome Assembly from Long Uncorrected Reads. *Genome Research* **2017**, *27*, 737–746.

[18] Walker, B. J.; Abeel, T.; Shea, T.; Priest, M.; Abouelliel, A.; Sakthikumar, S.; Cuomo, C. A.; Zeng, Q.; Wortman, J.; Young, S. K.; Earl, A. M. Pilon: An Integrated Tool for Comprehen-

sive Microbial Variant Detection and Genome Assembly Improvement. *PloS One* **2014**, *9*, e112963.

[19] Pedersen, B. S.; Quinlan, A. R. Mosdepth: Quick Coverage Calculation for Genomes and Exomes. *Bioinformatics* **2018**, *34*, 867–868.

[20] Gu, Z.; Gu, L.; Eils, R.; Schlesner, M.; Brors, B. Circlize Implements and Enhances Circular Visualization in R. *Bioinformatics (Oxford, England)* **2014**, *30*, 2811–2812.

[21] Rhie, A. et al. Towards Complete and Error-Free Genome Assemblies of All Vertebrate Species. *Nature* **2021**, *592*, 737–746.

[22] Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map Format and SAMtools. *Bioinformatics (Oxford, England)* **2009**, *25*, 2078–2079.

[23] Seemann, T. Prokka: Rapid Prokaryotic Genome Annotation. *Bioinformatics* **2014**, *30*, 2068–2069.

Chapter 3

Long read karyocounting: an assembly-free method to estimate the number of chromosomes in eukaryotic genomes

3.1 Introduction

Improvements to DNA sequencing technology and genome assembly algorithms have vastly improved the fundamental understanding of many organisms by facilitating more contiguous and higher quality genome assembly. Long-read sequencing technologies, like the Oxford Nanopore Technologies MinION platform, are increasing the contiguity of genome assemblies by overcoming previous limitations of read length and GC bias, generating reads that are hundreds of kilobases long. While contiguity is significantly improved thanks to long-reads, a fully complete eukaryotic genome would contain all telomere-to-telomere chromosomes. This is often not possible to obtain directly from an assembly algorithm's output, since partial chromosomes in the form of contigs are often produced due to assembly errors near repetitive regions like telomeres. Complex repeat regions often need to be manually resolved.

State-of-the-art long-read assemblers like Canu (1), Flye (2), Shasta (3) are capable of high quality eukaryotic genome assembly, but only report on the size and number of assembled contiguous DNA sequences produced, leaving it to the researcher to determine the number of chromosomes an organism may contain. Genome assemblies for novel eukaryotic organisms will often fail to answer the fundamental biological question of “how many chromosomes does this organism have?” without significant manual intervention. Novel eukaryotes are now being studied as potential platforms for use in synthetic biology applications, like the marine diatom *Phaeodactylum tricornutum*. Although a high quality draft genome assembly has been available since 2008 (4), the number of nuclear chromosomes in *P. tricornutum* was still unknown as of early 2021 (5–7), limiting potential genome engineering applications such as complete genome synthesis and replacement.

Here, we developed an assembly- and reference-free approach to estimate the number of linear chromosomes in small eukaryotic genomes directly from long nanopore reads.

3.2 Methods

3.2.1 Data

To develop this method, I used the sequencing reads from *Phaeodactlyum tricornutum*, fully described in Chapter 4 and elsewhere (8). The data is publicly available from published European Nucleotide Archive project ID: PRJEB42700. To summarize the previous study, high-molecular weight DNA was extracted and sequenced using an Oxford Nanopore Technologies MinION flow cell version R9.4.1, using the SQK-LSK109 library preparation kit according to the manufacturer’s protocol version GDE 9063 v109 revK 14 Aug 2019, with one alteration: for DNA repair and end-prep, the reaction mixture was incubated for 15 minutes at 20° C and 15 minutes at 65° C. Basecalling was performed with Guppy in high-accuracy mode (v3.6). A summary of the throughput and read length is shown in Figure 3.1. We achieve a read n50 of 35 kilobases, and collected approximately 7.8 gigabases of sequences. The full workflow is described in Chapter 4 and previous work (8).

3.2.2 Workflow

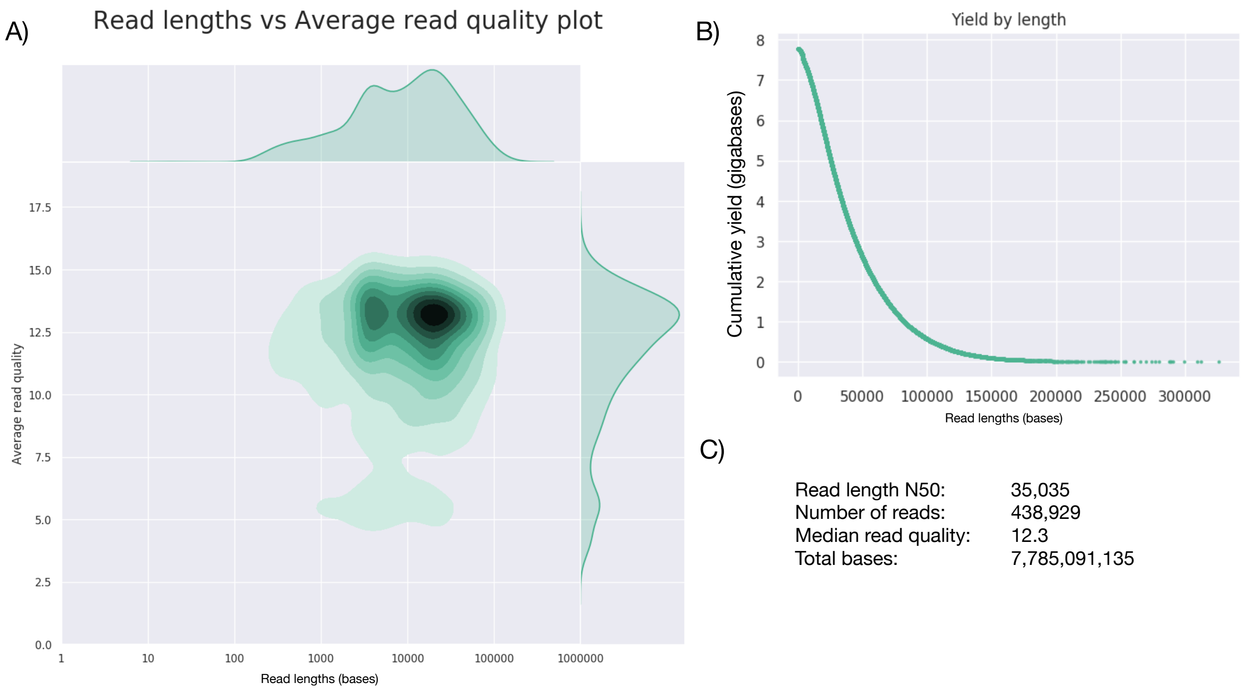
Each chromosome in a haploid set contains two unique telomeres, at the start and the end of the chromosome, while a pair of homologous chromosomes in a diploid set contains 4 unique telomeres. Therefore, the number of haploid chromosomes h, can be represented relative to the number of telomeres, t, as

$$h = \frac{t}{2} \tag{3.1}$$

and the number of diploid chromosomes, d, can be represented as

$$d = \frac{t}{4} \tag{3.2}$$

This approach relies on obtaining sequencing reads that contain both the telomeric repeats and unique sub-telomeric sequence for each chromosome to extract all unique telomeres in the dataset (Supplemental Figure 1A). Long telomere-containing reads are extracted using string matching (3 telomeric repeats or the reverse complement). The telomere-containing reads are then aligned in all vs. all mode using minimap2 (9), and filtered to retain only alignments with greater than 95% query coverage (i.e., full length alignments). The filtered target and query names are then used to build a network graph using iGraph, where each node represents a telomere-containing long sequencing read and each edge represents a filtered alignment between reads (10). In the ideal case, each component generated contains all long reads aligned to each other derived from a single telomere, meaning each network graph represents a single telomere. To ensure this is the case, all components are manually interpreted by visualization. The resulting components are enumerated, and the number of chromosomes is then estimated based on the expected ploidy of the sample. To demonstrate a use case, we use a publicly available dataset (ENA project PRJEB42700) to resolve the number of chromosomes in the marine diatom *Phaeodactylum tricornutum* (8).



Read length N50:	35,035
Number of reads:	438,929
Median read quality:	12.3
Total bases:	7,785,091,135

Figure 3.1: **Quality characteristics of ultra-long sequencing run generated by NanoPlot (11)**. A) Density plot of read length vs average Q score per read. B) Yield by read length. The cumulative yield refers to the total number of bases on a sequencing read larger than the denoted number on the x-axis. C) Various quality statistics.

3.3 Results

3.3.1 Each component of aligned reads represents a single telomere

We obtained 83 components with approximately 40 reads each, each representing a single telomere, and 9 components with more than 70 reads, that required further interpretation since they contained twice the number of expected reads (Figure 3.2 B). Six of components were composed of 2 clusters, where each cluster is a single telomere (Cluster 5 in Figure 3.2 C). Two of the remaining components were highly interconnected (similar to Cluster 2 in Figure 3.2 C), suggesting that there are no unique haplotypes from this chromosome distinguishable by sequence identity (i.e., there is only 1 haplotype). The single remaining component with high coverage was assigned as an individual telomere due to high interconnectedness like Cluster 2.

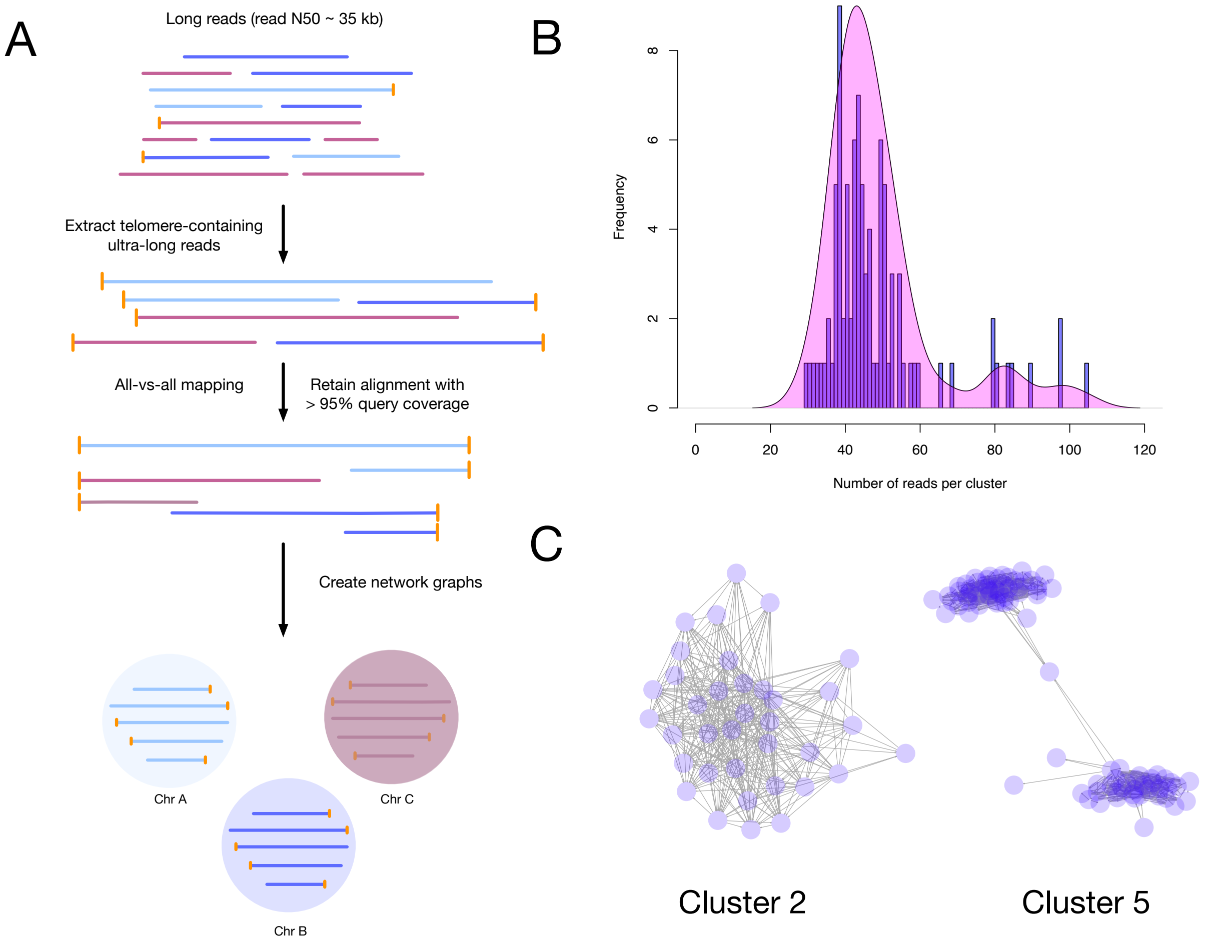


Figure 3.2: **Summary of network graph analysis**. A) General workflow for creating a network graph of telomere-containing reads. B) Histogram showing the frequency of the number of reads in each cluster. Overlayed is a density plot, showing three underlying distributions. C) Example clusters. Cluster 2 represents a a typical cluster of reads that have high inter-connectedness. Cluster 5 represents a component with two smaller clusters contained within it. Edges between vertices indicate that the read aligns to another with more than 95% query coverage.

3.3.2 Chromosome estimate from number of clusters

In total, we found 96 components that each represent a single telomere of a single haplotype, and 2 components that each represent 2 haplotypes of a single telomere (Table 3.1). This interpretation comes from the observation that one chromosome in the genome contains twice the expected sequencing coverage relative to other chromosomes (8) and loss of heterozygosity has been observed (12, 13). It's known that this species is diploid, so we therefore reasoned using equation 2 that the 96 single telomeres represent 24 chromosomes, and the 2 remaining components represent both haplotypes of a single chromosome, resulting in final count of 25 chromosomes. This orthogonal estimate agrees with our previous telomere-to-telomere assembly that comprised all previous large scaffolds from the initial draft assembly (8).

3.3.3 Validation using assembled genome

To confirm this, we aligned reads from each component against our previously proposed telomere-to-telomere genome assembly and assigned them to a chromosome in Table 3.2 (8). We found 22 of 25 assembled chromosomes contained the expected number of telomeres in the expected orientation (2 at start and 2 at end of a chromosome). However, we could assign only 3 telomeres for both chromosomes 3 and 8, and 5 telomere components were assigned to chromosome 23. Interestingly, a single telomere was placed in the middle of chromosome 4, in addition to the 4 expected components. Since telomeres were placed at both the start and end of the remaining 3 chromosomes, we believe at least a single haplotype of each of these chromosomes exists, for a total of 25 chromosomes. We hypothesize the unexpected number of telomeres and unexpected placement of a single telomere can be explained by mitotic recombination events, since recombination of a chromosomal arm may change the apparent location of a telomere (13).

Cluster description	Number of clusters	Number of chromosomes represented
Less than 70 reads	83	20.75
High coverage, 2 clusters	6	3
High coverage, chromosome 19	2	1
High coverage, 1 cluster	1	0.25
Total	92	25

Table 3.1: **Estimate of the number of chromosomes using the network components.** All components with less than 70 reads represent a single telomere of a single haplotype of a chromosome (2 telomeres and 2 haplotypes per chromosome, therefore there are 4 clusters per chromosome). Each components with 2 clusters represents 2 telomeres, and therefore takes only 2 dual-clusters per chromosome. Chromosome 19 has only two high coverage clusters, but the unique biology of *P. tricornutum* suggests only 1 haplotype exists for this specific chromosome. In total, there are 24 chromosomes with 2 haplotypes and 1 chromosome with 1 haplotype represented by these 92 telomere-containing long read components.

3.4 Discussion

Here, we developed an approach that we term “long-read karyotyping”, that is an assembly- and reference-free approach to estimate the number of chromosomes in eukaryotic microorganisms using only long reads. This approach also enables an orthogonal method to confirm the overall genome organisation proposed in the first telomere-to-telomere assembly for this species (8). We found that there are 24 chromosomes with two haplotypes and 1 chromosome with one haplotype for a total of 25 chromosomes, consistent with our previous telomere-to-telomere genome assembly (8). We show that long reads contain additional information about the chromosome number that previous sequencing technologies could not provide, enabled by the sequencing the full telomeres and sub-telomeric regions of each chromosome.

3.4.1 Consistency with known biology and genome assemblies

While 94 of 98 telomere clusters were uniquely assigned and agreed with the previously known biology of this organism, we believe the inconsistency of the remaining clusters are due to mitotic recombination events (13). For example, all of the long reads from one of the telomere clusters (cluster 11) aligns to the middle of chromosome 4. This is the only cluster to align in the middle of a chromosome. If this was due to a mis-assembly, and chromosome 4 was a combination of 2 chromosomes, we would expect to see additional telomere clusters representing the other haplotypes that also align at this location. Rather, we believe that this single cluster that does not align to the end of the chromosome represents a dynamic version of this chromosome that has recombined and is only present in a subset of the population of cells. We hypothesize the chromosomes with only an unexpected number (chromosome 3, 4, 8, 18) may be involved in more active recombination than other chromosomes.

3.4.2 Applications

Long-read karyotyping provides an orthogonal approach to validate complete genomes for eukaryotic organisms without a high-quality reference available. By applying this approach, it is now possible to answer the fundamental biological question of “how many chromosomes does this organism have?” to further the understanding of the chromosomal structure of novel organisms, such as fungi and diatoms. In addition, if the repeat sequences at the ends of the recently proposed linear genomic element “borgs” are known (14), we hypothesize this approach may be adapted to identify borgs from long read data in a metagenomic sequencing dataset.

3.4.3 Limitations

While long-read karyotyping may be easily applied to novel organisms with relatively small genomes, there are limitations that we anticipate. First, if the genome size is too large, it may be impractical to collect enough sequencing data. For example, the human genome is approximately 3 gigabases. To collect 100X sequencing coverage with current MinION flow cells would require 10 flow cells (assuming 30 gigabases per flowcell). However, it may be possible to use adaptive sequencing to enrich for telomere-containing reads using a program such such as ReadFish (15). Second, it is unknown how the ploidy of an organism affects this method. In the case of this diploid organism, it was manageable to resolve the 92 clusters according to known biology. For higher-ploidy organisms like plants or cell-lines, this may be more difficult. Mini-chromosomes may also cause additional complications in analysis. Furthermore, each telomere and sub-telomeric sequence must have enough unique sequence such that correct alignments can be retained by filtering by query coverage. For organisms with very large telomere repeats, this may not be possible to obtain enough reads that contain unique sub-telomere sequence.

3.5 Conclusions

Here, we developed an approach called long-read karyotyping to estimate the number of eukaryotic chromosomes present in an assembly- and reference-free method using only long reads. This approach can be applied to sequencing datasets generated from the Oxford Nanopore MinION platform for eukaryotic organisms with linear chromosomes. This will enable researchers to answer the fundamental biological question of "how many chromosomes does this organism have?" using only long DNA sequencing reads.

3.6 References

Bibliography

[1] Koren, S.; Walenz, B. P.; Berlin, K.; Miller, J. R.; Bergman, N. H.; Phillippy, A. M. Canu: Scalable and Accurate Long-Read Assembly via Adaptive k-Mer Weighting and Repeat Separation. *Genome Research* **2017**, *27*, 722–736.

[2] Kolmogorov, M.; Yuan, J.; Lin, Y.; Pevzner, P. A. Assembly of Long, Error-Prone Reads Using Repeat Graphs. *Nature Biotechnology* **2019**, *37*, 540–546.

[3] Shafin, K. et al. Nanopore Sequencing and the Shasta Toolkit Enable Efficient de Novo Assembly of Eleven Human Genomes. *Nature Biotechnology* **2020**, *38*, 1044–1053.

[4] Bowler, C. et al. The Phaeodactylum Genome Reveals the Evolutionary History of Diatom Genomes. *Nature* **2008**, *456*, 239–244.

[5] Bowler, C.; Falciatore, A. Phaeodactylum Tricornutum. *Trends in Genetics* **2019**, *35*, 706–707.

[6] Diner, R. E.; Noddings, C. M.; Lian, N. C.; Kang, A. K.; McQuaid, J. B.; Jablanovic, J.; Espinoza, J. L.; Nguyen, N. A.; Anzelmatti, M. A.; Jansson, J.; Bielinski, V. A.; Karas, B. J.; Dupont, C. L.; Allen, A. E.; Weyman, P. D. Diatom Centromeres Suggest a Mechanism for Nuclear DNA Acquisition. *Proceedings of the National Academy of Sciences* **2017**, *114*, E6015–E6024.

[7] Filloramo, G. V.; Curtis, B. A.; Blanche, E.; Archibald, J. M. Re-Examination of Two Diatom Reference Genomes Using Long-Read Sequencing. *BMC Genomics* **2021**, *22*, 379.

[8] Giguere, D. J.; Bahcheli, A. T.; Slattery, S. S.; Patel, R. R.; Flatley, M.; Karas, B. J.; Edgell, D. R.; Gloor, G. B. Telomere-to-Telomere Genome Assembly of Phaeodactylum Tricornutum. 2021.

[9] Li, H. Minimap and Miniasm: Fast Mapping and de Novo Assembly for Noisy Long Sequences. *Bioinformatics* **2016**, *32*, 2103–2110.

[10] Csárdi, G.; Nepusz, T. The Igraph Software Package for Complex Network Research. *undefined* **2006**,

[11] De Coster, W.; D'Hert, S.; Schultz, D. T.; Cruts, M.; Van Broeckhoven, C. NanoPack: Visualizing and Processing Long-Read Sequencing Data. *Bioinformatics* **2018**, *34*, 2666–2669.

[12] Russo, M. T.; Cigliano, R. A.; Sanseverino, W.; Ferrante, M. I. Assessment of Genomic Changes in a CRISPR/Cas9 Phaeodactylum Tricornutum Mutant through Whole Genome Resequencing. *PeerJ* **2018**, *6*, e5507.

