Norwegian University
of Life Sciences

Master's Thesis 2020 60 ECTS
Faculty of Biosciences

Exploring current practices and the potentials of Nanopore sequencing in metagenomics

Tina Johannessen
M-LUN

Acknowledgements

This thesis was written for the faculty of Biosciences at the Norwegian University of Life Sciences (NMBU) with Associate Professor Phillip B. Pope as the main supervisor and Associate Professor Simen R. Sandve as co-supervisor.

Firstly, I would like to thank my primary supervisor Phil, for providing the opportunity for me to work on this thesis, for insightful feedback and for always having a plan. I would also like to thank Dr. Live Heldal Hagen, for kind and patient guidance both in and out of the lab. Thank you, Dr. Sabina Leanti La Rosa, for allowing us to use your samples for the nanopore sequencing and sharing your work with us. To everyone in the Protein engineering and Proteomics group (PEP), thank you for being so welcoming, for sharing knowledge and for every-day conversations.

Thank you to all of my fellow master students, for sharing both successes and failures, and for meaningful discussions. A special thank you to my working partner for this thesis Alexsander Lysberg, without whom my days in the lab would have been much lonelier, for always finding new and terrible jokes to make me laugh, and for your constant companionship.

Finally, thank you to all of my friends and family for supporting me through this past year's ups and downs, your words of encouragement have meant the world.

Ås, July 2020

Tina Johannessen

Abstract

For billions of years, microorganisms were the sole inhabitants of planet Earth. As major drivers behind many essential geochemical cycles such as the carbon cycle, microbial communities are integral to the continued support of life on Earth and are found in everywhere from the deep seas to our own bodies.

In 1977, Frederick Sanger introduced a new method for determining the nucleotide sequences of DNA by chain-terminating inhibitors. This method later become known as Sanger sequencing and would go on to dominate the field for the next 30 or so years. In the mid-2000s, the development of high-throughput sequencing technology led to a revolution in microbial ecology. Often referred to as next-generation sequencing, these technologies were capable of generating tremendous amounts of data at much lower costs per sequenced base than traditional sequencing. This technology, however, rarely produces sequences above a few hundred bases in length, and thus genomes have to be reconstructed by piecing the small fragments back together like a jigsaw puzzle. As most genomes contain many repetitive regions of varying lengths, this reconstruction called assembly often cannot fully reconstruct the original genome due to the inability of short reads to resolve repeated sequences, and in response to this a third generation of sequencing is now on the rise, promising read lengths measured in kilobases and real-time output.

Continued improvements in sequencing technologies has allowed researches to study the function and structure of microbial communities in great detail. Through metagenomics, a culture-independent technique that directly investigates the DNA isolated from an environmental sample, the study of hard to cultivate species from a host of high-interest niches is now possible.

Sammendrag

I milliarder av år var mikroorganismer de eneste beboerne på planeten. Som viktige drivere bak mange geokjemiske sykluser slik som karbonsyklusen, er mikrobielle samfunn helt essensielle for fortsatt liv på jorda og er å finne overalt, fra havdyp til våre egne kropper.

I 1977 introduserte Frederick Sanger en ny metode for å sekvensere DNA ved hjelp av kjede-terminerende inhibitorer. Denne metoden ble senere kjent som Sanger sekvensering, og ville fortsette å dominere sitt felt i de neste 30 årene. På midten av 2000-tallet førte utviklingen av sekvenseringsteknologi med høy gjennomstrømning til en revolusjon i mikrobiell økologi. Disse teknologiene, ofte kalt neste generasjons sekvensering, var i stand til a generere enorme mengder med data til mye lavere pris per sekvenserte base en tradisjonelle metoder. Teknologien produserer sjelden sekvenser lenger enn et par hundre baser i lengde, og derfor måtte genomer rekonstrueres fra små biter som i et puslespill. Siden de fleste genomer inneholder repetitive sekvenser av varierende lengde, byr dette på problemer i monteringen («assembly») av genomer, siden de korte sekvensene ikke klarer å løse opp i disse. Som svar på dette er det nå en tredje generasjon av sekvensseringsteknologier som begynner å gjøre seg kjent, som lover lengre sekvenser, målt i kilobaser og sanntidsdata.