[13] Bulankova, P. et al. Mitotic Recombination between Homologous Chromosomes Drives Genomic Diversity in Diatoms. *Current Biology* **2021**, *31*, 3221–3232.e9.

[14] Al-Shayeb, B.; Schoelmerich, M. C.; West-Roberts, J.; Valentin-Alvarado, L. E.; Sachdeva, R.; Mullen, S.; Crits-Christoph, A.; Wilkins, M. J.; Williams, K. H.; Doudna, J. A.; Banfield, J. F. Borgs Are Giant Extrachromosomal Elements with the Potential to Augment Methane Oxidation. 2021.

[15] Payne, A.; Holmes, N.; Clarke, T.; Munro, R.; Debebe, B. J.; Loose, M. Readfish Enables Targeted Nanopore Sequencing of Gigabase-Sized Genomes. *Nature Biotechnology* **2021**, *39*, 442–450.

Chromosome	Start-haplotypes	End-haplotypes	Number-of-clusters
1	30, 38	15, 20	4
2	14, 64	47, 67	4
3	31_1, 40	92	3
4	34, 61	11, 73, 46	5
5	32, 71	24, 81	4
6	22, 4	72, 86	4
7	17,66	28,74	4
8	53	16, 8	3
9	49, 90	31_2, 58	4
10	55, 79	85, 9	4
11	52, 76	63, 75	4
12	7_1, 7_2	45_1, 45_2	4
13	12_1, 12_2	69, 6	4
14	23, 51	33, 82	4
15	60, 70	18, 68	4
16	42, 80	87, 89	4
17	65, 91	1, 44	4
18	26, 36	19	3
19	57	48	2
20	10_1, 29	59, 62	4
21	13, 3	50, 83	4
22	25, 35	10_2, 78	4
23	2, 88	43, 5	4
24	39, 84	41, 77	4
25	54, 56	27, 37	4

Table 3.2: **Chromosome assignment of 92 clusters of telomere-containing long reads**. Start-haplotypes represents the cluster of reads that align near the 0-base telomere repeat on the forward strand, End-haplotypes represents clusters of reads that align at the telomere at the full length of the chromosome. The sole exception is cluster 11, which aligns in the middle of chromosome 4. Clusters that have two distinct smaller clusters are denoted with an underscore.

Chapter 4

Telomere-to-telomere genome assembly of *Phaeodactylum tricornutum*

The work presented in this chapter is based on a manuscript conditionally accepted for publication at PeerJ. No permission is required for reproducing, re-mixing, or redistributing under PeerJ's standard CC BY 4.0 license.

4.1 Introduction

Phaeodactylum tricornutum is a marine diatom that is described as a "diatom cell factory" (1) because it can be used to manufacture valuable commercial products. Recent genetic toolbox expansions, such as delivering episomes by bacterial conjugation (2), CRISPR-editing tools (3–8), the generation of auxotrophic strains (9–11), and the identification of highly active endogenous promoters (12) are enabling rapid implementation of new product designs into commercial-scale production.

The genome of *P. tricornutum* CCAP 1055/1 was sequenced in 2008, and resulted in a scaffold-level assembly with 33 scaffolds (NCBI assembly ASM15095v2) (13), with the exact number of chromosomes unknown. Chloroplast and mitochondrial genomes have also been published (14, 15), and have previously been identified as targets for genetic engineering (16), as well as other chromosomes (17). Although the Bowler *et al.* assembly contains several telomere-to-telomere chromosomes, many scaffolds have only zero or one telomere, suggesting they are either incomplete or fragments of another chromosome. More recent work identifying centromeric sequences (18) in *P. tricornutum* has suggested that there may be less than 33 chromosomes, and the authors were only able to identify 25 unique centromeric DNA sequences.

While the current assembly is an excellent resource, it does not represent a completed genome assembly. The lack of a completed genome assembly for *P. tricornutum* means that synthetic biology researchers are unable to pursue generating artificial chromosomes with this model diatom, since the full sequence of each chromosome is required to rebuild them by DNA synthesis. It is also important to know the location and sequence of mobile genetic elements that could be removed to in order to simplify a potential fully synthesized chromosome sequence. A more complete understanding of the genome will be a resource to help researchers answer more fundamental biological questions about *P. tricornutum*.

To generate a telomere-to-telomere assembly of *P. tricornutum* CCAP 1055/1, we used a hybrid approach with ultra-long reads from the Oxford Nanopore MinION platform and highly accurate short reads from the Illumina NextSeq platform. We also introduce a novel graph-based approach to manually resolve telomere-related assembly errors. This approach identifies all unique telomere sequences and we demonstrate how it can be applied to manually correct assembly errors adjacent to chromosome ends. The full structural context of the *P. tricornutum* genome provides additional information for potential synthetic biology applications to manipulate the genome of this diatom cell factory.

4.2 Methods

4.2.1 Growth

Phaeodactylum tricornutum (Culture Collection of Algae and Protozoa CCAP 1055/1) was grown in L1 medium without silica at 18° C under cool white fluorescent lights (75 mE m^{-2} s^{-1}) and a photoperiod of 16 h light:8 h dark as described previously (7).

4.2.2 DNA extraction

200 mL of culture (approximately 5 x 10^8 cells) was spun at 3000 X g for 10 minutes at 4° C. The pellet was resuspended in 1 mL TE (pH 8.0) and added dropwise to a mortar (pre-cooled at -80° C) pre-filled with liquid nitrogen. The frozen droplets were ground into a fine powder with a mortar and pestle, being careful to keep the cells from thawing by adding more liquid nitrogen as necessary. The frozen powder was transferred to a 15 mL Falcon tube where 2 mL of lysis buffer was added (1.4 M NaCl, 200 mM Tris-HCl pH 8.0, 50 mM EDTA, 2% (w/v) CTAB, RNAse A (250 μg/mL) and proteinase K (100 μg/mL)). The solution was mixed very slowly by inversion, incubated for 30 minutes at 37° C (mixed very slowly halfway through incubation). Cellular debris was pelleted at 6000 X g for 5 minutes. Lysate was transferred to a new 15 mL Falcon tube. One volume of 25:24:1 phenol:chloroform:isoamyl alcohol was added, mixing slowly by inversion. The sample was centrifuged at 6000 X g for 5 minutes. The aqueous phase was transferred as slow as possible to a new Falcon tube. One volume of 24:1 chloroform:isoamyl alcohol was added, and mixed slowly with end-over-end inversion. The sample was centrifuged at 6000 X g for 5 minutes. Approximately 450 uL of the aqueous phase was transferred into new 1.5 mL Eppendorf tubes. To the Eppendorf tube, 1/10 volume of 3 M NaAc pH 5.2 and 2 volumes (final volume) of ice-cold 100% ethanol were added, mixing slowly by end-over-end inversion. The sample was centrifuged at 16 000 X g for 5 minutes, and washed twice with 500 uL 70% ethanol. Ethanol was decanted, and the pellet was dried for approximately 10 minutes by inverting on a paper towel. The pellet was resuspended in 100 uL 10 mM Tris-HCl pH 8.0, 0.1 mM EDTA pH 8.0. After resuspending overnight at 4° C, DNA fragments less than 20 kbp were then selectively removed using the Short Read Eliminator (SRE) kit from Circulomics (Baltimore). DNA from the same extraction was used for sequencing on both the Oxford Nanopore MinION and Illumina NextSeq 550 platform.

4.2.3 Sequencing

An Oxford Nanopore MinION flow cell R9.4.1 was used with the SQK-LSK109 Kit according to the manufacturer's protocol version GDE_9063_v109_revK_14Aug2019, with one alteration: for DNA repair and end-prep, the reaction mixture was incubated for 15 minutes at 20° C and 15 minutes at 65° C. Basecalling was performed after the run with Guppy (Version 3.6). NanoPlot (19) was used to generate Q-score versus length plots and summary statistics. The read N50 of the unfiltered reads was approximately 35 kb. For Illumina sequencing, the Nextera XT kit was used, and a 2X75 paired-end mid-output NextSeq 550 library was prepared according to the manufacturer's protocol, and run at the London Regional Genomics Center (lrgc.ca). Reads were trimmed using Trimmomatic v0.36 (20) in paired end mode with the following settings: AVGQUAL:30 CROP:75 SLIDINGWINDOW:4:25 MINLEN:50 TRAILING:15. SLIDINGWINDOW AND TRAILING were added to remove poor quality base calls. Raw sequencing signal and basedcalled reads are available on the European Nucleotide Archive project number PRJEB42700.

4.2.4 Assembly

4.2.5 Telomere identification

We first obtained sequences for the end of every linear chromosome. The sequence of the telomere repeats for *P. tricornutum* are known from the previous assembly (13) to be repeats of AACCCT. All long reads larger than 50 kilobases with 3 or more consecutive telomeric repeats (or the reverse complement) were extracted by filtering using NanoFilt (19) and by string matching using grep. All-versus-all mapping of the telomeric reads was performed using minimap2 (21). Only overlapping reads with a minimum query coverage of 95 % were retained.

To determine the sequence of unique telomeres for each chromosome, a network graph was generated with iGraph (22). Each read name was used as a vertex, and edges were generated between each overlapping read with more than 95% query coverage. Noise was filtered by removing any group of overlaps with less than 5X coverage. There were 93 vertices that had greater than

20X coverage; that is, there are 93 unique telomere sequence groups. Most groups had approximately 40X coverage (number of long reads per group), however, several outliers had more than 60X coverage. These represent duplicated regions in the telomeres that are not unique (i.e., more than one haplotype or chromosome contains this sequence). The longest read of each telomere group, typically greater than 100 kb in length, was retained as a representative telomere sequence for correction.

4.2.6 Assembly

Miniasm (21) was chosen for assembly to reduce computational power needed compared to other assemblers like Canu (23) or Flye (24). Nanopore reads longer than 75 kilobases were used for initial assembly with miniasm, using the parameters -s 30000 -m 10000 -c 5 -d 100000. From this initial assembly, the output from miniasm were manually completed with the following approach:

1) Mapping of telomeric reads against the unitig (high-confidence contig). If no telomere was present on the unitig and a high query coverage alignment was found, the unitig was extended to the telomere sequence of the mapped telomere. 2) After telomere extension (or confirmation), reads longer than 50 kb were mapped to the unitig to confirm overlapping coverage over the entire chromosome. Coverage was evaluated using only reads larger than 50 kb and with higher than 50% query coverage, with an alignment score:length ratio less than 2 (similar to previous validation methods)(25). A query coverage of only 50% was chosen to allow for potential haplotype divergence. 3) Telomere-to-telomere unitigs with overlapping ultra-long read coverage and no gaps were deemed validated and brought forward to improve base accuracy by read polishing.

The chloroplast and mitochondrial genomes were assembled using a reference based approach by first extracting all reads that aligned to the reference chloroplast and mitochondria with high query coverage. Reads were then *de-novo* assembled using miniasm.

4.2.7 Polishing

The assembly was polished using 4 rounds of Racon (26), two rounds of medaka (Oxford Nanopore) and two rounds of Pilon (27). Sequencing coverage was visualized after polishing to determine if large scale errors were introduced into any of the chromosomes, and manual corrections were made when sequencing coverage dropped to zero. For the chloroplast and mitochondria, only the subset of reads identified as either chloroplast or mitochondria were used for polishing.

4.2.8 Methylation

5mC methylation sites were predicted using Megalodon v2.2.1 (Oxford Nanopore Technologies) using the model res_dna_r941_min_modbases_5mC_CpG_v001.cfg from the Rerio repository (Oxford Nanopore Technologies) with Guppy 4.5.2. A default threshold of 0.75 was used as a minimum score for modified base aggregation (probability of modified/canonical base) to produce the final aggregated output.

4.3 Results

4.3.1 Workflow

We developed a sample preparation protocol that provided high-molecular weight DNA. We observed a read N50 of 35 kilobases, with the longest reads just over 300 kb. Of the 7.8 gigabases of raw sequence data, approximately 2.5 gigabases were from reads longer than 50 kilobases. We found that chromosomes assembled using standard approaches were often mis-assembled around telomeres, or were fragmented and only contained 1 telomere. To correct each contig, we used the unique ultra-long telomere reads as described in the methods and in Figure 1. This approach was used to manually identify a tiling path for each chromosome until each chromosome was contiguous from telomere to telomere, and validated by a tiling overlapping read path.

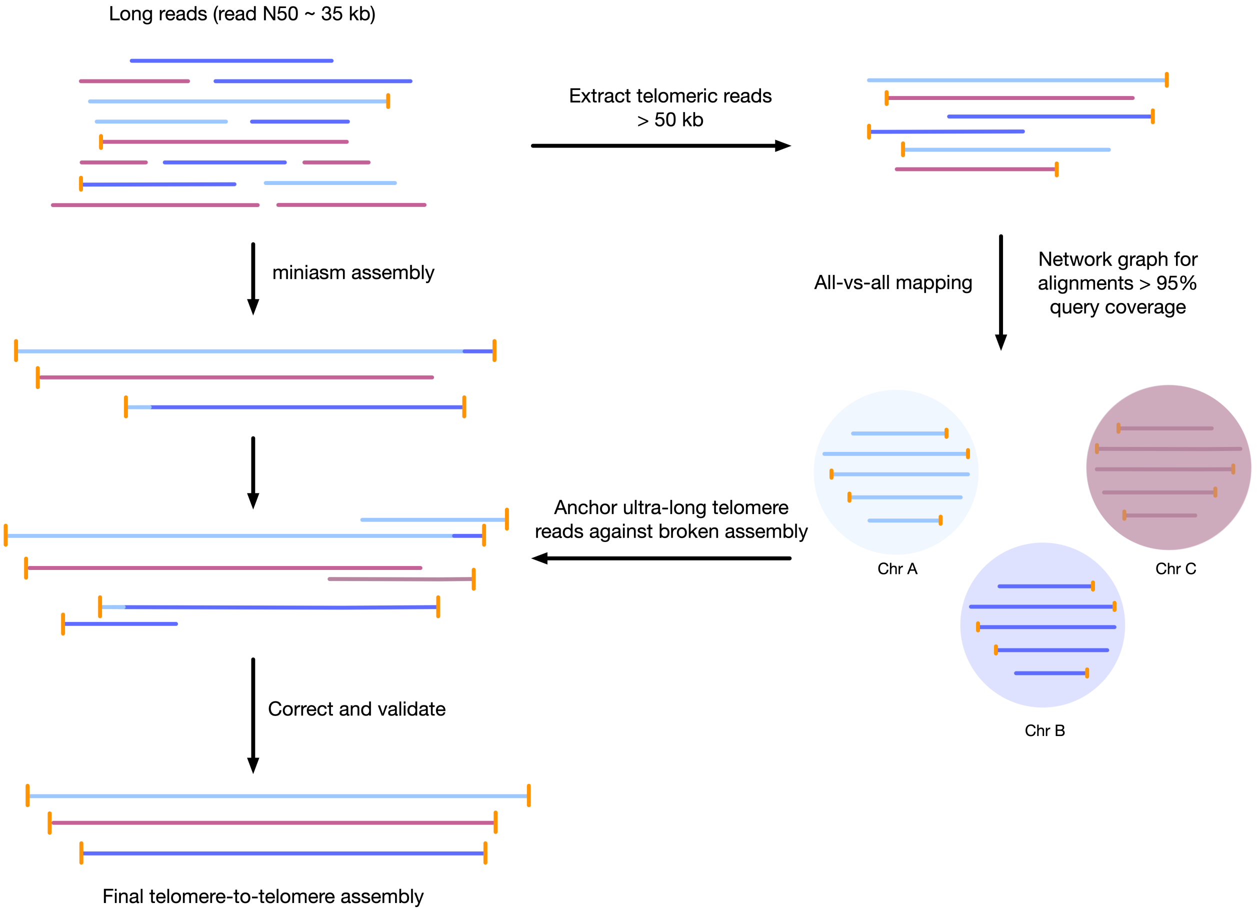


Figure 4.1: **Workflow for telomere-to-telomere genome assembly**. Telomere-containing nanopore reads larger than 50 kb are extracted and mapped in all-vs-all mode using minimap2. The resulting alignments are filtered by 95% query coverage, and a network graph is created using iGraph using read names as vertices, and alignments between reads as edges. Each resulting cluster represents one end of a chromosome. On a chromosome-by-chromosome basis, ultra-long read coverage is plotted. If an assembled chromosome is missing a telomere or has an assembly error revealed by a lack of overlapping read coverage, the longest read from each telomere cluster is mapped against the chromosome, and the resulting telomere is used to manually correct the assembly and extend to the telomere using an overlap-layout consensus approach.

4.3.2 Tiling path of overlapping reads verify contiguity

To ensure our genome assembly is contiguous, we generated multiple independent minimum tiling paths of overlapping long reads (one such path is shown in Supplementary Table B.1). Reads larger than 50 kb were mapped against the assembly using minimap2. To ensure no incorrect alignments were retained, any reads with less than 90% of the read aligned to the assembly were removed. From this subset, 5 independent minimum tiling paths that required at least 10 kb of overlap between each read were generated. All chromosomes have multiple independent (i.e., no common reads) tiling paths of reads with a minimum overlap of 10 kb in the final assembly, indicating that all chromosomes are contiguous.

In addition to overlapping reads, Figure 4.3 and Supplementary Figure B.1 also show the GC content for each chromosome. A previous study has proposed that centromeres could be identified by low GC content calculated in 100 bp windows (18). The 100 base window(s) with the minimum GC content are shown in these figures, highlighted in red. These windows represent putative centromere sequences as previously described (18).

4.3.3 Telomere-to-telomere assembly comprises previous scaffolds

We ultimately obtained 25 telomere-to-telomere chromosome assemblies that recruit 98% of long reads, and these chromosomes comprise all previously proposed chromosomes from Bowler *et al.* (2008), as well as circularized chloroplast and mitochondrial genomes. The median coverage for unfiltered long reads across the nuclear genome was 202X, while median coverage for the chloroplast and mitochondrion were approximately 6201X and 528X, respectively. This was calculated in 1000 base windows using mosdepth (28).

A key feature of this updated assembly is the consistency with previous sequencing efforts (13). Previously, 25 centromere sequences were identified (18), suggesting that there were fewer than the proposed 33 chromosomes. This agrees with our conclusion of 25 nuclear chromosomes. We independently resolved the location of all the previously proposed partial chromosomes without internal inconsistencies in Figure 2 (i.e., scaffolds with only 1 telomere were resolved into a telomere-to-telomere chromosome).

4.3.4 Estimating number of chromosomes using ultra-long reads

Previous studies have suggested that *P. tricornutum* has a minimum of 33 chromosomes using pulsed-field gel electrophoresis (29). Our orthogonal, reference-free method using network graphs of telomere-containing overlapping ultra-long reads revealed 25 chromosomes.

We used 2 properties of telomeres for this: first, telomeres on linear chromosomes can be identified by unique subtelomeric sequences, and second, that telomere-containing DNA fragments will begin or end with a telomere, representing the start or end of a chromosome. After aligning all telomere-containing reads and retaining only alignments with greater than 95% query coverage, we used iGraph to create network graphs, which resulted in two classes of independent graphs. The first class had 85 independent graphs, each with approximately 50 nodes (i.e., 50 ultra-long reads in each graph), and the second class had 8 graphs with approximately 100 nodes (Supplemental Figure B.2). In a diploid organism we expect 4 telomeres per chromosome if we assume that each chromosome has two haplotypes; i.e., (maternal + paternal) x haplotypes. Under this assumption, 85 independent graphs with approximately 50 nodes represents 21.25 telomeres. Some chromosomes will not have diverged sufficiently, meaning there will be only two telomeres with twice the sequencing coverage per chromosome (maternal + paternal). The remaining 8 graphs 100 nodes each therefore gives a further 4 chromosomes.

With this logic we estimate 25.25 chromosomes exist in *P. tricornutum*, which agrees very closely with our final assembly of 25 chromosomes. The additional 0.25 chromosome may be explained by mitotic recombination (30). Using the features of ultra-long reads at the ends of linear DNA elements (i.e., eukaryotic chromosomes) thus enables an orthogonal method for estimating the number of chromosomes in a reference-free manner.

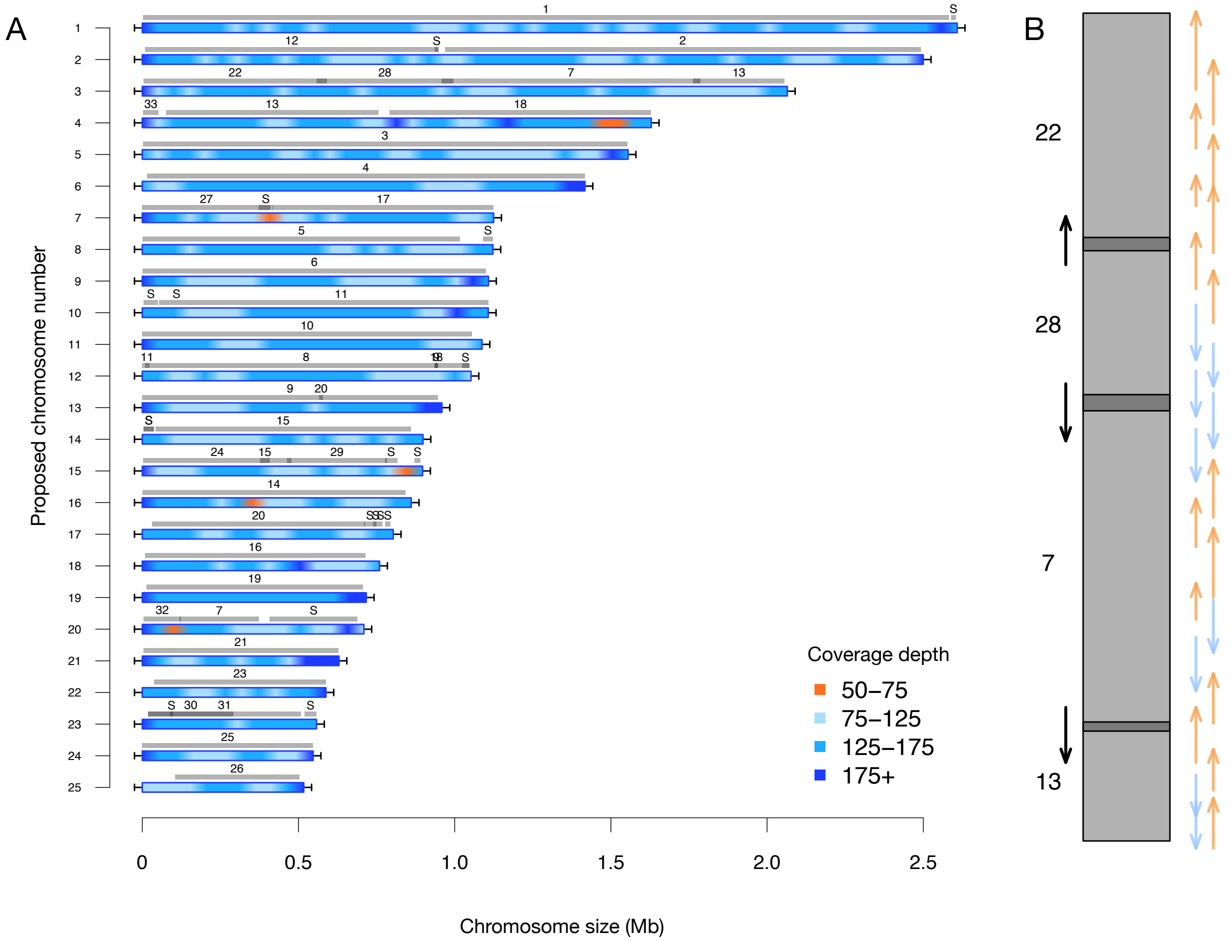


Figure 4.2: **Sequencing coverage and comparison to previous assembly**. A) Filtered long-read coverage and comparison to previous assembly. Reads longer than 20 kb were mapped against the assembly, filtered (minimum 20000 base alignment and 50 % query coverage), and genome coverage was calculated in 50 kb windows using mosdepth. The colours and ranges bottom-right) describe the coverage depth calculate for each 50 kb window. Newly proposed chromosomes names are indicated on the left (by length). Scaffolds from the previous genome assembly (ASM15095v2) are overlayed as grey bars, aligned using minimap2 in asm5 mode and filtered to retain minimum 10 kb alignments. Numbers on top of gray bars indicate which previous scaffold number, with S representing small “bottom drawer” scaffolds. Horizontal “T” bars on each end indicate telomere-repeat presence. B) Visualization of proposed chromosome 3 with alignments to previous chromosomes. Dark gray regions indicate overlap. Coloured arrows on the right indicate minimum overlapping read path (orange = negative strand, blue = positive strand), black arrows on left show ultra-long reads that completely span regions where previous assembly could not assemble through.

4.3.5 Assembly quality

To assess the quality of the assembly, we used Merqury (31) to estimate the base-level accuracy and completeness by k-mer frequency, shown in Supplemental Table B.2. We found that the estimated quality value (estimated log-scaled probability of error for the consensus base calls by Merqury) ranged from 27 - 53, depending on the chromosome. The mean quality value (QV) for nuclear chromosomes was 28.86, with chromosome 19 as an outlier at 43. The QV for all nuclear genomes except for 19 are likely lower because the chromosomes were polished using reads that are heterozygous. The chloroplast and mitochondrial genomes have a QV of 53 and 42, respectively. Importantly, the k-mer completeness estimate of 80% suggests that many k-mers in the Illumina reads are not represented in this genome assembly, implying significant haplotype variation. This was also the case when using the Bowler assembly as input for Merqury.

We also estimated the genome completion using a software package called Benchmarking Universal Single-Copy Orthologs (BUSCO) (32). Using the stramenopiles_odb10 model, we found our assembly was 95% complete, with only 3% of expected BUSCOs missing. When evaluating the chromosome scaffolds of the Bowler assembly, we found it was 96% complete with 3% of expected BUSCOs missing.

After removing Lambda spike-in reads with NanoLyse, we found that 98.12% of long reads are recruited by the assembly. When reads are filtered by removing any read that does not align over more than 90% of it's length (i.e., query coverage is higher than 90%), the number of reads recruited drops to 74%.

4.3.6 Filtered long-read coverage for Chromosome 19 is inconsistent with diploid state

We observed that chromosome 19 has remarkably consistent (i.e., no drops in coverage) filtered long-read coverage relative to the other chromosomes (Figure 4.3, Supplemental Figure B.1). While we initially predicted *P. tricornutum* would have two haplotypes since it is diploid, recent work has demonstrated that while each cell has two haplotypes, many haplotypes within a population arise due to mitotic recombination (30). The consistency of filtered long read coverage for chromosome 19 indicates that there is only a single haplotype, whereas the other chromosomes have 2 or more haplotypes present, which can be inferred from inconsistent read depth at regions where haplotype divergences occur in Figure 4.3 and Supplemental Figure B.1. This indicates that there are not two haplotypes for chromosome 19, suggesting a different recent history for this chromosome.

4.3.7 5mC methylation and transposable elements

The Extensive de-novo TE Annotator (EDTA) pipeline (33) was used to predict transposable elements in the genome. We found that the majority of transposable elements are long terminal repeat (LTR) retrotransposons (3.43% of the genome was found to be Copia-type, 5.86% were unknown, while terminal inverted repeats were only 1.17% of the genome, and helitrons were 0.54% of the genome). Each LTR region is represented as a shaded blue region in Figure 4.3 in blue, and density plots of the end locations are shown in the top quadrant. Chromosome 19 contained the fewest transposable elements at 50. The locations and density of LTR-retrotransposons are plotted in Figure 4.3 for proposed Chromosome 3 and Supplemental Figure B.1 for chromosome 19.

Previous studies have found that some tranposable elements were hypermethylated (34). Using chromosome scale nanopore methylation basecalling, we found a strong signal between many predicted LTR retrotransposons and methylation status (Figure 4.3, Supplemental Figure B.1). To test this, we enumerated all chromosome positions with methylated sites and transposons, and performed a Fisher's Exact Test, resulting in a p-value of 2.2e-16.

We examined the association between LTR transposon dense regions and regions where the previous assembly failed to generate overlapping regions. We observed that scaffolds with overlapping regions generally were not assembled into full chromosomes because of ambiguity in the placement of the LTR-rich regions at the ends of the scaffolds. These are now resolved by the long-read assembly identified here. Additionally, many of the low-coverage regions of our assembly overlap with the locations of the LTR-dense regions, consistent with chromosomal rearrangements being more likely in these regions. Further investigation at these regions is required.

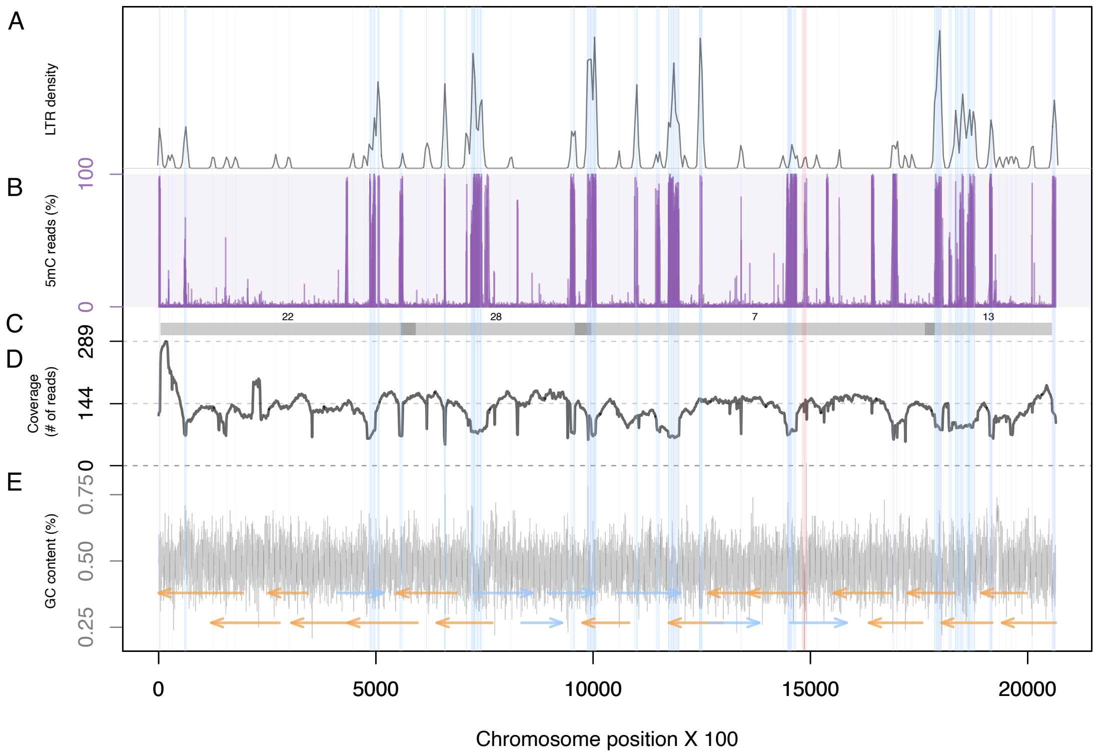


Figure 4.3: **Summary of genomic features for chromosome 3.** A) The density of LTR-retrotransposons as predicted by the EDTA pipeline. B) The proportion of reads that were called as methylated at each position along the chromosome. C) Scaffolds from the previous assembly are overlayed in gray bars, with dark grey representing overlapping regions. D) Filtered long-read coverage (minimum 20 kb length and 70% query coverage). E) GC content calculated and plotted in 100 base windows. An overlapping read tiling path, with a minimum overlap of 30 kb, is shown with orange indicating reads mapping to the negative strand and blue indicating reads mapping to the positive strand. The regions that are annotated at LTR-retrotransposons are highlighted in light blue.

4.4 Discussion

Here, we developed a graph-based approach to locate the unique telomere ends of all *P. tricornutum* chromosomes, and applied this information to generate a telomere-to-telomere assembly. The new assembly incorporates all chromosome fragments from the previous reference genome (13).

The chromosomes show marked variations in sequencing coverage that can be explained by haplotype variation. Where haplotype variation occurs, filtered long reads will not align against the assembly. This suggests that there are large regions of the chromosomes that have substantial haplotype differences. Strikingly, only chromosome 19 has completely consistent coverage between the telomeres. While this needs to be further investigated, we speculate that this chromosome in this strain may have undergone a recent sequence homogenization event. Previous work has also found that the same chromosome appears homozygous in the wild type strain (3, 30). It has previously been speculated that *Phaeodactylum tricornutum* may be capable of sexual reproduction (35, 36), but there has yet to be conclusive evidence of this occurring.

Chromosome 19 has a high quality value of 43, while the other nuclear chromosomes have lower quality values around 28. For all chromosomes except 19, this drop in per-base quality is due to polishing the nanopore assembly with a heterozygous read set. However, the high quality value and consistent filtered-long read coverage suggest that there are not highly divergent haplotypes for chromosome 19. Recently published data has demonstrated that mitotic recombination occurs frequently in *P. tricornutum* (30). They independently showed that there is a significantly lower SNP density on chromosome 19, agreeing with this finding (3). Interestingly, the high rate of mitotic recombination suggests that it is unlikely that a static haplotype-resolved diploid genome may be fully resolved for this species with the currently available technology. In this context, the k-mer completeness estimate we obtained from Merqury suggests that up to 20% of the Illumina k-mers result from SNPs arising from mitotic recombination events within the population, suggesting a high degree of haplotype divergence.

We demonstrate that nanopore sequencing can identify methylated regions, and the entire methylome of *P. tricornutum* is strongly associated with transposable elements (Figure 4.3, Supplemental Figure B.1). This agrees with previous work (34) that found a significant enrichment of DNA methylation at LTR retrotransposons, and we provide an updated map by predicting methylation sites directly from sequenced native DNA.

We have deposited all short and raw long-read data publicly for use by the community as Project PRJEB42700 on the European Nucleotide Archive. This telomere-to-telomere genome assembly will be a resource for designing and creating synthetic chromosomes in *Phaeodactylum tricornutum*, as well as answering fundamental biological questions for this species.

4.5 Conclusions

Here, we report a collapsed telomere-to-telomere genome assembly for *Phaeodactylum tricornutum* CCAP 1055/1. A combination of ultra-long nanopore sequencing reads (greater than 100 kb), a novel approach to correcting assembly errors near telomeres, and manual curation enabled the completion of a telomere-to-telomere genome. We also describe a method to estimate the number of chromosomes using the properties of ultra-long telomere-containing reads in a reference-free manner. We provide the signal level nanopore data as a resource to enable the community to further investigate 5mC methylation for this species. This work improves upon our current understanding of the model diatom *Phaeodactylum tricornutum* to enable further developments in synthetic biology.

4.6 Funding

DJG Mitacs number IT8360, Ontario Graduate Scholarship

SS: Mitacs number IT8360

GBG: Natural Sciences and Engineering Research Council of Canada (NSERC), grant number: RGPIN-03878-2015

BK: Natural Sciences and Engineering Research Council of Canada (NSERC), grant number: RGPIN-2018-06172

DE: Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant RPGIN-2015-04800

4.7 References

Bibliography

[1] Butler, T.; Kapoore, R. V.; Vaidyanathan, S. Phaeodactylum Tricornutum: A Diatom Cell Factory. *Trends in Biotechnology* **2020**, *38*, 606–622.

[2] Karas, B. J. et al. Designer Diatom Episomes Delivered by Bacterial Conjugation. *Nature Communications* **2015**, *6*, 6925.

[3] Russo, M. T.; Cigliano, R. A.; Sanseverino, W.; Ferrante, M. I. Assessment of Genomic Changes in a CRISPR/Cas9 Phaeodactylum Tricornutum Mutant through Whole Genome Resequencing. *PeerJ* **2018**, *6*, e5507.

[4] Moosburner, M. A.; Gholami, P.; McCarthy, J. K.; Tan, M.; Bielinski, V. A.; Allen, A. E. Multiplexed Knockouts in the Model Diatom Phaeodactylum by Episomal Delivery of a Selectable Cas9. *Frontiers in Microbiology* **2020**, *11*, 5.

[5] Sharma, A. K.; Nymark, M.; Sparstad, T.; Bones, A. M.; Winge, P. Transgene-Free Genome Editing in Marine Algae by Bacterial Conjugation – Comparison with Biolistic CRISPR/Cas9 Transformation. *Scientific Reports* **2018**, *8*, 14401.

[6] Stukenberg, D.; Zauner, S.; Dell'Aquila, G.; Maier, U. G. Optimizing CRISPR/Cas9 for the Diatom Phaeodactylum Tricornutum. *Frontiers in Plant Science* **2018**, *9*, 740.

[7] Slattery, S. S.; Diamond, A.; Wang, H.; Therrien, J. A.; Lant, J. T.; Jazey, T.; Lee, K.; Klassen, Z.; Desgagné-Penix, I.; Karas, B. J.; Edgell, D. R. An Expanded Plasmid-Based

Genetic Toolbox Enables Cas9 Genome Editing and Stable Maintenance of Synthetic Pathways in Phaeodactylum Tricornutum. *ACS Synthetic Biology* **2018**, *7*, 328–338.

[8] Serif, M.; Dubois, G.; Finoux, A.-L.; Teste, M.-A.; Jallet, D.; Daboussi, F. One-Step Generation of Multiple Gene Knock-Outs in the Diatom Phaeodactylum Tricornutum by DNA-Free Genome Editing. *Nature Communications* **2018**, *9*, 3924.

[9] Zaslavskaia, L. A.; Lippmeier, J. C.; Kroth, P. G.; Grossman, A. R.; Apt, K. E. Transformation of the Diatom Phaeodactylum Tricornutum (Bacillariophyceae) with a Variety of Selectable Marker and Reporter Genes. *Journal of Phycology* **2000**, *36*, 379–386.

[10] Buck, J. M.; Bártulos, C. R.; Gruber, A.; Kroth, P. G. Blasticidin-S Deaminase, a New Selection Marker for Genetic Transformation of the Diatom Phaeodactylum Tricornutum. *PeerJ* **2018**, *6*, e5884.

[11] Slattery, S. S.; Wang, H.; Giguere, D. J.; Kocsis, C.; Urquhart, B. L.; Karas, B. J.; Edgell, D. R. Plasmid-Based Complementation of Large Deletions in Phaeodactylum Tricornutum Biosynthetic Genes Generated by Cas9 Editing. *Scientific Reports* **2020**, *10*, 13879.

[12] Erdene-Ochir, E.; Shin, B.-K.; Kwon, B.; Jung, C.; Pan, C.-H. Identification and Characterisation of the Novel Endogenous Promoter HASP1 and Its Signal Peptide from Phaeodactylum Tricornutum. *Scientific Reports* **2019**, *9*, 9941.

[13] Bowler, C. et al. The Phaeodactylum Genome Reveals the Evolutionary History of Diatom Genomes. *Nature* **2008**, *456*, 239–244.

[14] Oudot-Le Secq, M.-P.; Grimwood, J.; Shapiro, H.; Armbrust, E. V.; Bowler, C.; Green, B. R. Chloroplast Genomes of the Diatoms Phaeodactylum Tricornutum and Thalassiosira Pseudonana: Comparison with Other Plastid Genomes of the Red Lineage. *Molecular Genetics and Genomics* **2007**, *277*, 427–439.

[15] Oudot-Le Secq, M.-P.; Green, B. R. Complex Repeat Structures and Novel Features in the Mitochondrial Genomes of the Diatoms Phaeodactylum Tricornutum and Thalassiosira Pseudonana. *Gene* **2011**, *476*, 20–26.

[16] Cochrane, R. R.; Brumwell, S. L.; Soltysiak, M. P. M.; Hamadache, S.; Davis, J. G.; Wang, J.; Tholl, S. Q.; Janakirama, P.; Edgell, D. R.; Karas, B. J. Rapid Method for Generating Designer Algal Mitochondrial Genomes. *Algal Research* **2020**, *50*, 102014.

[17] Karas, B. J. et al. Assembly of Eukaryotic Algal Chromosomes in Yeast. *Journal of Biological Engineering* **2013**, *7*, 30.

[18] Diner, R. E.; Noddings, C. M.; Lian, N. C.; Kang, A. K.; McQuaid, J. B.; Jablanovic, J.; Espinoza, J. L.; Nguyen, N. A.; Anzelmatti, M. A.; Jansson, J.; Bielinski, V. A.; Karas, B. J.; Dupont, C. L.; Allen, A. E.; Weyman, P. D. Diatom Centromeres Suggest a Mechanism for Nuclear DNA Acquisition. *Proceedings of the National Academy of Sciences* **2017**, *114*, E6015–E6024.

[19] De Coster, W.; D'Hert, S.; Schultz, D. T.; Cruts, M.; Van Broeckhoven, C. NanoPack: Visualizing and Processing Long-Read Sequencing Data. *Bioinformatics* **2018**, *34*, 2666–2669.

[20] Bolger, A. M.; Lohse, M.; Usadel, B. Trimmomatic: A Flexible Trimmer for Illumina Sequence Data. *Bioinformatics* **2014**, *30*, 2114–2120.

[21] Li, H. Minimap and Miniasm: Fast Mapping and de Novo Assembly for Noisy Long Sequences. *Bioinformatics* **2016**, *32*, 2103–2110.

[22] Csárdi, G.; Nepusz, T. The Igraph Software Package for Complex Network Research. *undefined* **2006**,

[23] Koren, S.; Walenz, B. P.; Berlin, K.; Miller, J. R.; Bergman, N. H.; Phillippy, A. M. Canu: Scalable and Accurate Long-Read Assembly via Adaptive k-Mer Weighting and Repeat Separation. *Genome Research* **2017**, *27*, 722–736.

[24] Kolmogorov, M.; Yuan, J.; Lin, Y.; Pevzner, P. A. Assembly of Long, Error-Prone Reads Using Repeat Graphs. *Nature Biotechnology* **2019**, *37*, 540–546.

[25] 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* **2020**, 2020.04.08.032540.

[26] Vaser, R.; Sović, I.; Nagarajan, N.; Šikić, M. Fast and Accurate de Novo Genome Assembly from Long Uncorrected Reads. *Genome Research* **2017**, *27*, 737–746.

[27] Walker, B. J.; Abeel, T.; Shea, T.; Priest, M.; Abouelliel, A.; Sakthikumar, S.; Cuomo, C. A.; Zeng, Q.; Wortman, J.; Young, S. K.; Earl, A. M. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PloS One* **2014**, *9*, e112963.

[28] Pedersen, B. S.; Quinlan, A. R. Mosdepth: Quick Coverage Calculation for Genomes and Exomes. *Bioinformatics* **2018**, *34*, 867–868.

[29] Filloramo, G. V.; Curtis, B. A.; Blanche, E.; Archibald, J. M. Re-Examination of Two Diatom Reference Genomes Using Long-Read Sequencing. *BMC Genomics* **2021**, *22*, 379.

[30] Bulankova, P. et al. Mitotic Recombination between Homologous Chromosomes Drives Genomic Diversity in Diatoms. *Current Biology* **2021**, *31*, 3221–3232.e9.

[31] Rhie, A.; Walenz, B. P.; Koren, S.; Phillippy, A. M. Merqury: Reference-Free Quality, Completeness, and Phasing Assessment for Genome Assemblies. *Genome Biology* **2020**, *21*, 245.

[32] Simão, F. A.; Waterhouse, R. M.; Ioannidis, P.; Kriventseva, E. V.; Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs. *Bioinformatics (Oxford, England)* **2015**, *31*, 3210–3212.

[33] Ou, S.; Su, W.; Liao, Y.; Chougule, K.; Agda, J. R. A.; Hellinga, A. J.; Lugo, C. S. B.; Elliott, T. A.; Ware, D.; Peterson, T.; Jiang, N.; Hirsch, C. N.; Hufford, M. B. Benchmarking

Transposable Element Annotation Methods for Creation of a Streamlined, Comprehensive Pipeline. *Genome Biology* **2019**, *20*, 275.