Fortsatt utvikling av sekvenseringsteknologi har tillatt forskere å studere funksjon og struktur hos mikrobielle samfunn i detalj. Gjennom metagenomikk, en kultur-uavhengig metode som direkte undersøker DNA fra en miljøprøve har nå tidligere ukultiverbare arter fra en rekke nisjer av høy interesse blitt mulig.

Abbreviations

A_{260}, A_{280}, A_{230}	Absorbance at 260 nm, 280 nm and 230 nm respectively
ASV	Amplicon Sequence Variant
bp	Base pairs
BSA	Bovine serum albumin
CAE	Capillary array electrophoresis
CAZyDB	The carbohydrate-active enzymes database
CAZyme	Carbohydrate active enzyme
CCS	circular consensus sequence
CLR	continuous long reads
contig	contiguous sequence
CTAB	cetyltrimethylammonium bromide
dNTP	Deoxynucleotide triphosphate
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
dsDNA	Double stranded DNA
E.C.	Enzyme Commission
HMP	Human microbiome project
HMW	High molecular-weight
kb	Kilobases
KEGG	The Kyoto Encyclopedia of Genes and Genomes
MAG	Metagenomic assembled genome
mV	Millivolts
NGS	Next generation sequencing
nm	Nanometers
OLC	Overlap layout consensus
ONT	Oxford Nanopore Technologies
ORF	Open reading frame
OTU	Operational taxonomic unit
PacBio	Pacific Biosystems
PCR	The polymerase chain reaction
RNA	Ribonucleic acids
SDS	Dodium dodecyl sulfate
SMRT	Single Molecule Real-Time
SOLiD	Oligo Ligation Detection
ssDNA	Single stranded DNA
Tb	Terabases
UV	Ultraviolet
ZMW	Zero-Mode Waveguides

Table of contents

Introduction

1.1 Background

Microorganisms are a fundamental part of life on earth and integral for several geochemical processes. The study of these microbial communities is therefore an essential part of understanding the natural world (Milanese et al., 2019). Historically, microorganisms were studied through culturing in the lab, and it was believed that unculturable organisms could not be classified. For many years, studies of microbial ecology operated on the premise that unless a microorganism could be cultured, it did not exist. By the mid-1980s however, it had become apparent that the diversity of microbial life was much higher than first anticipated, and that the vast majority of species were in fact unculturable (Handelsman, 2004). Microorganisms are found in just about every environment possible, and many have developed symbiotic relationships with larger multicellular life by colonizing the different parts of their hosts, such as mucosal membranes in animals. These symbiotic organisms make up what is known as the microbiota, which is defined as the sum of all microorganisms living within a host or in/on a specified of said host, and their combined genomes are referred to as microbiomes (Jun Wang & Jia, 2016). Of all host-associated microbiomes, those residing in the gastrointestinal tract have garnered the most attention, as these represent the most dense and diverse populations, often outnumbering the host both by number of cells and by number of genes. These microorganisms have been shown to be essential to host biology where they play an important role in the development of the immune system, aiding in metabolism by degrading otherwise indigestible polysaccharides and offering protection against pathogens (Sommer & Bäckhed, 2013).