[34] Veluchamy, A. et al. Insights into the Role of DNA Methylation in Diatoms by Genome-Wide Profiling in Phaeodactylum Tricornutum. *Nature Communications* **2013**, *4*, 2091.

[35] Mao, Y.; Guo, L.; Luo, Y.; Tang, Z.; Li, W.; Dong, W. Sexual Reproduction Potential Implied by Functional Analysis of SPO11 in Phaeodactylum Tricornutum. *Gene* **2020**, *757*, 144929.

[36] Patil, S.; Moeys, S.; von Dassow, P.; Huysman, M. J. J.; Mapleson, D.; De Veylder, L.; Sanges, R.; Vyverman, W.; Montresor, M.; Ferrante, M. I. Identification of the Meiotic Toolkit in Diatoms and Exploration of Meiosis-Specific SPO11 and RAD51 Homologs in the Sexual Species Pseudo-Nitzschia Multistriata and Seminavis Robusta. *BMC Genomics* **2015**, *16*, 930.

Chapter 5

Complete genome sequence of new species from the family *Saccharimonadaceae*, *UBA6175* sp., reveals acquisition of a genomic island

5.1 Introduction

The family *Saccharimonadaceae* are small, epibiotic bacteria with reduced genomes around 0.7-1 megabases long, that have been recently characterized in depth thanks to DNA sequencing improvements since they are currently unculturable as isolates (1, 2). Non-targeted high-throughput metagenomic sequencing and novel metagenome assembly algorithms have enabled partial and complete genome assembly of some of these enigmatic bacteria (2), which belong to the phylum Patescibacteria, recently resulting in a new proposal for the tree of life (3). This phylum contains unusual genomes from mostly uncultured bacteria (3). These bacteria have redefined what is considered the minimum set of genes needed for bacterial life (4), since bacteria within the Patescibacteria often lack biosynthetic pathways for core functions such as glycolysis or amino acid production. *Saccharimonadaceae* are typically found as ultra-small 200 - 300 nm coccus-shaped cells (5).

Many have these bacteria have no known cultured isolates since they are obligate epibionts; however, very recent advances have enabled the co-culturing of human-associated *Saccharimonadaceae* (5–9). From these studies, their hosts have only been found in the phylum Actinobacteriota, which are anaerobic gram-positive bacteria. Several species such as *Arachnia propionica* and *Actinomyces* sp.(7), *Gordonia amarae* (8) have been successfully used to co-culture *Saccharimonadaceae* derived from dental plaques. More recently, it has been suggested that some species from the *Saccharimonadaceae* family may play a protective role in dental caries - it was found that multiple isolates from periodontitis patients lead to reduced inflammatory bone disease by modulating their Actinobacteriota host's behaviour (10).

Patescibacteria have also been found in naturally occurring environments (2), and have been associated with hydrocarbon degradation (11, 12), although this is disputed by more recent evidence suggesting that these bacteria utilize the by-products of hydrocarbon degradation as building blocks (13). What role these bacteria may play in a hydrocarbon degrading bacterial community thus remains to be determined.

The number of complete genomes from the family *Saccharimonadaceae* has recently enabled phylogenomic analysis based on concatenated ribosomal protein sequences, and has been subdivided into 6 class-level clades (14, 15), labelled G1 through G6. Since these bacteria have mostly been discovered by high-throughput short-read metagenome sequencing, many of the genomes are contig-level assemblies, and could be missing sequence due to limitations of short read sequencing and metagenomic binning. Only 40 of the available 732 entries on NCBI's assembly database are completed, the rest being characterized as scaffold or contig-level assemblies (accessed September 13, 2021). While many of the available genomes are derived from the oral microbiome, there are several representative genomes assembled from environmental samples, including activate sludge samples, and aquifer sediments (2).

Many fundamental questions are still unanswered about these unusual bacteria, including the role they play in their communities, how they may contribute to human health in the oral microbiome, and if they contribute to hydrocarbon degradation. It is therefore critical to continue researching not just the family *Saccharimonadaceae*, but also many other bacteria from the phylum Patescibacteria. New long read sequencing technologies are progressing this research area by enabling complete genomes to be assembled directly from metagenomes, furthering this field of research (16–19).

Here, we build and investigate the complete genome of a novel species from the family *Saccharimonadaceae*, *UBA6175* sp., that was assembled using long read sequencing of DNA isolated directly from the metagenome of an algal-bacterial community known to degrade 1-adamantanecarboxylic acid, a surrogate for toxic naphthenic acids that are produced as a by-product during oil refining (20).

5.2 Methods

5.2.1 DNA extraction

An algal-bacterial co-culture initially enriched in 2012 from a northern Alberta oil sands tailings pond was used for this study. The culture had been routinely propagated since that time. For this study, a 5 mL aliquot was grown in Bold's Basal media, pelleted, and stored at -80° C until DNA was extracted (20).

DNA extraction was performed to maximize read length by preventing shearing (performed with wide-bore pipette tips, slow pipetting, mixing by slow end-over-end inversions). Pelleted cells were resuspended in buffer (50 mM Tris-HCl pH 8.0, 10 mM sodium EDTA, pH 8.0, RNAse A, hemicellulase, lysozyme, zymolyase) and incubated at 37° C for 1 hour, mixing by inversion every 10 minutes. Cetrimonium bromide was added to 2% and NaCl to 1.5 M. The sample was incubated at 50° C for 1 hour, mixing by inversion every 15 minutes. The sample was centrifuged at 6000 x G for 3 minutes. The supernatant was collected and transferred to a new tube. One volume of 25:24:1 phenol:chloroform:isoamyl alcohol pH 8.0 was added, and mixed by inversion. Phases were separated by centrifugation at 8000 X g for 3 minutes. The aqueous phase was transferred to a new tube, where 1 volume of chloroform was added and mixed by inversion. The phases were separated by centrifugation at 6000 X g for 3 minutes, and the aqueous phase transferred to a new tube. One quarter volume of Tris-EDTA pH 8.0 was added to the chloroform, mixed, and centrifuged as previously. The aqueous phase was removed and combined, and the chloroform extraction was repeated once. After collecting the aqueous phase, sodium acetate (pH 5.2) was added to 0.3 M, and 2 volumes of cold 70% ethanol was added, mixed by inversion. The mixture was centrifuged at 16 000 X G for 2 minutes, and washed once with cold 70% ethanol. The pellet was resuspended in TE buffer (10 mM Tris-Cl 1mM EDTA pH 8.0), and stored at 4° C until further use. Short DNA fragments were then selectively removed using the Short Read Eliminator (SRE) kit from Circulomics (Baltimore), and the sample was stored at 4° C until sequencing. DNA from the same extraction was used for sequencing on both the Oxford Nanopore minION and Illumina NextSeq 550 platforms.

5.2.2 DNA sequencing

An Oxford Nanopore minION flow cell R9.4.1 was used with the SQK-LSK109 Kit according to the manufacterer's protocol version GDE_9063_v109_revK_14Aug2019, with one alteration: for DNA repair and end-prep, the reaction mixture was incubated for 15 minutes at 20° C and 15 minutes at 65° C. Basecalling was performed after the run with Guppy (Version 3.3.0). NanoPlot (De Coster et al. 2018) was used to generate Q-score versus length plots and summary statistics. The read N50 of the unfiltered reads was approximately 24 kb. Nanopore reads were not filtered prior to assembly (as expected by the assembler).

For Illumina sequencing, the Nextera XT kit was used to prepare 2×75 paired-end mid-output NextSeq 550 run at the London Regional Genomics Center (lrgc.ca). Reads were trimmed using Trimmomatic v0.36 (Bolger, Lohse, and Usadel 2014) in paired end mode with the following settings: AVGQUAL:30 CROP:75 SLIDINGWINDOW:4:25 MINLEN:50 TRAILING:15. SLIDINGWINDOW AND TRAILING were added to remove poor quality base calls. Only paired end

Length	GC	C	R	Illumina	Nanopore	rRNA	tRNA	Taxonomy
3.14	62.8	98.59	0	392	193	3	48	*Parvibaculum*
4.47	64.4	100	2.82	402	94	3	49	*ZYF759*
3.73	63.9	100	0	68	32	3	46	*Hyphomonas*
3.84	63.5	98.59	1.41	184	67	6	45	*Blastomonas*
3.74	72.4	100	0	121	39	6	63	*UBA2363* sp.
5.16	42.6	98.59	0	180	97	9	40	*Algoriphagus*
3.98	66.9	100	1.41	41	16	6	51	*Tabrizicola*
3.96	71.9	98.59	2.82	125	57	3	53	*JAAZBK01* sp.
4.68	66.1	100	2.82	57	17	6	52	*Gemmobacter*
5.79	66.3	100	2.82	35	14	6	55	*Aquimonas*
0.79	51	85.92	4.23	88	36	3	40	*UBA6175* sp.
3.06	65.9	100	0	1107	355	3	48	*Brevundimonas*
2.88	64.4	100	0	94	35	3	44	*Glycocaulis*

Table 5.1: **Summary results for metagenomically-assembled circular genomes.** Genome completeness, redundancy, taxonomy was predicted using Anvi'o. *UBA6175* sp. corresponds to the novel genome described in this report. Illumina and Nanopore columns represent the average read depth of the genome. Length is reported in megabases, %C and %R refers to estimates of percent completion and redundancy of single-copy core genes, respectively.

reads were retained.

5.2.3 Genome assembly

The raw long reads were used for long-read metagenomic assembly using Flye v2.6 (21) with the parameters –meta -g 5m. Circularized contigs larger than 300 kb were extracted as putative genomes. Thirteen circular genomes were obtained, and taxonomy was predicted using anvi-estimate-genome-taxonomy (22), which uses the Genome Taxonomy Database (23).

To obtain a consensus sequence for each genome, long reads were first separated for each genome by mapping all reads against each genome using minimap2 (24), following by filtering long reads using Gerenuq with a minimum read length of 1000 and minimum query coverage of 90%. Short reads were filtered using samtools view -F 3848 (remove reads where the mate pair doesn't align, alignments are not primary, alignments are supplementary). Each group of reads was subsequently used to polish the corresponding genome first with long reads using Racon (25) and

Medaka (Oxford Nanopore Technologies), followed by polishing using the highly-accurate short reads with Pilon (26).

5.2.4 Genome annotation and taxonomic prediction

Ribosomal protein sequences were annotated using anvi-run-hmms. To reconstruct metabolic pathways, KEGG orthology (KO) terms were obtained for each amino acid sequence predicted by progidal (27) using anvi-run-kegg-kofams. The KO predictions were then analyzed using the KEGG mapper tool (28). Genomes were also annotated using prokka (29). Taxonomic prediction was performed using the Genome Taxonomy Database Toolkit (23), and the full taxonomy is reported in Table 5.1. The taxonomy between the Genome Taxonomy Database and NCBI differ slightly for this radiation. For example, GTDB-tk refers to this radiation as Patescibacteria, whereas NCBI-taxonomy refers to it as Saccharibacteria. We chose to use the Genome Taxonomy Database nomenclature because it is solely dependent on the sequence of the genome. To build a phylogenetic tree, the phylogenomic workflow was performed with Anvi'o (22). Concatenated ribosomal protein sequences were used for alignment with MUSCLE, excluding the L30 and L9 proteins (2).

5.3 Results

5.3.1 Genome quality

From this community, we obtained multiple circularized and complete genomes (Table 5.1). The genome for *UBA6175* sp. is 794,452 bp long, with a GC content of approximately 51%. Sequencing coverage obtained for this genome averaged 88X for Illumina and 36X for nanopore (Table 5.1). Annotation by prokka (29) revealed 40 tRNAs, a single rRNA operon (5S, 16S, 23S), and 818 predicted coding sequences. Genome completeness and redundancy estimates in Table 5.1 were calculated using Anvi'o (22), which refers to the number of expected single copy core genes

user_genome	classification
Algoriphagus	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Cytophagales;f__Cyclobacteriaceae;g__Algoriphagus;
Aquimonas	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Aquimonas;
Blastomonas	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__Blastomonas;
Brevundimonas	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Caulobacterales;f__Caulobacteraceae;g__Brevundimonas;
Glycocaulis	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Caulobacterales;f__Maricaulaceae;g__Glycocaulis
Parvibaculum	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Parvibaculales;f__Parvibaculaceae;g__Parvibaculum;
ZYF759	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Rhizobiaceae;g__ZYF759;s__ZYF759
Gemmobacter	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhodobacterales;f__Rhodobacteraceae;g__Gemmobacter_C;
Tabrizicola	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhodobacterales;f__Rhodobacteraceae;g__Tabrizicola;
UBA6175	d__Bacteria;p__Patescibacteria;c__Saccharimonadia;o__Saccharimonadales;f__Saccharimonadaceae;g__UBA6175;
UBA2363	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__UBA2363;g__Pseudofulvimonas;
JAAZBK01	d__Bacteria;p__Actinobacteriota;c__Acidimicrobiia;o__Acidimicrobiales;f__JAAYBP01;g__JAAZBK01;
Hyphomonas	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Caulobacterales;f__Hyphomonadaceae;g__Hyphomonas;

Table 5.2: **Taxonomic predictions of assembled genomes by the Genome Taxonomy Database tool-kit (23).**

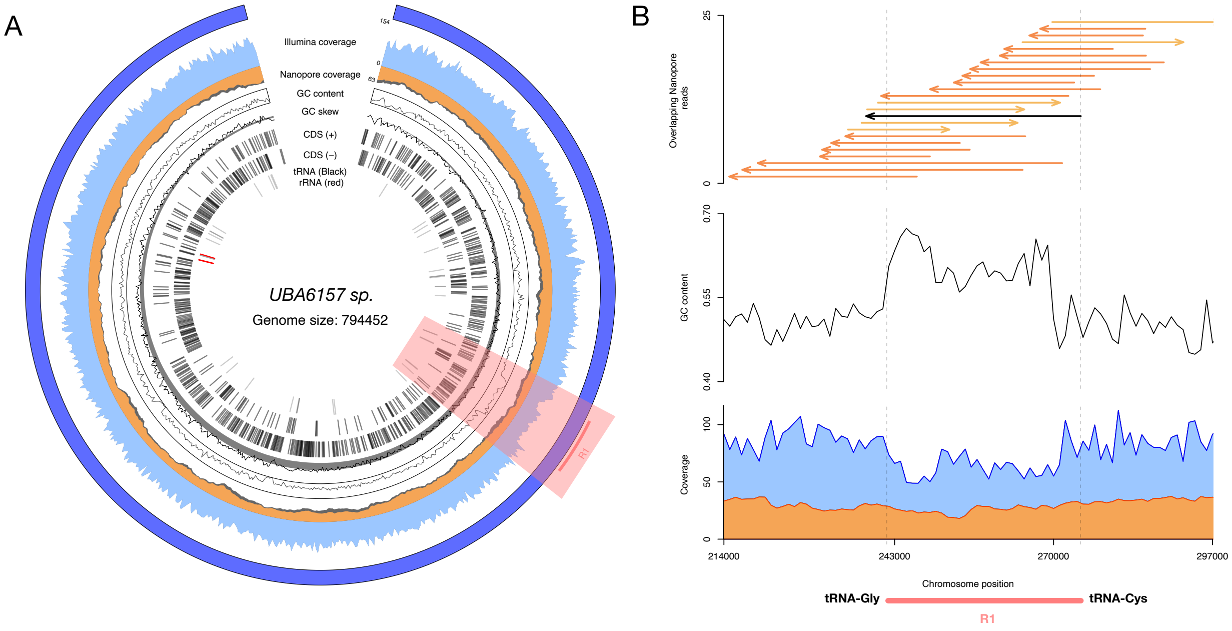


Figure 5.1: **Complete genome of *UBA6175* sp.** A) Summary of the genome. From out to in: Illumina sequencing depth (blue), Oxford Nanopore sequencing depth (filtered by 90% query coverage and minimum 5 kb length in orange, unfiltered in grey), GC content calculated in 1000 bp windows, GC skew, coding sequences predicted by prokka on positive and negative strand, tRNA and rRNA genes. B) Region R1 is flanked by tRNA genes. Grey dashed lines indicate the locations of tRNA-Gly and tRNA-Cys genes. Top) Ultra-long nanopore read alignments, negative strand alignments have left facing arrow heads, positive strand alignments have right-facing arrow heads. The black arrow is a read that spans both tRNA genes. Middle) GC content calculated in 100 bp windows. Bottom) Sequencing coverage, Illumina in blue, Nanopore in orange.

(4) that were found to be present. Although 85.92% completeness appears much lower than the other bacteria that are near 100%, this is due to the reduced biosynthetic capability of Patescibacteria that is commonly observed. A low completeness score may suggest a genome is incomplete, however, in this case it is expected given the predicted reduced metabolic capability.

To validate genome contiguity, nanopore reads were aligned and filtered to retain only reads where more than 90% of the read aligned against the genome with a minimum read length of 1000 bases (Figure 5.2), where no drops in coverage indicates complete contiguity. Highlighted in red is Region R1 in Figure 5.2 A). This region has a substantial increase in GC content to over 65% (Figure 5.2 B). In addition, Illumina sequencing depth drops at the region. To ensure this was a not an assembly error, we extracted and aligned long reads that partially (orange arrows) or completely (black arrow) spanned this region, confirming that contiguity is maintained, even though there is an unusual increase in GC content. The noticeable drop in Illumina sequencing depth is also typically indicative of an assembly error, however, the long overlapping reads demonstrates that there are no gaps at this region.

To estimate the per-base error rate in the final assembly, we used Merqury, which evaluates the k-mer content of the final assembly and compares it to high-quality Illumina reads (30). This resulted in a quality value (QV) of 53, which corresponds to an error rate of less than 1 in 100,000 bases. This genome can therefore be considered a finished metagenomically-assembled genome according to quality reporting standards previously established (a complete genome that is a gap-free genome with a consensus quality greater than Q50) (31).

5.3.2 Phylogenomic analysis reveals *UBA6175* sp. belongs to Clade G1

The *UBA6175* sp. genome lacks clearly defined orthologs of what was recently considered a universal ribosomal protein, L30 (32). In addition, it lacks an ortholog for a ribosomal protein that has been previously found in all bacteria, L9. These ribosomal proteins seem to be functionally compensated by unknown molecular mechanisms. These ribosomal protein sequences were therefore excluded from the phylogenomic analysis.

From the phylogenomic analysis, we found that *UBA6175* sp. belongs to the previously defined G1 clade (15). It appears to form a sister-clade with other genomes that were assembled from environmental sources (2) relative to the majority of the other G1 bacteria sequences which are associated with the oral microbiome.

5.3.3 *UBA6175* sp. has reduced functional capability

A core gene set for G1 and G6 clades of Saccharibacteria (NCBI taxonomy) has previously been shown that each clade contains unique and partial metabolic capabilities (33). The *UBA6175* sp. genome only encodes only a partial glycolysis pathway, lacking the ability to phosphorylate glucose in an ATP-dependent manner. Furthermore, it lacks the apparent ability to produce pyruvate.

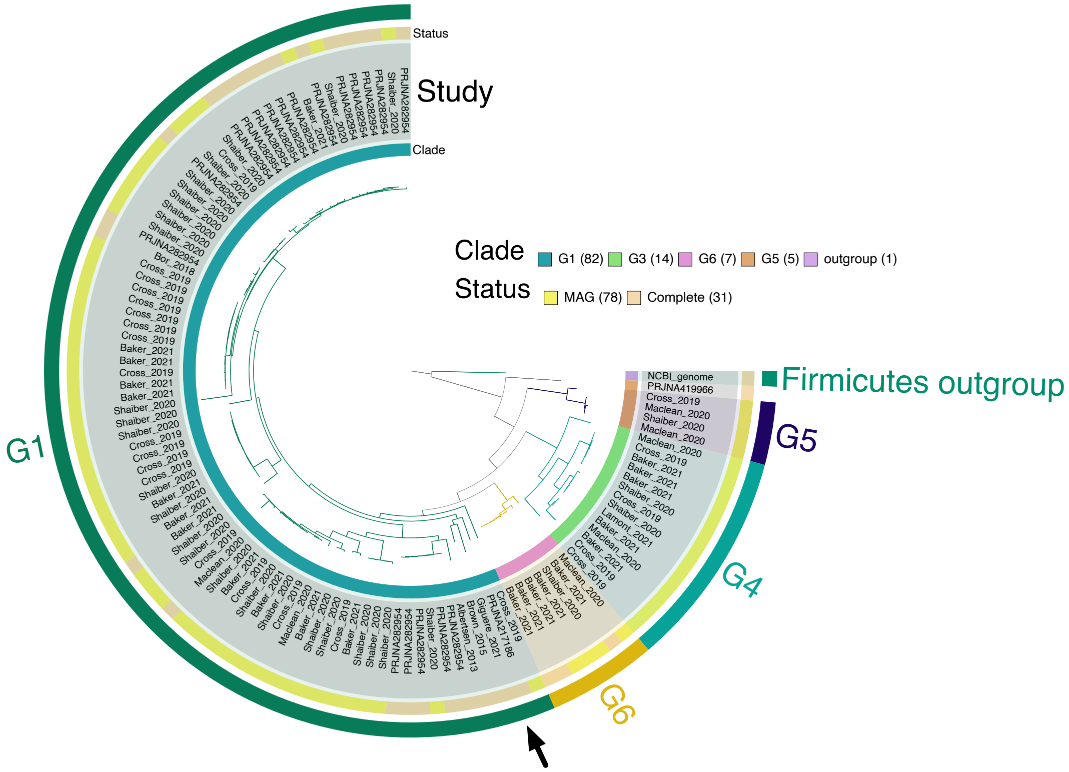

Status
Study
Clade
Clade 🟦 G1 (82) 🟩 G3 (14) 🟪 G6 (7) 🟧 G5 (5) 🟪 outgroup (1)
Status 🟨 MAG (78) 🟫 Complete (31)
Firmicutes outgroup
G1
G4
G5
G6

Figure 5.2: **Phylogenetic analysis of currently available Saccharibacteria (NCBI taxonomy) genomes and metagenomically-assembled genomes.**. The phylogenomic workflow from Anvi'o was performed using concatenated ribosomal protein sequences without the L9 or L30 proteins. A *Streptococcus* genome was used as an outgroup. Clades G1, G4, G5, and G6 are highlighted in gold, green, brown, and blue, respectively. Complete genomes are highlighted on the second outer track in light blue. The genome built in this chapter is highlighted with an arrow.

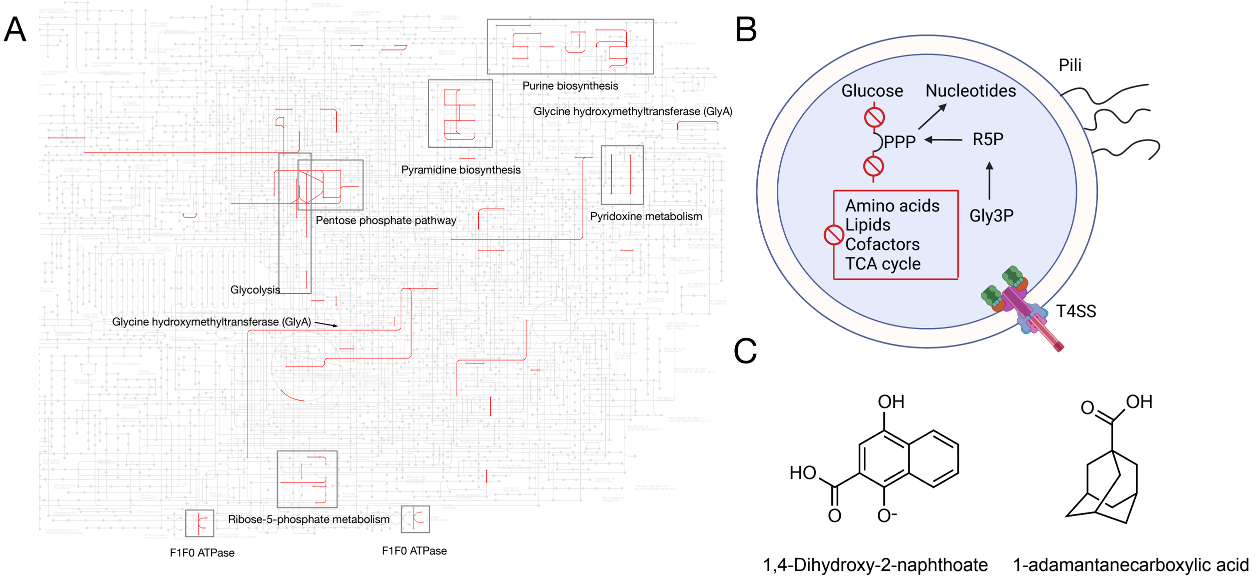

A
Purine biosynthesis
Glycine hydroxymethyltransferase (GlyA)
Pyramidine biosynthesis
Pentose phosphate pathway
Pyridoxine metabolism
Glycolysis
Glycine hydroxymethyltransferase (GlyA)
Ribose-5-phosphate metabolism
F1F0 ATPase
F1F0 ATPase
B
Pili
Glucose
Nucleotides
PPP
R5P
Gly3P
Amino acids
Lipids
Cofactors
TCA cycle
T4SS
C
OH
HO
O
O-
O
OH
1,4-Dihydroxy-2-naphthoate
1-adamantanecarboxylic acid

Figure 5.3: **Metabolic model for *UBA6175* sp.** A) KEGG pathway analysis. KO terms for annotated genes were predicted using anvi-estimate-metabolism, and KEGG mapper was used to analyze the pathways found. B) Metabolic model of *UBA6175* sp. Figure created with Biorender.com. Gly3P; glyceraldehyde-3P, R5P; ribulose-5-phosphate, PPP; pentose phosphate pathway, T4SS; type IV secretion system. C) Chemical structures of 1,4-Dihydroxy-2-naphthoate and 1-adamantanecarboxylic acid.

The genome appears to completely lack genes encoding the oxidative phase of the pentose phosphate pathway, however it does encode a portion of the reductive pentose phosphate pathway. The genome does not encode the TCA cycle, and also does not encode amino acid biosynthetic pathways. Interestingly, the genome does encode both purine and pyrimidine biosynthetic pathways, with additional ribose-5-phosphate metabolic capabilities. Similar to other G1 genomes, *UBA6175* sp. encodes an F1F0 ATPase. Taken together, *UBA6175* sp. contains few of the (previously considered) essential pathways for bacterial life.

Annotation using prokka of the Region R1 of the genome revealed the presence of the gene *menH*, demethylmenaquinone methyltransferase, which is one component of the pathway to synthesize menaquinol from 1,4-dihydroxy-2-naphthoate. The chemical structure is shown in Figure 5.4 C) with 1-adamantanecarboxylic acid, a compound that this algal-bacterial community is known to degrade (20).

protein id	description	taxa	query coverage	e value	percent identity
246	hypothetical protein	*Chryseoglobus* sp. 28M-23	99%	0	93.28%
247	hypothetical protein	*Chryseoglobus* sp. 28M-23	98%	9.00E-56	92.63%
248	replication-relaxation family protein	*Chryseoglobus* sp. 28M-23	99%	0	93.59%
250	type IV secretory system conjugative DNA transfer family protein	*Chryseoglobus* sp. 28M-23	99%	0	90.14%
251	hypothetical protein	*Chryseoglobus* sp. 28M-23	98%	5.00E-38	90.14%
253	methyltransferase	*Chryseoglobus* sp. 28M-23	99%	3.00E-141	94.55%
255	hypothetical protein	*Chryseoglobus* sp. 28M-23	99%	4.00E-63	95.10%
257	hypothetical protein	*Acidithrix ferrooxidans*	87%	2.00E-18	55.71%
258	hypothetical protein	*Agreia* sp. VKM Ac-1783	94%	2.00E-49	54.43%
259	Scr1 family TA system antitoxin-like transcriptional regulator	Acidithrix ferrooxidans	95%	7.00E-71	46.85%
260	hypothetical protein	*Chryseoglobus* sp. 28M-23	99%	3.00E-69	97.27%
261	DUF3846 domain-containing protein	*Chryseoglobus* sp. 28M-23	99%	1.00E-118	95.98%
262	type I restriction-modification system subunit M	*Mycobacteroides abscessus*	99%	0	82.86%
263	restriction endonuclease subunit S	*Arthrobacter* sp. FB24	98%	2.00E-124	50.91%
264	type I restriction endonuclease subunit R	*Chryseoglobus* sp. 28M-23	99%	0	94.91%
265	hypothetical protein	*Subtercola boreus*	99%	1.00E-143	94.42%
266	hypothetical protein	*Subtercola boreus*	99%	0	87.03%
267	hypothetical protein	*Leifsonia psychrotolerans*	90%	5.00E-32	57.84%
268	cadmium-translocating P-type ATPase	*Chryseoglobus indicus*	96%	0	93.66%
269	metal-sensitive transcriptional regulator	*Chryseoglobus indicus*	99%	8.00E-62	96.97%
270	DUF305 domain-containing protein	*Chryseoglobus indicus*	99%	7.00E-113	87.50%
271	alpha/beta hydrolase	*Chryseoglobus indicus*	99%	0	86.67%
272	hypothetical protein	*Chryseoglobus frigidaquae*	99%	7.00E-59	87.59%
273	DUF305 domain-containing protein	*Chryseoglobus* sp. 28M-23	99%	7.00E-107	91.84%
274	hypothetical protein	*Chryseoglobus indicus*	96%	0	92.00%
276	HD domain-containing protein	*Chromobacterium* sp. Panama	46%	1.00E-10	32.74%

Table 5.3: **Region R1 in the *UBA6175* sp. genome contains predicted proteins that are from Actinobacteriota.** Region R1 was extracted and genes were predicted and converted to amino acid sequencing using prodigal. BLASTP was then performed against the RefSeq database.

5.3.4 Region R1 may have been recently acquired

Region R1 was found to have an increase in GC content from approximately 51% to over 65%. While this may be considered evidence for a mis-assembly when a separate contig has an unexpectedly high GC content, ultra-long overlapping reads span the entire region, and many other reads anchor the region to both sides, demonstrating that this region is fully contiguous. To investigate this region further, we predicted the genes in this region using prodigal (27), and performed a BLAST search against the RefSeq database. We found that region R1 contained many proteins that align with very high percentage identity from a recently deposited genome from Chryseoglobus sp. 28M-23 (NCBI assembly accession: ASM1973919v1), a member of the Actinobacteriota phylum. According to the taxonomic predictions in Table 5.1, only 1 genome was predicted to belong to the Actinobacteriota genus, *JAAZBK01*.

Predicted proteins for the region of the *Chryseoglobus* genome that region R1 shares high protein similarity encodes a type IV secretory system conjugative DNA transfer protein and a replication-relaxosome protein, which are typically found as components of conjugative plasmids (34). It also encodes an accompanying type I complete restriction system (subunit M, S, and R). Interestingly, this regions appears to contain a cadmium-translocating P-type ATPase adjacent to a metal-sensitive transcriptional regulator (Table 5.3). Furthermore, annotation of tRNA genes using Aragorn (35) revealed tRNA gene predictions immediately flanking region R1 on both sides, which is a common characteristic of genomic islands (36).

Taken together, the increased GC content, flanking tRNA genes, the annotations of proteins related to conjugative plasmids, and the high similarity to another Actinobacteriota genome provides strong evidence that this strain of *Saccharimonadaceae* contains a novel genomic island (36).

5.4 Discussion

Here, we have generated a high-quality genome of a novel strain of the enigmatic *Saccharimonadaceae* from the recently described Patescibacteria, directly from the metagenome of a hydrocarbon-degrading community. We use a novel long-read filtering method to demonstrate contiguity, as well as a k-mer-based approach to estimate the per-base error rate for a *de novo*-assembled genome. We found that this genome belongs an environmentally-derived G1 clade that is a sister to other Patescibacteria derived from the oral microbiome, and conclude that this is a complete, finished-quality genome. To the best of our knowledge, we also conclude that this is the first *Saccharimonadaceae* genome reported with a genomic island.

The region R1 highlighted in Figure 5.2 contains several known criteria consistent with genomic islands (36). First, the tetranucleotide frequency is drastically different from the rest of the core genome, and region R1 has an average GC content over 65%, whereas the remainder of the genome has a GC content around 51%. Second, annotation of tRNAs revealed that region R1 is immediately flanked by tRNA genes. Integration into the 3' end of tRNA genes is known to occur by mobile genetic elements to generate such genomic islands (37, 38). Third, the presence of proteins related to conjugative transfer suggest that this may be a self-mobilizable element, or may have been mobilized in trans. We found a type IV secretory system conjugative DNA transfer protein, a full type 1 restriction modification system (subunit M, S, and R), a heavy-metal translocating P-type ATPase, and a gene involved in a biosynthetic pathway of compounds structurally similar to naphthenic acids, in addition to several proteins with unknown function that are highly conserved with a gram-positive Actinobacteriota. Lastly, the cadmium-translocating ATPase may improve fitness. Cadmium is known to be extremely toxic to micro-organisms (39) and cadmium is produced as a by-product from mining bitumen in the Albertan oil sands, resulting in airborne and riverborne contaminants released into the environment (40). The presence of a cadmium-translocating resistance gene in this *Saccharimonadaceae* strongly suggests that this is a recently acquired genomic island to improve fitness in its natural environment. To the best of our knowledge, this is the first observation of a high-GC genomic island acquisition event in *Saccharimonadaceae*. While no direct repeats were found, many of the characteristics of genomic islands (36) are present, suggesting this is a recently acquired element to improve fitness. Given the high protein sequence identity to *Chryseoglobus*, a gram positive Actinobacteriota, it is tempting to speculate that this genomic island may even be derived from its host bacterium, suggesting that direct DNA transfer between the basibiont and epibiont may be possible under stress conditions. The presence of this genomic island represents a major functional difference between other G1 Patescibacteria, whereas many of the G1 genomes have previously been shown to be highly syntenic (41).

5.5 Conclusions

We built a reference-quality metagenomically-assembled whole genome of a novel species of the family *Saccharimonadaceae* derived from a Northern Albertan tailings pond. To the best of our knowledge, this is the first observation of a high-GC genomic island acquisition event in this species. The genomic island contains genes that likely improve fitness in its natural environment, and genes that may be involved in hydrocarbon degradation. Long-read sequencing technologies and new metagenomic bioinformatic algorithms are enabling detailed genomic investigation of previously uncharacterized bacteria that may contribute to a wide variety of relevant problems, such as human oral health and even hydrocarbon degradation.

5.6 References

Bibliography

[1] Rheims, H.; Rainey, F. A.; Stackebrandt, E. A Molecular Approach to Search for Diversity among Bacteria in the Environment. *Journal of Industrial Microbiology* **1996**, *17*, 159–169.

[2] Brown, C. T.; Hug, L. A.; Thomas, B. C.; Sharon, I.; Castelle, C. J.; Singh, A.; Wilkins, M. J.; Wrighton, K. C.; Williams, K. H.; Banfield, J. F. Unusual Biology across a Group Comprising More than 15% of Domain Bacteria. *Nature* **2015**, *523*, 208–211.

[3] Hug, L. A. et al. A New View of the Tree of Life. *Nature Microbiology* **2016**, *1*, 1–6.

[4] Campbell, J. H.; O'Donoghue, P.; Campbell, A. G.; Schwientek, P.; Sczyrba, A.; Woyke, T.; Soll, D.; Podar, M. UGA Is an Additional Glycine Codon in Uncultured SR1 Bacteria from the Human Microbiota. *Proceedings of the National Academy of Sciences* **2013**, *110*, 5540–5545.

[5] He, X.; McLean, J. S.; Edlund, A.; Yooseph, S.; Hall, A. P.; Liu, S.-Y.; Dorrestein, P. C.; Esquenazi, E.; Hunter, R. C.; Cheng, G.; Nelson, K. E.; Lux, R.; Shi, W. Cultivation of a Human-Associated TM7 Phylotype Reveals a Reduced Genome and Epibiotic Parasitic Lifestyle. *Proceedings of the National Academy of Sciences* **2015**, *112*, 244–249.

[6] Bor, B.; McLean, J. S.; Foster, K. R.; Cen, L.; To, T. T.; Serrato-Guillen, A.; Dewhirst, F. E.; Shi, W.; He, X. Rapid Evolution of Decreased Host Susceptibility Drives a Stable Relationship between Ultrasmall Parasite TM7x and Its Bacterial Host. *Proceedings of the National Academy of Sciences* **2018**, *115*, 12277–12282.

[7] Murugkar, P. P.; Collins, A. J.; Chen, T.; Dewhirst, F. E. Isolation and Cultivation of Candidate Phyla Radiation Saccharibacteria (TM7) Bacteria in Coculture with Bacterial Hosts. *Journal of Oral Microbiology* **2020**, *12*, 1814666.

[8] Batinovic, S.; Rose, J. J. A.; Ratcliffe, J.; Seviour, R. J.; Petrovski, S. Cocultivation of an Ultrasmall Environmental Parasitic Bacterium with Lytic Ability against Bacteria Associated with Wastewater Foams. *Nature Microbiology* **2021**, *6*, 703–711.

[9] Collins, A. J.; Murugkar, P. P.; Dewhirst, F. E. Establishing Stable Binary Cultures of Symbiotic Saccharibacteria from the Oral Cavity. *Journal of Visualized Experiments: JoVE* **2021**,

[10] Chipashvili, O.; Utter, D. R.; Bedree, J. K.; Ma, Y.; Schulte, F.; Mascarin, G.; Alayyoubi, Y.; Chouhan, D.; Hardt, M.; Bidlack, F.; Hasturk, H.; He, X.; McLean, J. S.; Bor, B. Episymbiotic Saccharibacteria Suppresses Gingival Inflammation and Bone Loss in Mice through Host Bacterial Modulation. *Cell Host & Microbe* **2021**,

[11] Luo, C.; Xie, S.; Sun, W.; Li, X.; Cupples, A. M. Identification of a Novel Toluene-Degrading Bacterium from the Candidate Phylum TM7, as Determined by DNA Stable Isotope Probing. *Applied and Environmental Microbiology* **2009**, *75*, 4644–4647.

[12] Xie, S.; Sun, W.; Luo, C.; Cupples, A. M. Novel Aerobic Benzene Degrading Microorganisms Identified in Three Soils by Stable Isotope Probing. *Biodegradation* **2011**, *22*, 71–81.

[13] Figueroa-Gonzalez, P. A.; Bornemann, T. L. V.; Adam, P. S.; Plewka, J.; Révész, F.; von Hagen, C. A.; Táncsics, A.; Probst, A. J. Saccharibacteria as Organic Carbon Sinks in Hydrocarbon-Fueled Communities. *Frontiers in Microbiology* **2020**, *11*, 3343.

[14] Camanocha, A.; Dewhirst, F. E. Host-Associated Bacterial Taxa from Chlorobi, Chloroflexi, GN02, Synergistetes, SR1, TM7, and WPS-2 Phyla/Candidate Divisions. *Journal of Oral Microbiology* **2014**, *6*, 25468.

[15] Baker, J. L. Complete Genomes of Clade G6 Saccharibacteria Suggest a Divergent Ecological Niche and Lifestyle. *mSphere 6*, e00530–21.

[16] Moss, E. L.; Maghini, D. G.; Bhatt, A. S. Complete, Closed Bacterial Genomes from Microbiomes Using Nanopore Sequencing. *Nature Biotechnology* **2020**, *38*, 701–707.

[17] 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* **2020**, 2020.04.08.032540.

[18] Nicholls, S. M.; Quick, J. C.; Tang, S.; Loman, N. J. Ultra-Deep, Long-Read Nanopore Sequencing of Mock Microbial Community Standards. *GigaScience* **2019**, *8*.

[19] Stewart, R. D.; Auffret, M. D.; Warr, A.; Wiser, A. H.; Press, M. O.; Langford, K. W.; Liachko, I.; Snelling, T. J.; Dewhurst, R. J.; Walker, A. W.; Roehe, R.; Watson, M. Assembly of 913 Microbial Genomes from Metagenomic Sequencing of the Cow Rumen. *Nature Communications* **2018**, *9*, 870.

[20] Paulssen, J. M.; Gieg, L. M. Biodegradation of 1-Adamantanecarboxylic Acid by Algal-Bacterial Microbial Communities Derived from Oil Sands Tailings Ponds. *Algal Research* **2019**, *41*, 101528.

[21] Kolmogorov, M.; Bickhart, D. M.; Behsaz, B.; Gurevich, A.; Rayko, M.; Shin, S. B.; Kuhn, K.; Yuan, J.; Polevikov, E.; Smith, T. P. L.; Pevzner, P. A. metaFlye: Scalable Long-Read Metagenome Assembly Using Repeat Graphs. *Nature Methods* **2020**, *17*, 1103–1110.

[22] Eren, A. M.; Esen, Ö. C.; Quince, C.; Vineis, J. H.; Morrison, H. G.; Sogin, M. L.; Delmont, T. O. Anvi'o: An Advanced Analysis and Visualization Platform for 'omics Data. *PeerJ* **2015**, *3*, e1319.

[23] Chaumeil, P.-A.; Mussig, A. J.; Hugenholtz, P.; Parks, D. H. GTDB-Tk: A Toolkit to Classify Genomes with the Genome Taxonomy Database. *Bioinformatics* **2020**, *36*, 1925–1927.

[24] Li, H. Minimap and Miniasm: Fast Mapping and de Novo Assembly for Noisy Long Sequences. *Bioinformatics* **2016**, *32*, 2103–2110.

[25] Vaser, R.; Sović, I.; Nagarajan, N.; Šikić, M. Fast and Accurate de Novo Genome Assembly from Long Uncorrected Reads. *Genome Research* **2017**, *27*, 737–746.

[26] Walker, B. J.; Abeel, T.; Shea, T.; Priest, M.; Abouelliel, A.; Sakthikumar, S.; Cuomo, C. A.; Zeng, Q.; Wortman, J.; Young, S. K.; Earl, A. M. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PloS One* **2014**, *9*, e112963.

[27] Hyatt, D.; Chen, G.-L.; Locascio, P. F.; Land, M. L.; Larimer, F. W.; Hauser, L. J. Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification. *BMC bioinformatics* **2010**, *11*, 119.

[28] Kanehisa, M.; Sato, Y.; Kawashima, M. KEGG Mapping Tools for Uncovering Hidden Features in Biological Data. *Protein Science n/a*.

[29] Seemann, T. Prokka: Rapid Prokaryotic Genome Annotation. *Bioinformatics* **2014**, *30*, 2068–2069.

[30] Rhie, A.; Walenz, B. P.; Koren, S.; Phillippy, A. M. Merqury: Reference-Free Quality, Completeness, and Phasing Assessment for Genome Assemblies. *Genome Biology* **2020**, *21*, 245.

[31] Bowers, R. M. et al. Minimum Information about a Single Amplified Genome (MISAG) and a Metagenome-Assembled Genome (MIMAG) of Bacteria and Archaea. *Nature Biotechnology* **2017**, *35*, 725–731.

[32] Korobeinikova, A. V.; Garber, M. B.; Gongadze, G. M. Ribosomal Proteins: Structure, Function, and Evolution. *Biochemistry. Biokhimiia* **2012**, *77*, 562–574.

[33] Baker, J. L.; Morton, J. T.; Dinis, M.; Alvarez, R.; Tran, N. C.; Knight, R.; Edlund, A. Deep Metagenomics Examines the Oral Microbiome during Dental Caries, Revealing Novel Taxa and Co-Occurrences with Host Molecules. *Genome Research* **2021**, *31*, 64–74.

[34] Smillie, C.; Garcillán-Barcia, M. P.; Francia, M. V.; Rocha, E. P. C.; de la Cruz, F. Mobility of Plasmids. *Microbiology and Molecular Biology Reviews* **2010**, *74*, 434–452.

[35] Laslett, D.; Canback, B. ARAGORN, a Program to Detect tRNA Genes and tmRNA Genes in Nucleotide Sequences. *Nucleic Acids Research* **2004**, *32*, 11–16.