Due to their impact on health and development, microbiomes of both humans and livestock are subject to many studies. Modern agriculture faces two major challenges in the form of growing populations and climate change. Herbivorous livestock like ruminants are important to global food security as these are capable of producing meat and dairy of high nutritional value from complex carbohydrates (Seshadri et al., 2018). This conversion of biomass is made possible by the rumen microbiome, which is a highly complex and diverse microbial community comprised of bacteria, archaea, fungi, protozoa and phages that ferment indigestible plant biomass into short-chain fatty acids which in turn can be utilized by the host (Stewart et al., 2019). As land constraints limit the capacity for increased ruminant numbers, and efforts must therefore be made to increase the efficiency of present production to meet rising demands. Understanding the underlying mechanisms of microbial lignocellulosic biomass degradation may therefore play an important role in the development of future ruminant production (Huws et al., 2018). In addition, methane, a potent greenhouse gas, is a common byproduct of ruminant production. The fermentation process leads to the production of hydrogen gas which is subsequently utilized by methanogenic archaea to reduce carbon dioxide into methane. However, the relative abundance of these methanogens in the rumen have been found to be closely linked to the level of methane production, thus indicating a potential to reduce methane emissions through manipulation of the rumen microbiome (Wallace et al., 2015).

Unlike the herbivores, omnivores such as humans derive less of their total energy from their respective microbiota, however, these communities still hold great importance for the health of the host (Flint, Bayer, Rincon, Lamed, & White, 2008). The human body is home to what is estimated to be trillions of microbial cells, consisting of bacteria, archaea, and eukaryotic microbes as well as both eukaryotic viruses and bacteriophages. In 2007 the human microbiome project (HMP) was launched in an attempt to characterize and understand the influence of the microbiome on health and disease (Proctor, 2011; Turnbaugh et al., 2007). Of all microbiomes associated with the human body, the gut

represents the most dense and diverse community, with an estimated 100 trillion cells and 5 million genes, the structure of which has been shown to vary with both host genetics, age and environmental factors such as diet (Spanogiannopoulos, Bess, Carmody, & Turnbaugh, 2016). Several complex diseases such as diabetes 2 and obesity as well as some forms of cancer have all been associated with the microbiome (Jun Wang & Jia, 2016).

In addition to their effects on the health and metabolism of their hosts, microbiomes represent a large reservoir of enzymes of significant economic interest. This is especially true for cellulose degrading communities such as those found within the gastrointestinal tract. These populations display some of the most rapid natural rates biomass decomposition, and there is therefore considerable interest in mining these microbiomes for enzymes that may be used for biotechnological applications such as the production of biofuels from renewable plant sources (Baldrian & López-Mondéjar, 2014; Flint et al., 2008).

1.2 Carbohydrate digestion by microbes

1.2.1 Lignocellulosic biomass

Lignocellulosic biomass is the most abundant organic compound on Earth, and consists of mainly cellulose, hemicellulose and lignin (Moraïs et al., 2012). Along with pectin, these polymers are the main components of the plant cell wall (Gibson, 2012).

Cellulose is a linear polysaccharide composed of monomers of D-glucose linked by β-1,4-glucosidic bonds. Due to the linearity of the molecule, hydrogen bonds can be formed both within and between adjacent chains, forming a crystalline structure, making it mostly insoluble and difficult to hydrolyze (Jørgensen, Kristensen, & Felby, 2007; Malherbe & Cloete, 2002). In contrast, hemicellulose is a heteropolymer, and is made up of several monosaccharides, such as glucose, mannose and xylose (McKendry, 2002). Hemicellulose has an amorphous structure, and is generally less polymerized than cellulose, with chain lengths in the range of 500-3000 monomers. Xyloglucans, xylans, glucomannans and galactoglucomannans are all examples of hemicellulose (Gibson, 2012). Lignin is also an amorphous polymer, consisting of several aromatic compounds called phenyl-propanes (Jørgensen et al., 2007; M. Li, Pu, & Ragauskas, 2016). Pectin is another heteropolysaccharide: it has a high content of galacturonic acid, but it may also contain as many as 17 different monosaccharides (Mohnen, 2008; Voragen, Coenen, Verhoef, & Schols, 2009).

The exact composition of the plant cell walls varies with different plant types, tissue types and stages of development, but are generally comprised of cellulose chains embedded within a matrix of hemicellulose, lignin or pectin and a number of proteins (Flint et al., 2008). This matrix combined with the relative recalcitrance of its separate components make plant cell walls notoriously difficult to degrade (Lynd, Weimer, Van Zyl, & Pretorius, 2002).

1.2.2 Microbial digestion of carbohydrates

Insoluble substrates like those found in the plant cell wall are largely indigestible to most animals (Russell, Muck, & Weimer, 2009), and are mostly degraded by microorganisms living in the soil, or in the gastrointestinal tract, thus making accessible the highly stable, fixated carbon in these compounds, and closing the loop of the carbon cycle (Lynd et al., 2002).

Enzymes and other proteins involved in either assembling, modifying or breaking down oligo- and polysaccharides are collectively referred to as carbohydrate active enzymes, or CAZymes. The

carbohydrate-active enzymes database (CAZyDB) is a comprehensive, specialized database dedicated to characterizing these enzymes (Lombard, Golaconda Ramulu, Drula, Coutinho, & Henrissat, 2014) that are divided into families based on their amino acid sequence, structure and enzymatic mechanisms. These families include: glycoside hydrolases, polysaccharide lyases, carbohydrate esterases, glycosyltransferases, carbohydrate binding modules, and auxiliary activity enzymes (Levasseur, Drula, Lombard, Coutinho, & Henrissat, 2013).

Microorganisms utilize CAZymes in a variety of different ways. For example, *Ruminococcus flavefaciens*, a cellulose degrading anaerobic bacteria living in the rumen can for complex structures called cellulosomes comprised of multiple catalytic, structural and substrate binding domains. These complexes allow for close contact between the substrate and enzymatic machinery necessary for severing the linkages of polysaccharides bound in the plant cell wall as well as preventing diffusion of products away from the cell by forming a scaffolds docked to the cell surface (Flint et al., 2008). Further approaches involve the secretion of free enzymes directly into the environment that catalyze the breakdown of polysaccharides which can be readily absorbed for further degradation and polysaccharide utilization loci, a cluster of genes that encode enzyme systems associated with the cell envelope that facilitates the response, ability to bind to and degrade glycans and import the freed oligosaccharides (Naas et al., 2014). The abundance of these machineries in natural cellulolytic degrading microbiomes are an integral part of the natural carbon cycle, the understanding of which is essential in relation to the effects of global climate change. Therefore, culture-independent methods such as metagenomics (Section 1.4) are often used to examine the full enzymatic potential of microbial communities (Baldrian & López-Mondéjar, 2014).

1.3 Sequencing technologies

1.3.1 Gel-based methods

In the mid-1970s, several new methods for DNA sequencing were developed. The Nobel Prize in Chemistry was in 1980 awarded to Paul Berg, along with Frederick Sanger and Walter Gilbert for their work with nucleic acids, and in the case of the two latter, particularly the development of new sequencing methods (https://www.nobelprize.org/prizes/chemistry/1980/press-release/). Both Sanger's method of chain-terminating inhibitors, and the Maxam-Gilbert chemical cleavage method used gel electrophoresis to separate fragments by size, allowing the sequence to then be read off the gel. (Maxam & Gilbert, 1977; Frederick Sanger, Nicklen, & Coulson, 1977)

The Maxam-Gilbert method of sequencing takes a fragment of single stranded DNA, labeled on one end with radioactive phosphorus-32 and induces breakage of the molecule at specific bases through chemical treatment. In total, four cleavage reactions take place. Purines are first methylated, and preferentially cleaved in two reactions: one that favors the cleavage of guanine, and one that favors the cleavage of adenine. Treatment with hydrazine and piperidine leads to cleavage of both pyrimidines, whereas adding sodium chloride to the reaction suppresses the reaction with thymine, leading to cleavage of only cytosine. Reaction conditions are controlled in such a way that only one base is attacked on each molecule, and by running fragments from all four reactions side by side on a polyacrylamide gel, separating each fragment by size, from the labeled end to the point of cleavage, a pattern of bands that can be used to read the sequence directly is revealed (Maxam & Gilbert, 1977).