[36] Juhas, M.; van der Meer, J. R.; Gaillard, M.; Harding, R. M.; Hood, D. W.; Crook, D. W. Genomic Islands: Tools of Bacterial Horizontal Gene Transfer and Evolution. *FEMS Microbiology Reviews* **2009**, *33*, 376–393.

[37] Williams, K. P. Integration Sites for Genetic Elements in Prokaryotic tRNA and tmRNA Genes: Sublocation Preference of Integrase Subfamilies. *Nucleic Acids Research* **2002**, *30*, 866–875.

[38] Liu, H.-l.; Zhu, J. Analysis of the 3′ Ends of tRNA as the Cause of Insertion Sites of Foreign DNA in Prochlorococcus. *Journal of Zhejiang University. Science. B* **2010**, *11*, 708–718.

[39] Trevors, J. T.; Stratton, G. W.; Gadd, G. M. Cadmium Transport, Resistance, and Toxicity in Bacteria, Algae, and Fungi. *Canadian Journal of Microbiology* **1986**, *32*, 447–464.

[40] Kelly, E. N.; Schindler, D. W.; Hodson, P. V.; Short, J. W.; Radmanovich, R.; Nielsen, C. C. Oil Sands Development Contributes Elements Toxic at Low Concentrations to the Athabasca River and Its Tributaries. *Proceedings of the National Academy of Sciences* **2010**, *107*, 16178–16183.

[41] McLean, J. S.; Bor, B.; Kerns, K. A.; Liu, Q.; To, T. T.; Solden, L.; Hendrickson, E. L.; Wrighton, K.; Shi, W.; He, X. Acquisition and Adaptation of Ultra-Small Parasitic Reduced Genome Bacteria to Mammalian Hosts. *Cell Reports* **2020**, *32*, 107939.

Chapter 6

General discussion

A method to determine the sequence of DNA bases in an organism's genome was first published in 1977, enabling targeted and accurate sequencing of short regions (1). The improvements to throughput, mainly enabled by automation, allowed humans to understand the genetic basis of countless diseases by completing a draft of the first human genome in the early 2000s (2, 3). However, DNA sequenced by nanopores enables ultra-long sequencing reads to be obtained, and this has fundamentally changed what insights can be obtained in DNA sequencing data. The overall goal of this thesis has been to explore this new type of sequencing data, develop new methods for analysis, and directly apply them to better understand microorganisms, both novel and known.

6.1 On the definition of a "complete" genome

The question of whether a static genome assembly accurately recapitulates an organism's genome as found in nature has now become subject to debate since long read technology enables routine "complete" genome reconstruction (4, 5). For many bacterial species, a "complete" genome may be a single circular chromosome, and an accompanying small plasmid (if present). In eukaryotes, a complete genome is often a collection of linear nuclear chromosomes, such as the 22 pairs of autosomes and 1 pair of sex chromosomes in humans (6). A static representation of these genomes as a text file generated from genome assemblies is the current standard (7). However, as researchers further investigate non-model organisms with unique biology, such as mixed bacterial communities that cannot be cultured in a laboratory and non-human eukaryotes, it is becoming evident that a static genome assembly may not be sufficient to accurately represent what exists naturally.

In mixed bacterial communities, integrative conjugative elements like conjugative plasmids or genomic islands, are known to transfer functional capability throughout the community (8). When the rate of horizontal gene transfer is high enough, it is possible that multiple versions of a genome may exist for the same species within the same environment, each with unique mobile elements. Now that circular bacterial genomes can be regularly assembled directly from DNA extracted from bacterial communities without enrichment or culturing, if multiple genomic versions of the same bacteria are present, which one is complete, and which one is incomplete? If two genomes differ by only the presence or absence of a horizontal transfer event such as a genomic island, should this be considered two unique genomes? Indeed, debate as to whether a bacterial genome should be considered a mobile genetic element itself now exists in the literature (5).

Similarly, in eukaryotic organisms with high rates of mitotic recombination, like *Phaeodactylum tricornutum*, there can exist more than the 2 canonical haplotypes of the diploid genome since the DNA sequenced is derived from a population of actively recombining cells, yet it is still inherently the same species with nearly the same functional potential. What constitutes “a complete genome sequence” in such an organism where many variations naturally occur, even from a population of cells that asexually reproduce from a single cell (9)? Likely owing to this recombination, the number of chromosomes in the species has not been resolved since the draft assembly was published in 2008, and recent estimates suggest it contains between 20 and 33 chromosomes (10, 11). Using the novel approach described in Chapter 3 of this thesis, I was able to resolve the number of chromosomes independently from our telomere-to-telomere genome assembly presented in Chapter 4 to provide two orthogonal, agreeing estimates for the number of chromosomes for this species. Interestingly, while most of the chromosomes contained the expected 4 telomeres per chromosome (2 per haplotype), there were 3 chromosomes that contained an unexpected number of telomeres, and even one telomere that was found to align to the middle of a chromosome. This suggests that on top of mitotic recombination producing many gene-level haplotypes (like single nucleotide polymorphisms), there may be chromosomal-level recombination at the telomeres between different pairs of chromosomes. While we were able to determine a base number of chromosomes for *Phaeodactylum tricornutum* of 25, new questions arise. Is the number of chromosomes variable depending on recombination activity? Are large portions of chromosomes recombined often? Are certain chromosomes more likely to recombine than others? Where and why do the recombinations occur? Using the approaches I developed in Chapter 3 and Chapter 4, these questions could be answered in the future using ultra-long read sequencing.

The work presented in this thesis provides further evidence that genomes are much more dynamic than what a static text file containing the order of deoxyribonucleotides may suggest, especially in understudied non-model organisms.

6.2 Long reads enable deeper biological inference from genome assemblies

Genome assembly algorithms have become incredibly powerful as high-throughput sequencing technologies have evolved, and we are now at the cusp of being able to routinely generate "complete" genome assemblies for nearly every sample sequenced without significant manual curation (12). However, an evolving understanding of genomes requires an evolution of bioinformatic tools. For example, it is no longer informative to report on the number and size of contigs if the output sequences are complete circular bacterial genomes; the size of the genome and number of plasmids should be reported. In addition, since assembly of linear chromosomes can now be completed telomere-to-telomere, it is much more biologically informative to report on the number and size of the chromosomes than the number of contigs assembled. Ideally, these two pieces of information would be congruent in a telomere-to-telomere genome assembly.

In addition, it has been shown that even the most recent genome assembly algorithms are not perfect (13), highlighting the need to ensure that each assembled genome should be individually investigated. We found that alternating haplotypes in both bacteria and eukaryotes can be visualized by evaluating long-read sequencing coverage after filtering by read length and query coverage (14, 15). Such a visualizations enables a strategy for identifying where mis-assemblies have occurred, where alternate functionally important haplotypes may exist, and allows researchers to ensure a genome assembly is fully contiguous.

Nanopore sequencing enables biological inference that was not possible with previous technologies. I showed in Chapter 3 that is possible to estimate the number of chromosomes directly from long sequencing without doing a genome assembly. Since this strategy only depends on the presence of linear dsDNA elements and the telomere sequence, I hypothesize that this approach could also be used to identify and validate other linear dsDNA elements, such as the recently proposed extra-chromosomal element "borgs" (16). While eukaryotic nuclear chromosomes have telomere repeats that are essential for this approach, borgs are hypothesized to contain kilobase-sized terminal inverted repeats at the start and end of the elements that could be used instead. This hypothesis remains to be tested.

Furthermore, visualizing filtered long-read sequencing depth enables a deeper understanding than just the sequence itself. In Chapter 4, the long-read sequencing depth of chromosome 19 for *Phaeodactylum tricornutum* was shown to be consistent across the entire chromosome, whereas the observed sequencing depth for all other nuclear chromosomes varied due to alternating haplotypes. In addition, only 2 telomeres were identified for this chromosome in Chapter 3. A new question arises: why does chromosome 19 have no observed haplotype variation, unlike all other chromosomes? Diatoms are known to have a wide variety of reproductive mechanisms, often initiated by environmental cues such as starvation conditions (17). While *Phaeodactylum tricornutum* is theorized to be capable of sexual reproduction, it has not been observed experimentally (18). Brown algae (including diatoms) are known to contain a diverse set of reproductive mechanisms, yet *Phaeodactlyum tricornutum* has only been observed to reproduce asexually. The single observed haplotype of chromosome 19 is reminiscent of chromosomes with suppressed recombination, like regions of sex chromosomes in other brown algae (19). A filamentous brown algae, *Ectocarpus*, contains a sex chromosome that includes a small sex-determining region, where recombination is completely suppressed. Is there suppressed recombination over the entirety of chromosome 19 because the entire chromosome is sex-determining? This is a question that remains unanswered for this species and requires further investigation. To the best of my knowledge, Chapter 3 and 4 provides the first demonstration that sequencing coverage alone can be used to determine that only a single haplotype exists for chromosome 19 in *Phaeodactylum tricornutum*.

On the surface, long read sequence data may appear similar to short read sequence data with the main differences being read length and sequence quality. However, this thesis demonstrates that long read sequencing data provides more than just the sum of assembled sequences. It encodes biological information that was not possible to obtain with previous sequencing technologies.

6.3 Fully understanding genomes is required for synthetic biology

Methods to improve genome assemblies for both wet-lab and dry-lab have been presented in this thesis, enabling deeper biological insights for both novel and previously known species, across two kingdoms of life. Understanding as much as possible about genomes will enable future investigation into genetic modifications and synthetic biology applications.

P. tricornutum is gaining traction as a chassis for protein production since it can perform N-linked glycosylation similarly to humans, meaning it can be used as a platform to produce biologically active human antibodies (20, 21). While Cas9 editing has previously been used to knockout uracil and histidine biosynthetic pathways that can be rescued with a complementary plasmid in *P. tricornutum* (22), the addition of genetic material to nuclear chromosomes may not be stable due to mitotic recombination. If recombination in the middle of an operon occurs, it could become non functional. However, data presented in Chapter 4 presents a unique target for the introduction of exogenous pathways to *P. tricornutum*. Since recombination on chromosome 19 appears to be suppressed, it may be a more suitable location for the integration of genetic material for synthetic biology applications. Integrating functional pathways onto this chromosome would likely provide additional stability relative to other chromosomes since mitotic recombination does not occur as frequently.

The complete genome of *UBA6175* sp. presented in Chapter 5 revealed that a genomic island was recently acquired. Engineering the genomes of *Saccharimonadaceae* has yet to be explored, but may present an opportunity for a novel chassis for various synthetic biology applications. For example, it was recently found that *Saccharimonadaceae* suppresses gingival inflammation by modulating their hosts in the oral microbiome (23). This presents a unique, targeted application that could benefit human health. Given the presence of a cadmium-resistance gene in the genomic island, and the known heavy metal toxicity of tailings ponds in Northern Alberta (24), this finding suggests that the genomic island was recently acquired to improve fitness. This means that it may be possible to transfer functional genetic elements to this bacterium under selective pressure. Is it possible to transform *Saccharimonadaceae* with exogenous DNA to introduce functional genetic elements that could be beneficial to the human oral microbiome by further modulating their host? Similar to the plasmid-based complementation approach used in *P. tricornutum*, could such an approach be used to create a symbiosis between *Saccharimonadaceae* and a designer host? *Saccharimonadaceae* remain largely unexplored, especially in the context of genetic editing. The data presented in this thesis suggests that it may be possible to transfer and integrate exogenous DNA into these enigmatic bacteria via selective pressure.

6.4 Future work and final thoughts

A large portion of this thesis involved analyzing, interpreting, and developing a deeper understanding the newest type of DNA sequencing data: long reads. Nanopore sequencing is a rapidly developing technology, and the analysis of the underlying data is evolving as well. A significant question remains: how can one estimate the per-base quality of a *de novo* nanopore-only assembly that does not have a comparable reference sequence available? This will be important to answer moving forward.

The over-arching reasons to completely and accurately understand the genomes of microorganisms is twofold: 1) understanding, diagnosing, and improving human health; and 2) understanding micro-organisms that can be engineered and applied industrially to improve human health. Synthetic biology has the capability to vastly improve the quality of human life by enabling biological manufacturing of compounds to treat disease, reduce environmental impact of manufacturing, and even has the ability to correct human negligence, such as bioremediation of oil spills and plastics. However, to engineer biology to benefit humans has one fundamental requirement: an accurate and complete understanding of the DNA sequence of the biological chassis being used.

I strongly believe this thesis advances the methods needed to understand long read sequencing data, and I demonstrated applications of these methods to projects to advance the field of synthetic biology. I showed that filtering long reads by query coverage can identify where additional bacterial haplotypes exist. I developed the first approach to estimate the number of chromosomes directly from long DNA reads only. Using these methods, I completed the first collapsed telomere-to-telomere genome assembly for *Phaeodactylum tricornutum*, revealing additional information about the fundamental biology of this species. I also showed for the first time that a bacterium from the recently described phylum, Patescibacteria, is capable of acquiring a genomic island, suggesting genetic manipulation may be possible. The methods and results presented in this thesis push the boundaries of analysis in the DNA sequencing field, and enable further development in the field of synthetic biology.

6.5 References

Bibliography

[1] Sanger, F.; Nicklen, S.; Coulson, A. R. DNA Sequencing with Chain-Terminating Inhibitors. *Proceedings of the National Academy of Sciences* **1977**, *74*, 5463–5467.

[2] Venter, J. C. et al. The Sequence of the Human Genome. *Science* **2001**, *291*, 1304–1351.

[3] Lander, E. S. et al. Initial Sequencing and Analysis of the Human Genome. *Nature* **2001**, *409*, 860–921.

[4] Nicholls, S. Computational Recovery of Enzyme Haplotypes from a Metagenome. Ph.D. thesis, Aberystwyth University, 2018.

[5] Hall, J. P. J. Is the Bacterial Chromosome a Mobile Genetic Element? *Nature Communications* **2021**, *12*, 6400.

[6] Ford, C. E.; Hamerton, J. L. The Chromosomes of Man. *Nature* **1956**, *178*, 1020–1023.

[7] Nurk, S. et al. The Complete Sequence of a Human Genome. 2021.

[8] Zatyka, M.; Thomas, C. M. Control of Genes for Conjugative Transfer of Plasmids and Other Mobile Elements. *FEMS Microbiology Reviews* **1998**, *21*, 291–319.

[9] Bulankova, P. et al. Mitotic Recombination between Homologous Chromosomes Drives Genomic Diversity in Diatoms. *Current Biology* **2021**, *31*, 3221–3232.e9.

[10] Bowler, C. et al. The Phaeodactylum Genome Reveals the Evolutionary History of Diatom Genomes. *Nature* **2008**, *456*, 239–244.

[11] Bowler, C.; Falciatore, A. Phaeodactylum Tricornutum. *Trends in Genetics* **2019**, *35*, 706–707.

[12] Cheng, H.; Concepcion, G. T.; Feng, X.; Zhang, H.; Li, H. Haplotype-Resolved de Novo Assembly Using Phased Assembly Graphs with Hifiasm. *Nature Methods* **2021**, *18*, 170–175.

[13] Wick, R. R.; Holt, K. E. Benchmarking of Long-Read Assemblers for Prokaryote Whole Genome Sequencing. *F1000Research* **2021**, *8*, 2138.

[14] 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* **2020**, 2020.04.08.032540.

[15] Giguere, D. J.; Bahcheli, A. T.; Slattery, S. S.; Patel, R. R.; Flatley, M.; Karas, B. J.; Edgell, D. R.; Gloor, G. B. Telomere-to-Telomere Genome Assembly of Phaeodactylum Tricornutum. 2021.

[16] Al-Shayeb, B.; Schoelmerich, M. C.; West-Roberts, J.; Valentin-Alvarado, L. E.; Sachdeva, R.; Mullen, S.; Crits-Christoph, A.; Wilkins, M. J.; Williams, K. H.; Doudna, J. A.; Banfield, J. F. Borgs Are Giant Extrachromosomal Elements with the Potential to Augment Methane Oxidation. 2021.

[17] Frenkel, J.; Vyverman, W.; Pohnert, G. Pheromone Signaling during Sexual Reproduction in Algae. *The Plant Journal* **2014**, *79*, 632–644.

[18] Mao, Y.; Guo, L.; Luo, Y.; Tang, Z.; Li, W.; Dong, W. Sexual Reproduction Potential Implied by Functional Analysis of SPO11 in Phaeodactylum Tricornutum. *Gene* **2020**, *757*, 144929.

[19] Ahmed, S. et al. A Haploid System of Sex Determination in the Brown Alga Ectocarpus Sp. *Current Biology* **2014**, *24*, 1945–1957.

[20] Baïet, B.; Burel, C.; Saint-Jean, B.; Louvet, R.; Menu-Bouaouiche, L.; Kiefer-Meyer, M.-C.; Mathieu-Rivet, E.; Lefebvre, T.; Castel, H.; Carlier, A.; Cadoret, J.-P.; Lerouge, P.; Bardor, M. N-Glycans of Phaeodactylum Tricornutum Diatom and Functional Characterization of Its N-Acetylglucosaminyltransferase I Enzyme. *The Journal of Biological Chemistry* **2011**, *286*, 6152–6164.

[21] Hempel, F.; Maier, U. G. An Engineered Diatom Acting like a Plasma Cell Secreting Human IgG Antibodies with High Efficiency. *Microbial Cell Factories* **2012**, *11*, 126.

[22] Slattery, S. S.; Wang, H.; Giguere, D. J.; Kocsis, C.; Urquhart, B. L.; Karas, B. J.; Edgell, D. R. Plasmid-Based Complementation of Large Deletions in Phaeodactylum Tricornutum Biosynthetic Genes Generated by Cas9 Editing. *Scientific Reports* **2020**, *10*, 13879.

[23] Chipashvili, O.; Utter, D. R.; Bedree, J. K.; Ma, Y.; Schulte, F.; Mascarin, G.; Alayyoubi, Y.; Chouhan, D.; Hardt, M.; Bidlack, F.; Hasturk, H.; He, X.; McLean, J. S.; Bor, B. Episymbiotic Saccharibacteria Suppresses Gingival Inflammation and Bone Loss in Mice through Host Bacterial Modulation. *Cell Host & Microbe* **2021**,

[24] Kelly, E. N.; Schindler, D. W.; Hodson, P. V.; Short, J. W.; Radmanovich, R.; Nielsen, C. C. Oil Sands Development Contributes Elements Toxic at Low Concentrations to the Athabasca River and Its Tributaries. *Proceedings of the National Academy of Sciences* **2010**, *107*, 16178–16183.

Appendix A

Supplemental information for Chapter 2

A.1 Supplemental figures

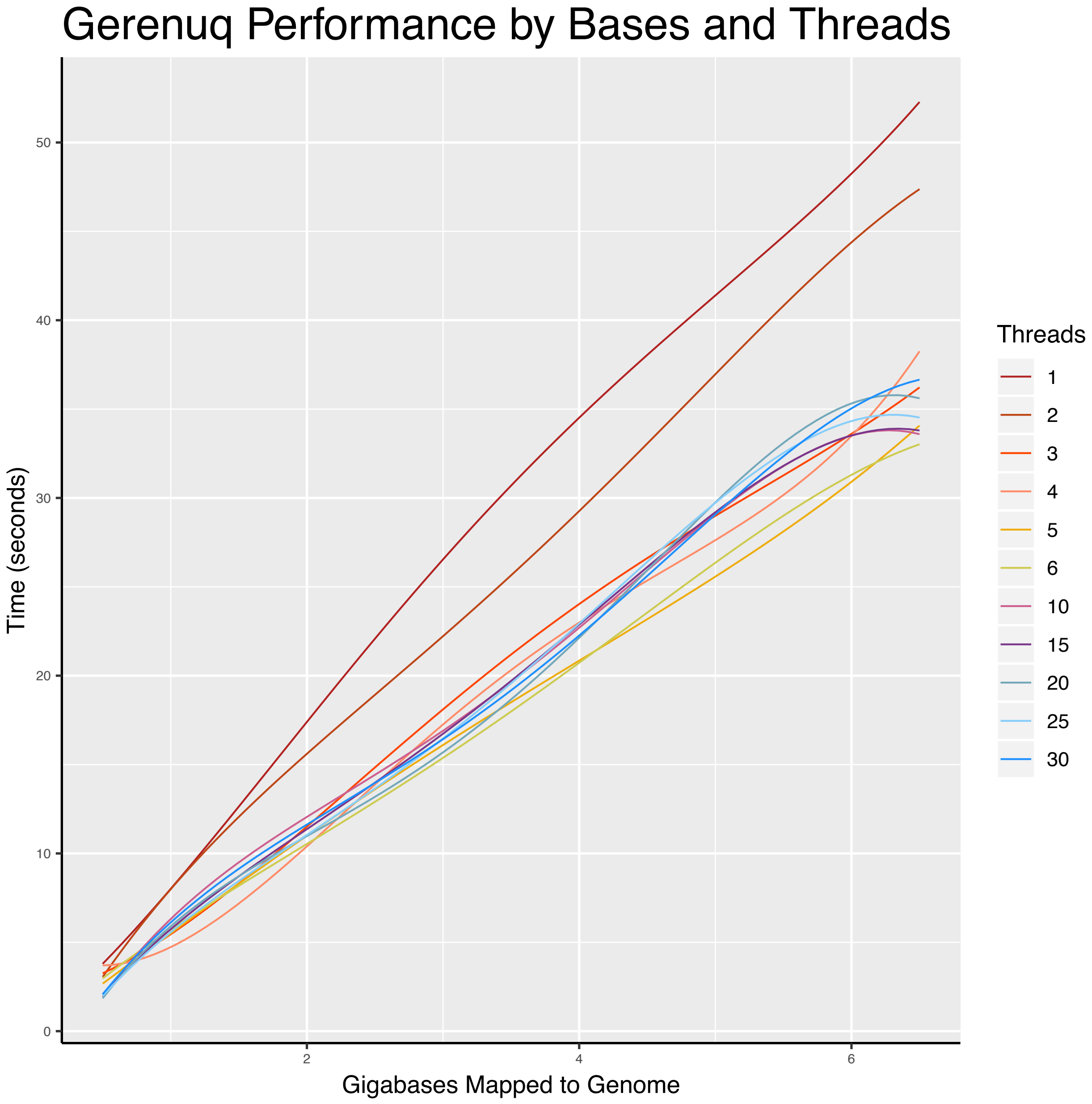


Figure A.1: Performance of Gerenuq. Reads were filtered using a different number of threads.