Frederick Sanger first developed what he called the "plus and minus" method of DNA sequencing prior to the introduction of the chemical cleavage method (Fred Sanger & Coulson, 1975), but it was what became known as the dideoxy method that later became known as Sanger sequencing. As with

the Maxam-Gilbert method, the dideoxy method required four different reactions to run in parallel, but instead of breaking apart the existing molecules, Sanger's method used primed synthesis with specific chain terminating inhibitors to make fragments of varying length. In each reaction, the fragment to be sequenced was mixed with a primer (in the form of viral or complementary strand), polymerase, and a mixture of triphosphates: dCTP, dTTP, dGTP, phosphorus-32 labeled dATP as well as the dideoxy or arabinosyl derivative of one triphosphate in each of the four reactions. The lack of the 3'-hydroxyl group on the dideoxy derivatives terminates chain extension, and in the case of arabinosyl derivatives, the orientation of said hydroxyl group does not allow further synthesis with the polymerase that was being used. In each reaction, the deoxynucleotides and the corresponding dideoxy or arabinosyl derivatives were added in such a ratio that not all incorporations of a given nucleotide would end in chain termination, thus ensuring fragments of different lengths where the final added nucleotide would be known. Fragments from each reaction would then be run side by side on an acrylamide gel, and the order of the bands in each lane would correspond to the location of the terminating nucleotide relative to the beginning of the fragment, allowing the sequence to be read off the gel. (Frederick Sanger et al., 1977) Sanger's method was less technically demanding, requiring less use of toxic chemicals, and held greater potential for upscaling than the Maxam-Gilbert method, and thus became the go-to method for sequencing and further development. (Schadt, Turner, & Kasarskis, 2010; Van Dijk, Auger, Jaszczyszyn, & Thermes, 2014)

By the 2000s, Sanger sequencing was mostly performed on capillary array electrophoresis (CAE) instruments that allowed for up to 96 capillaries to run in parallel. Relatively long, high quality reads of more than 700 bases could be produced, and this technique was used to successfully complete the first full sequence of the human genome (Hert, Fredlake, & Barron, 2008; Van Dijk et al., 2014).

1.3.2 Next generation sequencing

Towards the end of the Human Genome Project, it had become clear that the high cost and rather low throughput of traditional gel-based methods were major obstacles for answering complex biological questions (Goodwin, McPherson, & McCombie, 2016). In 2004 the "1000$ Genome" project was launched, providing funding for the development of new technology, which given time would hopefully achieve the gold standards of sequencing: high accuracy, long reads, high throughput and low cost (Yue Wang, Yang, & Wang, 2015). Common for these so-called next generation technologies, is massively parallel amplification of template DNA, creating high throughput of reads of relatively short length, most only a few hundred bases in length, as well as direct detection of output from the sequencer (Van Dijk et al., 2014).

Van Dijk et. al. Mentions four different technologies when they look back at the first ten years of next generation sequencing (NGS). In 2005, Life Sciences/Roche released the 454 Genome Sequencer, the very first next generation platform. This device uses what is referred to as pyrosequencing: where the DNA-library is loaded into wells, along with primer and enzymes. The wells are then exposed to only one type of dNTP at a time, the incorporation of a given nucleotide to the primer releases pyrophosphate, and the resulting light emission is captured by a charge-coupled device camera (Metzker, 2010; Van Dijk et al., 2014). Another technology similar to pyrosequencing is Ion Torrent semiconductor sequencing: DNA is loaded into wells and only one dNTP is added at a time, however, instead of a camera registering the light-signal of pyrophosphate, the proton released by hydrolysis in chain extension causes a slight shift in pH, which is then detected by sensors in each well (Quail et al., 2012; Rothberg et al., 2011). The Sequencing by Oligo Ligation Detection (SOLiD) technology was developed by Applied Biosystems, and uses a repeating cycle of octamer hybridization probes that when ligated to the sequencing primer can be identified by specific fluorescent labels. (Hert et al.,

2008). The dominant next generation technology, however, was developed by Illumina. With a range of different platforms, Illumina sequencing would provide the lowest cost per base, as well as the highest throughput (Goodwin et al., 2016; Van Dijk et al., 2014).

Illumina released their first sequencing platform in 2006, and as with the 454, it was based on a type of sequencing by synthesis (Van Dijk et al., 2014). DNA fragments are immobilized on the flow cell surface by annealing to one of two oligonucleotides complementary to adapter sequences added to both ends in the library preparation step. Clusters of the same fragments are generated by bridge amplification, where extension primed by the flow cell oligonucleotide generates two complementary strands. Denaturing removes the original template strand, and the newly synthetized strand is annealed to the oligo complementary to the adapter on the opposite end of the strand, forming the shape of a bridge. The bridged strand is copied by polymerase and the strands are separated by denaturing. This process is then repeated over and over for all fragments attached to the flow cell, creating clonally amplified regions for each fragment. Sequencing begins with hybridization of a sequencing primer to the template, before strand extension by cyclic reversible termination. Synthesis is halted after incorporation of each nucleotide, which is fluorescently labeled with a reversible terminator, and unused dNTPs are washed away before imaging is used to determine which nucleotide has been added to each cluster. The terminator is then cleaved, allowing the incorporation of the next nucleotide. The cycle is repeated the same number of times as the total read length of the forward read, and the read product is washed away. To generate reverse reads, the template strand is once again folded in a bridge formation, and a complimentary strand is synthesized before the original template is removed, allowing the same process to take place on the opposite end of the template (Illumina; Metzker, 2010). Although the first sequencer by Illumina only generated reads of 35 base pairs (bp), further improvements to the technology now allows for read lengths of up to 300 bp (Van Dijk et al., 2014).

Next-generation sequencing was developed to tackle the issue of low throughput and high cost associated with first-generation methods: however, it did face challenges of its own. Because these methods rely on template amplification, they are vulnerable to copying errors and bias (Schadt et al., 2010). The short reads, in the range of 25-250 bases initially presented difficulties for assembly, and because of this *de novo* sequencing was likely to remain exceedingly expensive. (Hert et al., 2008). However, significant improvements were made, both in the laboratory, and in data analysis. These advancements, in the form of new sequencing machines, and in the chemistry utilized, produced read lengths of several hundred bp less than ten years later. A host of new algorithms were developed to handle the massive amounts of short-read data, and NGS is now used both for *de novo* assembly and metagenomics (Van Dijk et al., 2014). Genomic analysis of more complex structural variation such as haplotypes or repetitive regions, however, remain challenging for these short-read platforms, and the cost of sequencers is still high (Mikheyev & Tin, 2014).

1.3.3 Third generation sequencing

In recent years, new technologies involving the sequencing of single molecules without the need for amplification have been referred to as *third generation sequencing*. These technologies, unlike next generation sequencing, can produce average read lengths of several thousand bases, and maximum read lengths of more than 100 kilobases (kb). Although these methods are promising and hold a great deal of potential for easier assembly as well as expanding the areas of application, one drawback is a relatively high error rate of sometimes up to 40%. Pacific Biosystems (PacBio) and Oxford Nanopore Technologies (ONT) are currently the major players when it comes to development of third gen sequencers, using Single Molecule Real-Time (SMRT) - , and nanopore sequencing respectively (Bleidorn, 2016; Ye, Hill, Wu, Ruan, & Ma, 2016).