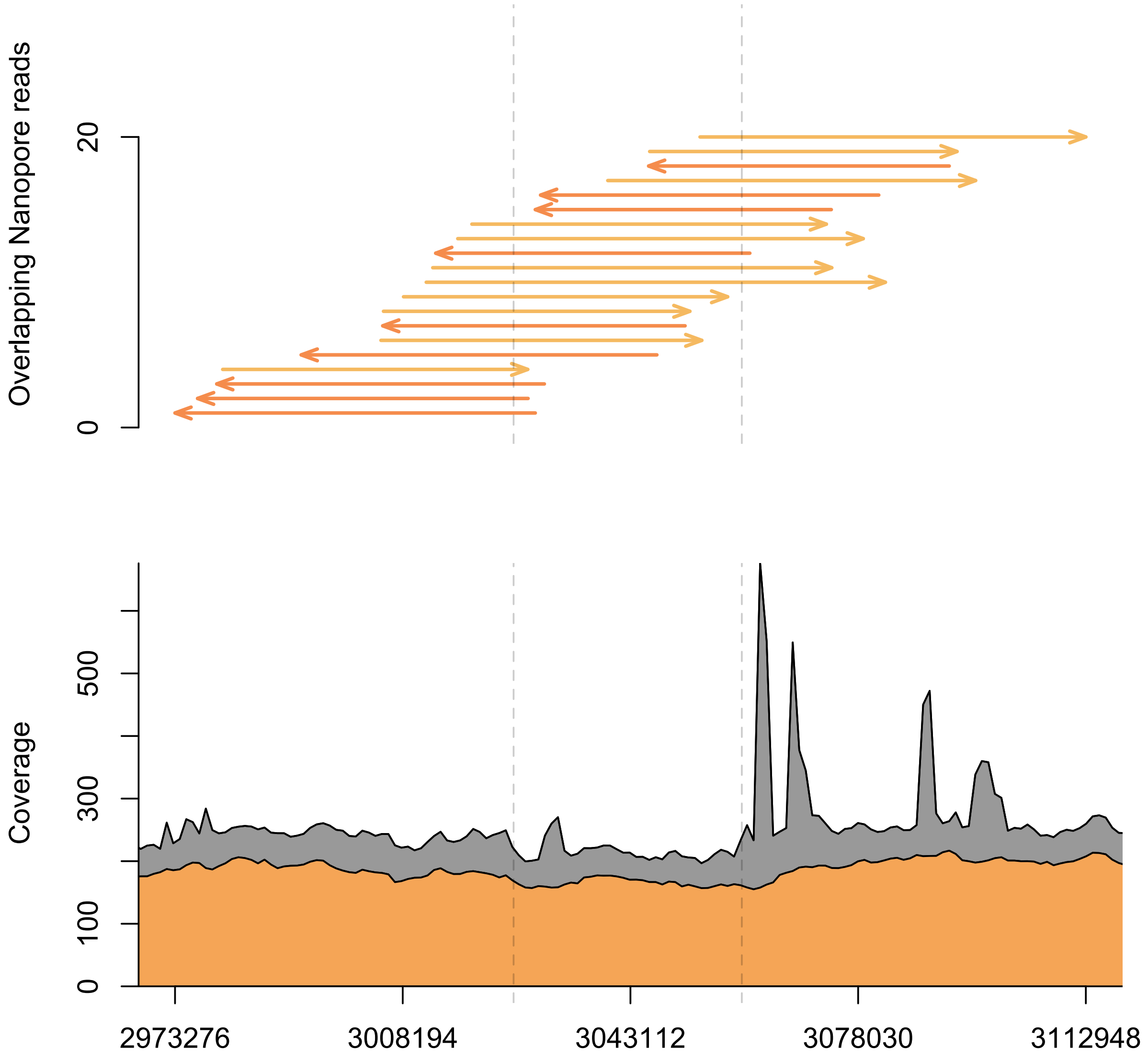


Figure A.2: Long read coverage of an alternate haplotype. Top, long reads were mapped against the genome with the haplotype included. Right facing arrows indicate reads aligned to the positive strand, left facing arrows indicate reads aligned to the negative strand. Bottom, observed sequencing coverage before (grey) and after (orange) filtering by Gerenuq.

Appendix B

Supplemental information for Chapter 4

B.1 Supplemental figures

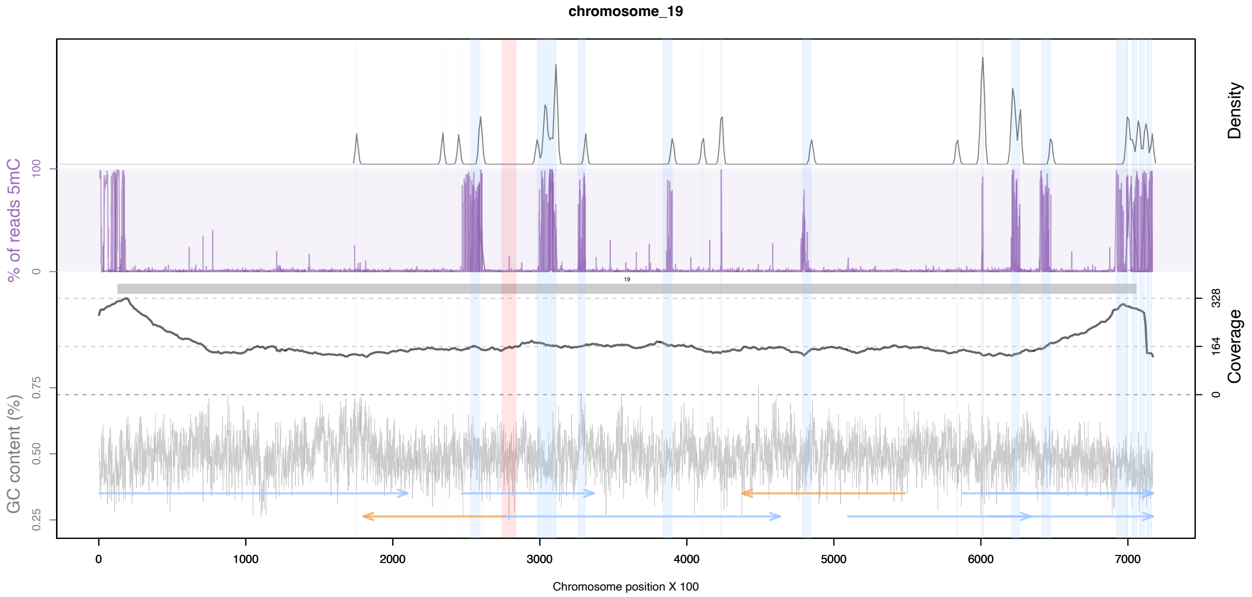


Figure B.1: Summary of genomic features for chromosome 19. A) The density of LTR-retrotransposons as predicted by the EDTA pipeline. B) The proportion of reads that were called as methylated at each position along the chromosome. C) Scaffolds from the previous assembly are overlayed in gray bars, with dark grey representing overlapping regions. D) Filtered long-read coverage (minimum 20 kb length and 70% query coverage). E) GC content calculated and plotted in 100 base windows. An overlapping read tiling path, with a minimum overlap of 30 kb, is shown with orange indicating reads mapping to the negative strand and blue indicating reads mapping to the positive strand. The regions that are annotated at LTR-retrotransposons are highlighted in light blue.

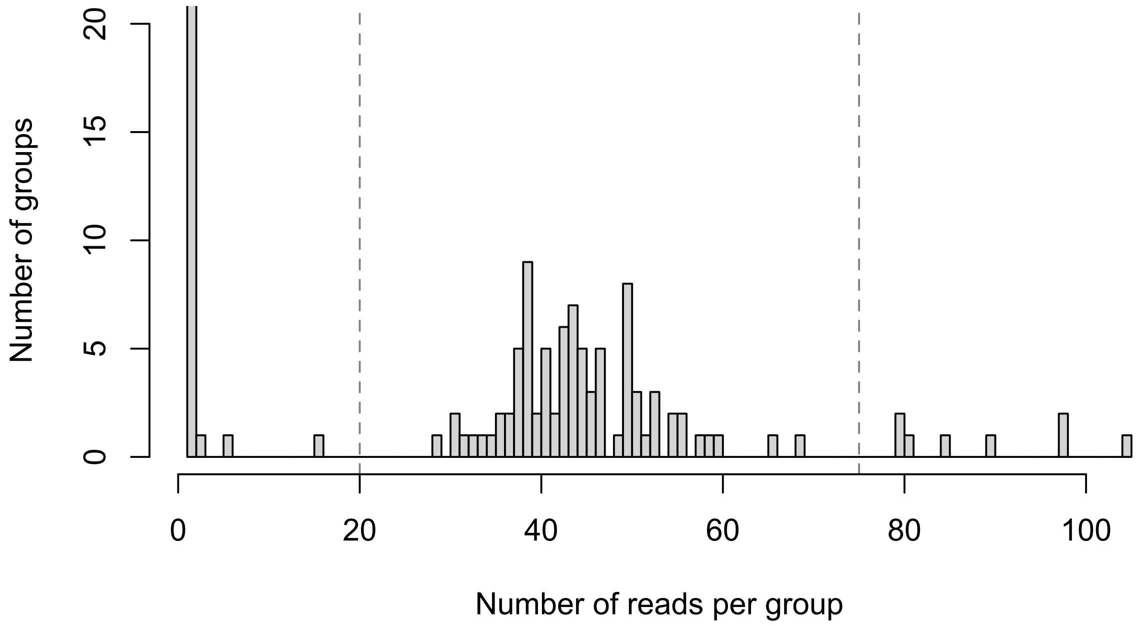


Figure B.2: Histogram of telomere clusters.

B.2 Supplemental tables

Table B.1: Overlapping read path for long reads aligning to all chromosomes. Alignment positions of individual reads are shown, with minimum 10 000 base overlaps for all chromosomes.

Read name	Chromosome	Start alignment position	End alignment position
12f00f00-e76d-4849-a8df-03c55ddbcd53	chromosome_1	114	136747
8ddefd10-857a-48d7-b326-67e291254ea1	chromosome_1	86768	234048
5ad0cc7c-4bba-4f6e-b3a1-30efa627589f	chromosome_1	199274	336314
e8a72bec-d1da-48b8-917d-f9b252a801ef	chromosome_1	323715	431785
c869636e-895c-4f13-9ac0-a34f8ff03a22	chromosome_1	416134	549894
02ab0713-ad0f-457a-887e-cc29dd685bd7	chromosome_1	526114	667469
849096b1-06c4-4333-95b0-34fd0ddfcad3	chromosome_1	646247	777917
39bd608c-0c65-4f12-8dae-15cba2d1b6ef	chromosome_1	763020	902285
5755c8e4-5214-48e6-b9a5-0b39e07f2de9	chromosome_1	884700	1042388
e8584232-3ea1-4246-842b-1e77925b5906	chromosome_1	1010926	1112377
89027817-abae-46d7-9ec6-1ca675ccc304	chromosome_1	1096440	1217430
df5910c6-cfdd-4d5d-a603-154798bc42f0	chromosome_1	1181833	1303815
4971e5ad-03f1-4faa-ac8e-ed53be25871e	chromosome_1	1290299	1402507
9f1ff485-9261-4db4-8882-74d3c00f038d	chromosome_1	1351938	1532201
67ec7a07-dd6c-4e11-bb59-dfbd3728e92a	chromosome_1	1516334	1750877
adb55539-3eb6-4fee-8521-e09c88fd9574	chromosome_1	1689076	1801365
322f83be-3cbd-40cd-aa2d-a91413d6b004	chromosome_1	1774411	1909797
2a7f0757-9b65-40d3-a9e0-f91516ead697	chromosome_1	1895451	1989450
b7ae7052-843a-42cc-81bb-632a28e3a44d	chromosome_1	1949833	2086699
0d261bbb-37b8-46fc-af79-8671c6ecba59	chromosome_1	2051189	2182598
89077d2d-03a2-4298-8de0-de9ffe398dbb	chromosome_1	2152901	2302746
621035be-fe12-4ea3-ae3f-8148c464388b	chromosome_1	2260991	2402245

e975b3bd-c40c-4358-94be-4751a5bdd066	chromosome_1	2378661	2479980
c5b91fb2-a7fa-4810-ac73-d4884695820f	chromosome_1	2461612	2608159
3c694eb0-0f29-47b9-9b47-49698c992542	chromosome_10	23	151093
804a1af2-c165-4237-9fb3-7535e41e4878	chromosome_10	124996	240931
2ce972c5-b8c5-4b35-b9dc-62ed4b85c051	chromosome_10	221861	336680
e0cdee9e-847e-46b5-9a09-554f373c08b6	chromosome_10	325532	426973
404aefba-52a7-411d-8b30-3900c038bf8f	chromosome_10	413523	524727
67059b42-e049-40a5-87a2-a871550ce9a8	chromosome_10	511635	690826
253192b7-9745-4e56-a98f-525908b70ef6	chromosome_10	665132	765799
1bcc20da-8de6-47f7-b2da-9d5e54ca3bda	chromosome_10	750360	893299
560948ef-1ab8-4492-970d-2992852382ba	chromosome_10	853705	988919
e05e1e95-ecef-478f-8c1c-693f56e9b451	chromosome_10	924671	1107313
fd32ceba-f7c6-4006-96c1-d234a636bca9	chromosome_11	276	173234
08a17caa-c446-46d7-a24d-bef7fee5024d	chromosome_11	149183	266697
ac2aabd6-a8eb-473f-b6bd-1c25191ba7a6	chromosome_11	235927	350750
7435b6cf-e856-4236-ba08-8d233137a555	chromosome_11	333432	432324
e2a27806-c3f5-47e3-ad6d-1315a959d6a2	chromosome_11	419331	520375
8aa01b2b-b2a8-4b7f-8536-36927fd3da92	chromosome_11	470671	623243
556a3fa8-5808-42bf-9882-a5202b47770b	chromosome_11	611168	732271
23e3d146-7dee-45d6-acdf-80e9a322a459	chromosome_11	664359	817175
f7914d7e-3e9b-470f-8dcd-bd0743b689ba	chromosome_11	789847	938650
e502fff8-bb07-40c4-8116-82585a8fdf45	chromosome_11	927214	1087283
fd64b582-6dba-4e69-accd-08aa3449c7f4	chromosome_12	115	85562
8d898dfd-ae94-4d2b-b91f-35c8196c30ca	chromosome_12	70317	186817
48cd2f70-63de-493f-a135-cd6f793a17d9	chromosome_12	163852	259688
83f7a48e-ff07-4c95-8559-2d2addb65781	chromosome_12	245249	374713
46df773b-5108-45dd-9cc8-7474c3f50a52	chromosome_12	350096	473995
3b3f3a16-bf60-4751-8eb2-abb5c93d2461	chromosome_12	457625	619824

5aa3e23c-1236-4838-8aca-d9f96093a48c	chromosome_12	584213	730312
9a352b56-fbd6-4773-a0b6-747be5310b26	chromosome_12	707543	868869
9c312f33-3124-4d77-8208-c9f1b6105dc1	chromosome_12	829326	962049
ec3ac61f-c7b6-438b-b68b-badcbec7ac4a	chromosome_12	896670	1052199
5cecf314-a28b-4f56-ac87-11ff72199e76	chromosome_13	23	166678
c91b6704-5aa9-4f37-b2a3-3b32e1aae334	chromosome_13	149281	249725
cc95c95e-fa91-43fe-b3ff-6fdf23b561e5	chromosome_13	223099	333023
8a572e36-a7be-4111-9cdb-5cbe3aeb1179	chromosome_13	314088	452288
82e19b5e-0260-42eb-bf09-6d248a9743b5	chromosome_13	418815	567741
af15c4cb-ff70-4188-9cc7-7aaaa546d972	chromosome_13	535742	721913
ed4b6eaa-6636-4b94-9edd-4c225584ac6e	chromosome_13	691888	811121
8b8b6fe4-9ca9-48e3-86ea-6ae2dfd82486	chromosome_13	742389	959195
6b35a049-221c-49b2-8852-fa9f9c1681ba	chromosome_13	791947	959195
608aca5e-be1b-4ee4-81f9-1920a49949b1	chromosome_14	277	180092
79f4a2c1-a83e-4812-9dc7-d6629088a59a	chromosome_14	114805	299165
5364b93e-66c4-489b-b8ae-447b7dddfa8a	chromosome_14	282282	377464
695517e6-6acc-490e-b8bd-2fb89a068424	chromosome_14	321148	447701
cedd0312-f98d-4a7b-9b5a-f9ddcb28d697	chromosome_14	430009	532304
c3e3ea87-e08e-4c6d-b792-5593ff229c8d	chromosome_14	505657	630128
79222fe8-0183-496b-a49d-a96c49890b87	chromosome_14	601106	700867
6df530f0-c6e9-4d4e-9651-f33c1921325e	chromosome_14	665120	806315
b60a0834-e03e-4b59-9905-91a19d50a9bf	chromosome_14	783196	898569
a59ef52a-a591-4efd-bc8c-e31e85eea867	chromosome_14	783802	898569
0319f14c-ec17-44ee-b7b3-a91c86a714fb	chromosome_15	118	167265
aa3a32bd-5daf-4f27-b1cb-51341b1be694	chromosome_15	75021	223190
a3ac3cba-aad0-4db5-b8e1-f89e55d68006	chromosome_15	209461	312101
9540128d-9230-4419-8980-80906d8f2427	chromosome_15	286891	392930
d996585b-b3b2-4df7-b3fb-6a0947b46811	chromosome_15	376529	479502

a4f2784c-7181-4072-b02a-205fd124b918	chromosome_15	456881	566293
1cf06f06-3ee6-4312-845c-9796e88ef1f7	chromosome_15	541540	666635
816fd6be-24c5-4468-a237-835505893935	chromosome_15	635758	782946
00272977-eefb-4433-aef5-08b9875837ef	chromosome_15	740777	897220
5a78fdf2-e3ee-4c95-8a28-cd5973bebf84	chromosome_16	7	129040
2bbd1e91-0cc9-4173-8ca2-4bd29e4c46c8	chromosome_16	107457	222099
10486b3d-833a-4ac7-8a68-dc1c0d45b400	chromosome_16	209715	330236
37116640-302b-42d1-9c3a-24b06168dd0d	chromosome_16	313180	412904
e738bd89-543a-4175-a48e-318a7fb27584	chromosome_16	393694	485590
5a26ff32-e6ca-4e0f-b9b8-1ee817b49752	chromosome_16	469852	566970
e5d3879c-0897-4cf8-afe2-c73b4cae1355	chromosome_16	550311	664516
0618c102-a712-4db5-9536-e5ed32557dc8	chromosome_16	652132	799293
f2854750-a9a7-46de-bca4-fa31eb541001	chromosome_16	679626	860792
2d3cc462-ed7a-48e5-b644-190c2ec805aa	chromosome_17	13	153402
dfbeb09c-13e8-4dfa-b143-c35a21ced137	chromosome_17	121241	251502
14204815-c12a-4b80-955b-1b0a75e7019f	chromosome_17	238696	362986
6259d1ef-7094-4ee5-946b-46d5ef884597	chromosome_17	334304	462074
87b64719-cc61-4ec7-9057-ad13b8fcd82a	chromosome_17	428354	545134
5db7afba-16fa-4ee7-9f8d-4dd3f9a52613	chromosome_17	491315	622065
3b1a45a9-85a7-401d-bede-ff197de3426d	chromosome_17	601518	725862
382399c3-3c1d-445c-bb06-7c78ef5552a4	chromosome_17	646840	803256
7349c3af-22ec-4e8c-b058-e812252fc34a	chromosome_18	21	143486
6c7e2361-8089-4f6a-b017-10cd463a7f20	chromosome_18	129407	262616
9c585b96-ebdd-4ebc-a906-38e5912ee321	chromosome_18	228810	355086
2aa07c11-9128-43b7-ba0b-5c5ca1a57d52	chromosome_18	273058	436104
6d905d41-c836-4ead-b3b8-df78fd25bec8	chromosome_18	415063	568497
e44ae5fb-ec74-4130-a879-8978f87e8100	chromosome_18	538581	691109
6358442f-df16-4a55-9673-2f683e7abca5	chromosome_18	620587	759491

c1516352-fa7c-43ec-958a-df2a0a19da31	chromosome_19	254	209878
be10ef4a-e889-4e17-923e-79b804a0a38c	chromosome_19	189393	331315
5caed681-b5f7-41f3-ad2e-55fd53e1d65c	chromosome_19	275394	463038
518a6d52-2299-4d38-9fb8-3e5279500622	chromosome_19	443988	555670
84f0d0f5-0ea1-454a-a1f2-87942784f1f6	chromosome_19	525235	634173
56f4a83a-c65b-41ea-806d-8e170e4b8f22	chromosome_19	587246	716698
23b9a525-c8c8-4b82-818c-7ce8ef1444bf	chromosome_2	243	161521
c8a1f040-7f9d-48f2-bd99-27212a6d0673	chromosome_2	137112	273086
7fc19c4b-4cbf-4e6d-92a4-88e0fa0c6b19	chromosome_2	249533	384553
f9930e75-9d20-40e6-903d-3f8c95d550fd	chromosome_2	366838	508025
06a8c0c7-1ae2-410c-8472-6a7fe4a90191	chromosome_2	496715	628189
dd35d79e-6816-4273-9ed5-3d2f2d00c845	chromosome_2	584761	727489
a7f19f1a-e9ca-4931-8932-6501ad8e6a99	chromosome_2	710715	809649
6a879493-28a2-4cbd-8976-017c21d23853	chromosome_2	784055	863988
53d86224-746a-44a6-bff0-c59ecb98ce93	chromosome_2	794010	863988
0cd900f1-0e72-44c5-bb76-e77e1d8430e2	chromosome_2	836457	979102
e0f36c4d-2853-45f8-bc8d-658376ec2670	chromosome_2	968281	1110417
276c9d00-265c-44f2-85fe-65e985c819be	chromosome_2	1072604	1207483
8fa9f46f-0ef3-46e8-8dba-40e76132539d	chromosome_2	1195322	1336141
9b419501-2703-4c05-ab37-40cdf8142d24	chromosome_2	1299544	1418016
515daccf-c98f-49fa-8fe4-21286f0325e4	chromosome_2	1384896	1527963
56bebddb-9f72-4383-8f54-31435ac99a00	chromosome_2	1509554	1634366
0043feb7-6d31-4cdf-b86a-2b16db1e7dcc	chromosome_2	1602716	1733943
67d56bf8-2e92-4f58-bdd3-2c569f048741	chromosome_2	1720401	1824441
e613e658-284b-471e-b942-5c6b7e93e69e	chromosome_2	1791377	1982710
3fe7b8f6-cfda-4fc3-ae32-d56778ff967d	chromosome_2	1874866	2059771
87f48e33-5c0d-47ef-9201-b6480d7d460c	chromosome_2	2043847	2186350
8679c0da-2852-450e-93ae-c1af1af6dfc4	chromosome_2	2171292	2285476