PacBio SMRT sequencing, like most mentioned next generation technologies relies on a type of sequencing by synthesis and fluorescently labeled nucleotides. But unlike e.g. Illumina, SMRT sequencing gives real-time output and does not include any cyclic processes. This is achieved by creating circular DNA templates through the addition of hairpin adapters. Primers and polymerase are then added to the library, before loading onto the sequencers SMRT cell. The SMRT cell consists of small wells, called Zero-Mode Waveguides (ZMW) where single molecules are immobilized by fixing the polymerase to the bottom of the well. Incorporation of a given nucleotide by the polymerase emits a characteristic light signal, which is recorded by a camera. SMRT sequencing generates two types of reads: continuous long reads (CLR), linear reads of high length – or if the template is shorter, the polymerase can traverse the template several times generating a circular consensus sequence (CCS). Because the sequencing errors in SMRT sequencing is randomly distributed, each pass of the polymerase over a template lowers the overall error rate, achieving accuracies of more than 99% by increasing coverage (Ardui, Ameur, Vermeesch, & Hestand, 2018; Goodwin et al., 2016; PacBio, 2020).

The technology developed by ONT, in contrast does not involve any form of fluorescence or synthesis, but directly detects the sequence of ssDNA molecules in real-time (Goodwin et al., 2016). As early as 1996, Kasiannowicz et. al. demonstrated the translocation of single-stranded RNA or DNA through a biological nanopore with the help of an electric field. *Staphylococcus aureus* α-hemolysin – an ion channel with a diameter of 2.6 nm – was embedded in a bilayer membrane separating two compartments of buffer at pH 7.5. By applying a potential of - 120 mV, several current blockages directly proportional to polynucleotide molar concentration were observed. They hypothesized that the decrease in detected ionic current could possibly be used to determine the sequence of nucleotides as they passed through the pore (Kasianowicz, Brandin, Branton, & Deamer, 1996).

For nanopore technology to be able to be developed for the purpose of sequencing, several requirements needed to be met. In 1999, differences in blockage amplitude, blockage duration and pattern were shown for different RNA homopolymers, as well as copolymers of poly A and poly C. This confirmed sensitivity to chemically distinct parts of the molecule as it passed through the pore (Akeson, Branton, Kasianowicz, Brandin, & Deamer, 1999). Although significant challenges remained, especially regarding increased resolution, nanopore sequencing held massive potential. If it could be achieved, the advantages included minimal sample preparation, low cost, and very long read lengths (Branton et al., 2010; D. W. Deamer & Akeson, 2000).

Today, nanopore sequencing is a reality. Oxford Nanopore revealed their first DNA sequencing device: the small, portable MinION in 2012, and it became available for early-access users in April 2014. Preparation of the library includes ligation of an adapter sequence and a motor protein to one end of the sequencing library, along with a hairpin adapter which allows for sequencing of both strands of dsDNA. The MinION flow cell uses 512 membrane embedded protein nanopores to sequence separate DNA molecules. DNA moves through the nanopore as a single strand guided by the motor protein, and changes in voltage are monitored before being translated into k-mers corresponding to the bases present in the pore. Reads of either the forward or reverse strand are called 1D reads, and by sequencing both, a more accurate consensus sequence, termed a 2D read can be generated (D. Deamer, Akeson, & Branton, 2016; Goodwin et al., 2016; ONT, 2019a).

Since its release, the MinION has proven itself as a massively promising technology. Single base resolution is yet to be accomplished, but as with PacBio, increased coverage and 2D reads reduces error rate significantly. Continuous improvements to the technology has led to higher read-lengths, better base-call accuracy and detection of base modifications, as well as higher throughput: mostly

thanks to development of new platforms such as the PromethION (Jain, Olsen, Paten, & Akeson, 2016).

In the over 40 years since the work of Sanger, Maxam and Gilbert made large scale DNA sequencing possible, the field has gone through massive development. With the advent of third generation sequencing, all four of the gold standards of sequencing set forth by the "1000$ genome project" might finally be within reach (Y. Wang et al., 2015). The long reads will help resolve complex genomic regions, as well as greatly simplify *de novo* assembly of non-model organisms. In the field of metagenomics, the longer reads can also help achieve better species assignment, although this is reliant on high accuracy of consensus sequences. As sequences are retrieved in real-time, third generation sequencing also offers a new tool for clinical application, where the MinION has already shown it can quickly produce results in the field. Of these new technologies, nanopore sequencing in particular holds great promise; the low cost and portability afforded by the MinION, as well as minimal required sample preparation could potentially make sequencing available to much smaller laboratories and institutions (Bleidorn, 2016).