6dba74de-68be-49d1-bf4a-754ebfc1b2ca	chromosome_2	2261185	2499617
bab5401c-a984-41a1-8640-e1c669ac363a	chromosome_20	8	132179
d7a2a8ae-f903-450a-98c6-46cb085fd032	chromosome_20	108717	219383
c48e15cd-61c2-43d4-98de-de54b8ee8bde	chromosome_20	208378	350723
95ff1dcb-f261-42d6-b884-99618a6cd75c	chromosome_20	340592	470209
ee4a80bb-a2fd-41ab-bb10-02c3e363c56a	chromosome_20	459898	562358
a05e7863-5a4e-4a87-b72b-d2b67ea37ff0	chromosome_20	547442	639179
9cffd23c-b5d4-4a1e-a953-8457b0fe9064	chromosome_20	588436	709261
156c4bed-e40f-445c-b302-8fe3818efeec	chromosome_20	591736	709261
66e9ee60-4578-4690-8419-274e02afb64f	chromosome_20	606085	709261
2659e7bc-d427-4365-ad41-b5b16eb947f0	chromosome_21	7	121509
68b1c0bd-5cf9-41f7-a2e8-1961bdd8ebc9	chromosome_21	110024	256071
4e31ae16-aa15-4ec3-b9de-eb4701f28c10	chromosome_21	213813	351018
c80d1aef-ef18-4bb8-ae78-8e1c6677db36	chromosome_21	338324	470014
f5519510-52ca-4b51-bcba-45ca83059715	chromosome_21	454386	518226
65c32665-1e9c-4bbb-b874-a517d677684d	chromosome_21	454409	518226
967f9512-5117-4fe6-86fd-97ab1fb467b1	chromosome_21	502558	629488
6642ab40-f256-4bde-b187-a02417c1c863	chromosome_22	6	155427
d99099d1-633c-4e7c-aabe-f972e0ecaa4b	chromosome_22	108837	255877
5184a123-581c-467c-a086-c48325fdca40	chromosome_22	245259	338430
cd62b7fc-5073-449d-9981-e5f6fa45a851	chromosome_22	321877	469799
4aca853a-4e19-4c9f-83e8-eabfc691c64f	chromosome_22	455831	539480
411ef6c2-9f98-4092-8297-a8c1c29ea80f	chromosome_22	528371	587838
4a1587b0-3750-43a7-bbfb-28b86d81c188	chromosome_22	529096	587838
f86f7aee-7313-4bd7-b015-f2a480d0a0f5	chromosome_23	237	144930
7740e20d-efc1-4b0f-9583-3b8858139b28	chromosome_23	105509	269152
fd61220d-2a9d-4320-a684-a26b46cea5ed	chromosome_23	257295	352483
6931b23b-211a-4df5-b219-7bf9f64fdc70	chromosome_23	301884	444163

8ed4a931-e2f9-4424-8bd0-299e3213c9c6	chromosome_23	428963	557587
bae7ee75-28fc-46a3-9e0a-bfda6c39aa78	chromosome_24	225	173784
7aaf5650-91e5-4fb2-9338-2dfd09341fa5	chromosome_24	151555	271818
3e6b469c-6e92-44c6-ba3c-e77629987c8b	chromosome_24	240823	336942
906934e0-045c-4704-81bb-ae480eb747b5	chromosome_24	302376	428307
8ee4821e-e794-4523-a5de-e91829a73689	chromosome_24	403399	546597
69f4cadb-db35-447a-a4b0-5ed9f99203b5	chromosome_25	7758	176334
96a3e0a6-44a1-4676-a4dc-b278bad22309	chromosome_25	163319	284218
799baf0e-eb00-4eb3-8bd5-e928a6c71c5d	chromosome_25	163322	284218
486066ef-803e-4f49-a12c-2cca949613be	chromosome_25	264162	372482
e468c1f4-8144-48e9-a57a-561dfc2e1022	chromosome_25	330356	442942
59eddacf-6bb2-42c7-a656-033304284c51	chromosome_25	393239	516877
331c489e-1073-48ab-bd89-f2e56147e81b	chromosome_3	324	192911
989455aa-29ef-4ca3-9566-656c89fce083	chromosome_3	169538	287647
1c7ad41d-1d78-471c-b1a5-7f54fecfb2db	chromosome_3	262693	357246
26910f91-25e2-49f5-b4ab-bd42bd364987	chromosome_3	305563	438328
d06f6d72-52fb-41ff-bcef-e6de97bbc139	chromosome_3	424722	567980
150f276a-3a9b-4447-94d7-22df86727c6c	chromosome_3	547991	685015
4a4d2a61-37ad-4beb-ae52-06e017030635	chromosome_3	666113	798674
489656b9-fa00-470a-8b05-af6bb02457fd	chromosome_3	773989	921175
42178f6f-bdf7-4864-b09e-da4c7f2b00e4	chromosome_3	895386	1002884
7a9aa205-26dd-45b2-9561-d451554a3f74	chromosome_3	975675	1082935
e99d32d3-3152-48d6-b700-b84ea2f1e7ca	chromosome_3	1055343	1200920
1d01e563-833f-4163-82a6-955068c223c2	chromosome_3	1174330	1298509
ef3e1328-4490-47b2-b958-8b74cef3d134	chromosome_3	1276295	1446243
6b799515-fb66-446d-8cd1-8b710a276f84	chromosome_3	1428129	1511230
304f93ae-9e44-475f-93ed-f4ddf8733dc8	chromosome_3	1491392	1633015
299f73ef-5c1d-433a-9f9e-fd0dd6450057	chromosome_3	1617255	1718061

503eca2a-ba25-480c-9c5b-ae1bb2cc9c35	chromosome_3	1699799	1782530
38d4c46b-9383-43b6-9f51-bcbd6b02ccbb	chromosome_3	1746039	1854171
1688cfd1-a4b7-4a78-bf28-0d61a13730d8	chromosome_3	1836150	1951814
0b5abe69-5031-41e1-82ca-2cf43ea42b98	chromosome_3	1940830	2064735
f43955a0-1ecd-40c5-a513-b8c1ad9aadc9	chromosome_4	68	129623
3f861d0e-c22c-42fc-9731-007ab0616106	chromosome_4	115433	240549
7890e17f-1958-4d31-a4e7-83f54c70ff29	chromosome_4	208978	361937
8f0697a4-556c-4228-9638-98cc7a72441e	chromosome_4	325057	431775
b9d51e62-a489-4b13-bc0e-fc697f336ed3	chromosome_4	408636	536604
22fabd25-2ae0-4309-973f-d35ff731ac44	chromosome_4	508106	631124
48b23dca-6b77-496f-84a1-09ab05603bd3	chromosome_4	600919	735413
a91d8ff2-9a94-4c3a-bf30-2e7677b01260	chromosome_4	663785	817344
5ff9a519-7760-4e45-bc0f-9fe4b6c5a2d4	chromosome_4	806649	912523
a3804d2f-21b6-4c56-8daa-6cd87b3378bf	chromosome_4	897405	1000693
eeb6abdb-7938-4c76-aec3-615439be53b5	chromosome_4	989307	1064558
cfd804bc-0cc4-4eb1-8730-decffafdee66	chromosome_4	1044495	1167527
d7b21e6b-093d-4b69-8919-11b9b26935bc	chromosome_4	1152362	1290587
c8b3ce54-b3ac-456c-b15b-0a477ddab691	chromosome_4	1278393	1461254
a689a14d-d655-40dd-9a10-09caf2edef58	chromosome_4	1450998	1525394
0a6beea3-a5a4-4e8b-aa99-ce66c66c1db8	chromosome_4	1488013	1629129
da6c68c8-cf05-4e44-af50-eb803bf37467	chromosome_5	118	134742
3ef61844-89bb-4ed6-9999-8b60b0ed63ec	chromosome_5	102255	231687
7c296fd4-5c14-4b24-93e4-463d6eaf2212	chromosome_5	211385	365477
ef1d3ae4-bdad-40cd-8eef-c0bf7754b8fd	chromosome_5	283562	468306
2ce5bf92-3ddd-4de8-9798-f78fa1d8805a	chromosome_5	435592	594126
5a59f3db-a8ac-4671-82d6-d6aa9f87f827	chromosome_5	576008	715463
cd331167-4d49-4360-82be-c734beeb22ce	chromosome_5	675228	816374
7e3ce048-a6b2-4d1d-8338-066ee321601f	chromosome_5	769522	916733

6094ccd3-296b-49cd-804a-a09d49f47d5e	chromosome_5	893755	1003331
5f6855ed-19cd-4a36-b901-bbec0f5f9d41	chromosome_5	987604	1112477
a13425cc-c991-454b-b1bd-3fa08b5549c6	chromosome_5	1072370	1210018
58a500e7-18f9-4bac-9a30-ddb316951da7	chromosome_5	1199633	1316915
eac54856-bed7-478f-8e60-cba6610e6bd8	chromosome_5	1299262	1439487
c55cdef9-aa9b-4bca-8664-18e0b026daa2	chromosome_5	1411315	1554809
90f6f461-64dd-46c9-b660-3fb0301d0055	chromosome_6	158	148399
21a36c85-aeb3-4add-98ec-74dc76993476	chromosome_6	135971	245763
1ea18722-9f12-4ef2-a8d9-1df2bfd5ba32	chromosome_6	214225	328547
3e888b72-8c5c-40c9-ae2d-e14b4b0cd947	chromosome_6	314122	398982
9f37b53f-d299-4e07-a828-72656f347167	chromosome_6	383792	499382
82b941dc-e029-470b-b4ba-00dcf69684ac	chromosome_6	458264	576468
7da496c6-a104-4f8c-b836-6214bbba0b07	chromosome_6	522287	714888
ca3e7f07-507f-4a87-99d6-763a1073d67c	chromosome_6	681100	857745
7de32eca-52e3-42cb-abd6-9009a0743e0f	chromosome_6	846967	941407
0a14ee38-302f-4bab-b12c-90c2db22bd24	chromosome_6	914867	991437
12e72327-cd41-44a0-b3e8-8ea3fb347f9f	chromosome_6	961535	1099373
cd79b6dd-f4ac-459c-9203-4b2ffa30cad0	chromosome_6	1076731	1185590
c16adc6b-d1c8-4a3a-8b80-7aa0a68f3103	chromosome_6	1154576	1273426
71072cfc-246c-474f-884f-fe2e44360243	chromosome_6	1223399	1394473
c25a15e8-5bf2-405b-8402-f172cd7105fa	chromosome_6	1301551	1417080
57b77647-a9cf-4c65-b838-b56144fa9999	chromosome_7	102	181816
2f7ab6b0-dfc4-4968-8996-9c20a33568f0	chromosome_7	150208	262514
4c085c96-2ac8-48f3-8f82-88d4297a9098	chromosome_7	237176	373499
e54137c4-868a-45c6-8519-a57a0ba38b55	chromosome_7	338738	486183
6c60aed9-e47c-4a8c-9123-f059e3db3f77	chromosome_7	460869	592894
717dffb9-5a8d-4353-8e45-bdccfd661ef5	chromosome_7	541219	705237
296e26a3-57d3-4ec7-b24b-30ab172fb363	chromosome_7	690701	798913

e1253bbe-7e77-4789-85cf-2c73c5269aef	chromosome_7	760234	932493
a1792b82-69c6-4281-8070-56a4a5fcf1af	chromosome_7	886385	1019103
15bf5e80-7a75-4eb6-a2f9-0831d9e22ffe	chromosome_7	988285	1124615
35c95a35-025d-4cb6-be29-12fd5cf1b238	chromosome_8	9	144388
b4fef834-54e4-47ad-852d-f3fed8a4b6af	chromosome_8	128582	231423
da370162-9ec4-49c0-b6ce-6f6fdfc56156	chromosome_8	209735	315450
535364c8-72a8-478a-829c-5cd4a9962db8	chromosome_8	278337	410214
70e146fa-04f7-4bb0-a3c4-2e147888a73d	chromosome_8	365211	506223
5373a7be-8809-46ad-8cbc-12ff47b53963	chromosome_8	451560	597661
af778c23-bc26-473f-85b1-bfa297bb6987	chromosome_8	580511	664133
5e5674d0-d15b-4631-bf79-a57a86a6665f	chromosome_8	602139	762161
a24018a0-1567-426a-9917-d43ee341d957	chromosome_8	740307	820291
b06b35f9-3c38-485c-be77-d1c7db96ddb3	chromosome_8	805390	927686
98415a1a-fc70-4ed8-82eb-68af65ddc9f4	chromosome_8	910488	1007230
da19e7d2-5e90-4ff2-8ad7-8001f6c9d30d	chromosome_8	976850	1122323
31de7c71-8d2c-4a11-a11b-254f8541732b	chromosome_9	3	157978
8f6b2e24-e37a-4e55-84e7-165522d8897e	chromosome_9	147876	257744
de180526-e9d1-429d-b299-9ccbc2f57d24	chromosome_9	241949	339822
762b011e-0f94-4593-bbe3-98d7439e3d07	chromosome_9	316334	398015
2eb3ab7a-1050-49be-91a9-2539f11b4941	chromosome_9	384735	462047
94638ecc-70b2-41fc-b3cd-dfef01ec6cf8	chromosome_9	442317	620200
78202699-32ac-41af-9ae5-743b46f3fa96	chromosome_9	576141	744220
c6729300-e66a-46df-8256-e5f4c0f66bf9	chromosome_9	731241	858221
0460d05e-2ad0-4e93-8862-2790f89ded4f	chromosome_9	831034	944725
a101f7f1-7984-4102-a049-619767540fb2	chromosome_9	934696	1065587
d27d72e6-a9a5-44e5-ad39-aa527000bae9	chromosome_9	968928	1108070

Table B.2: Quality value estimates obtained from Merqury and Illumina reads

Chromosome	k-mers (assembly only)	All k-mers found	QV	Error rate
Chromosome 1	58052	2597765	28.7662	0.00132855
Chromosome 2	54615	2507151	28.8782	0.00129472
Chromosome 3	52103	2057091	28.2161	0.00150797
Chromosome 4	41719	1629113	28.1678	0.00152484
Chromosome 5	39504	1555004	28.2029	0.00151254
Chromosome 6	43532	1417141	27.3669	0.0018336
Chromosome 7	31066	1124607	27.8345	0.00164644
Chromosome 8	30254	1122370	27.9423	0.00160609
Chromosome 9	30223	1110319	27.8993	0.00162206
Chromosome 10	23142	1107373	29.0603	0.00124156
Chromosome 11	15495	1087430	30.7373	0.000843861
Chromosome 12	21797	1052218	29.0989	0.00123059
Chromosome 13	22976	959307	28.4619	0.00142499
Chromosome 14	21001	898560	28.5693	0.00139017
Chromosome 15	24054	897214	27.9662	0.00159729
Chromosome 16	24998	860814	27.6145	0.00173203
Chromosome 17	26223	803240	27.0985	0.00195053
Chromosome 18	20184	759555	28.0051	0.00158303
Chromosome 19	659	716913	42.6684	5.40951e-05
Chromosome 20	15323	709249	28.9145	0.00128396
Chromosome 21	14661	629742	28.5864	0.0013847
Chromosome 22	15406	587823	28.0659	0.00156102
Chromosome 23	18157	557573	27.1096	0.00194554
Chromosome 24	15073	546827	27.844	0.00164286
Chromosome 25	12389	516868	28.4584	0.00142612
Chloroplast	9	117354	53.4569	4.5114e-06
Mitochondrion	110	92768	41.5621	6.97892e-05

Curriculum Vitae

Daniel J. Giguere
Department of Biochemistry
University of Western Ontario
London, Ontario N6A 5C1

Education:

PhD Candidate, Biochemistry, 2017-current
Western University, London, Ontario

Master of Science, Biochemistry, 2016-2017
Western University, London, Ontario

Bachelor of Science, Biomedical Science, 2012-2016
University of Ottawa, Ottawa, Ontario

Publications in press:

Brumwell, S.L., Van Belois, K.D., **Giguere, D.J.**, Edgell D.R., Karas B.J. Conjugation-based genome engineering in *Deinococcus radiodurans*. ACS Synthetic biology, 2d28ed6c-1eec-43b5-8f20-46e4872c2d56 (Submitted November 19 2021, accepted with major revisions December 11 2021).

Giguere, D.J, Bahcheli, A.T., Slattery, S.S., Patel, R.R., Flatley, M., Karas, B.J., Edgell, D.R., Gloor G.B. Telomere-to-telomere genome assembly of Phaeodactylum tricornutum. PeerJ, Reference 60781 (Accepted with minor revisions, Submitted May 10 2021).

Publications submitted for peer review:

Slattery, S.S., **Giguere, D.J**., et al. Phosphate-regulated expression of the SARS-CoV-2 receptor-binding domain in the diatom *Phaeodactylum tricornutum* for pandemic diagnostics. Scientific Reports, 2d28ed6c-1eec-43b5-8f20-46e4872c2d56 (Submitted November 19 2021).

Peer-reviewed publications:

Slattery, S.S., Wang, H., **Giguere, D.J**. et al. Plasmid-based complementation of large deletions in Phaeodactylum tricornutum biosynthetic genes generated by Cas9 editing. Sci Rep 10, 13879 (2020). https://doi.org/10.1038/s41598-020-70769-6

Cochrane, R.R., Brumwell, S.L., Shrestha, A., **Giguere, D.J.**, Hamadache, S., Gloor, G.B., Edgell, D.R., Karas, B.J. Cloning of Thalassiosira pseudonana's Mitochondrial Genome in Saccharomyces cerevisiae and Escherichia coli. Biology. 2020; 9(11):358. https://doi.org/10.3390/biology9110358

Berg, M.D., **Giguere, D.J.,**, Dron, J.S., Lant, J.T., Generaux, J., Liao, C., Wang, J., Robinson, J.F., Gloor, G.B., Hegele, R.A., O'Donoghue, P., Brandl, C.J., (2019) Targeted sequencing reveals expanded genetic diversity of human transfer RNAs, RNA Biology, 16:11, 1574-1585
DOI: 10.1080/15476286.2019.1646079

Giguere, D.J., Macklaim, J.M., Lieng, B.Y., Gloor, G.B. omicplotR: visualizing omic datasets as compositions. BMC Bioinformatics. 2019 Nov 15;20(1):580. doi: 10.1186/s12859-019-3174-x.

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:

Modern DNA sequencing and applications in synthetic biology. Biochemistry 3392, Virtual. January, 2022 (invited lecture)

Completing the *Phaeodactylum tricornutum* genome with ultra-long sequencing reads. Great Lakes Bioinformatics Conference, Virtual. May, 2021 (abstract/poster)

Modern DNA sequencing and applications in synthetic biology. Biochemistry 3392, Virtual. February, 2021 (invited lecture)

In-house full plasmid sequencing to enable rapid synthetic biology. Maud L Menten Fall Symposium, London, Ontario. December, 2020 (invited speaker)

Characterizing naphthenic acid degrading bacterial communities. 69th Conference of the Canadian Society of Microbiologists, Sherbrooke, Canada. June, 2019 (abstract/talk)

omicplotR: A Shiny app for exploring omic datasets as compositions. Bioc 2018: Where Software and Biology Connect, Toronto, Ontario. July 2018 (abstract/poster)

Transcriptional response of C. difficile to bacteriotherapy. 7th Annual Course on Compositional Data Analysis, Girona, Spain. July 2018 (talk)

Awards and Honors:

Award	Value	Level	Period Held
MITACS Accelerate - Lab2Market	$15,000	National	2021/05-2021/09
Department of Biochemistry Best Talk	$50	Departmental	2020/12
Ontario Gradudate Scholarship	$15,000	Provincial	2020/09-2021/09
MITACS Accelerate Internship	$15,000	National	2020/05-2020/09
Ontario Gradudate Scholarship	$15,000	Provincial	2019/05-2020/05
Department of Biochemistry Best Poster	$50	Departmental	2020/05
Ontario Gradudate Scholarship	$15,000	Provincial	2018/05-2019/05
Canadian Society of Microbiology Travel Award	$250	National	2019/05
Direct Entry PhD Scholarship	$10,000	Departmental	20179/09
Western Graduate Research Scholarship	$32,333	Institutional	2017/09-2021/12

Experience:

Extra-curricular

Lab2Market Entrepreneurial Program, 2021

Biochemistry Graduate Student Association Co-Chair, 2020-2021

Biochemistry Graduate Studies committee student representative, 2018-2019, 2020-2021

Society of Graduate Students Finance Committee Co-Chair, 2018-2020

Teaching Assistantships

Biochemistry 3380G, Western University 2020

Biochemistry 3390, Western University 2019

Biochemistry 3380G, Western University 2018

Biochemistry 3380G, Western University 2017

Biochemistry 2280A, Western University 2016

Teaching Assistant Training Program 2016

Society Memberships

Canadian Society for Microbiologists, 2019-present

International Society for Computational Biology 2019-present

Journal reviews

NAR Genomics and Bioinformatics *ad hoc*, 2020