1.4 Metagenomics workflow

Metagenomics is a culture-independent method for analyzing microbial communities in environmental samples. Our traditional understanding of many microbial populations has mostly been based on the relatively few species that have been culturable in the lab, thus giving limited insights into the complexity of these communities (Hugenholtz & Tyson, 2008). The metagenome, which is the sum of genetic material in an environment, can be studied at different levels, depending on the purpose of the study. Marker gene analysis, such as 16S rRNA gene amplicon sequencing, is a quick and relatively cheap way of gaining a low-resolution taxonomic overview of microbial communities. For a more detailed insight into these communities, whole metagenome analysis, where all DNA in a given sample is sequenced can be applied. Along with other omics-based methods, including metatranscriptomics, metaproteomics and metabolomics, a deeper understanding of the composition and function of microbial communities can be achieved (Knight et al., 2018).

1.4.1 Extraction protocols:

Extracting high molecular-weight microbial DNA from natural cellulose-degrading communities presents a unique challenge, due to adsorption between cells and biomass, as well as the presence of host cells, potential enzymatic inhibitors and biofilms (Kunath, Bremges, Weimann, McHardy, & Pope, 2017). These factors can potentially lead to reductions in concentration, integrity and diversity of DNA during extraction, and it is therefore important to consider when working with metagenomic samples (Du, Guo, Li, Xie, & Yan, 2018). When working with environmental samples, DNA extraction methods can be divided into two categories. Extraction where cells are lysed within original sample material is termed direct extraction, whereas methods that first remove cells from the sample material prior to lysis is referred to as indirect extraction (Courtois et al., 2001).

Both direct and indirect extraction methods have their advantages and disadvantages. Direct methods are typically viewed as appropriate for determining prokaryotic taxonomic diversity due to their ability to capture more of the complete genomic material in the sample and higher yield with less sample material. This, however, comes at the price of possibly retaining extracellular DNA from the sample material, as well as reducing the fragment length obtained through the extraction due to shearing. (Williamson, Kan, Polson, & Williamson, 2011). Indirect methods generally produce larger fragments of specifically microbial DNA, but may decrease sample diversity, due to extraction biases (Robe, Nalin, Capellano, Vogel, & Simonet, 2003). It has been shown however, that although diversity

was affected by choice of extraction method, the relative diversity of each method was comparable, albeit accessing slightly different populations within the total community, and that indirect extraction using higher volumes of sample material did not seem to be more biased than the direct method (Delmont, Robe, Clark, Simonet, & Vogel, 2011).

All extraction methods, both direct and indirect, can be separated into six steps: sample pre-processing, cell lysis, purification, concentration, fragmentation and quality control, of which pre-processing and fragmentation are viewed as optional, and the rest required (Quick & Loman, 2019). For each of these six steps, several options are available, however here, only a few of the most common methods will be discussed.

For cell lysis, both chemical, enzymatic and mechanical methods are utilized (Quick & Loman, 2019). Chemical lysis generally involves the use of a detergent such as sodium dodecyl sulfate (SDS) which help dissolve cell membranes, whereas enzymatic treatment most commonly includes lysozymes that break down linkages within the peptidoglycan layer of cell walls. Mechanical lysis is independent on cell wall structure, and thus can achieve access to the entire bacterial community. One commonly used method of mechanical lysis is bead beating, in which glass or zirconium beads are added to the sample mix and shaken vigorously on a homogenizer. This method, and mechanical methods in general, can show quite high total yields, but at the cost of DNA shearing (Robe et al., 2003).

Common methods for purifying extracted DNA include column filtration, and the use of phenol: chloroform (Henderson et al., 2013). Spin columns selectively bind and separate nucleic acids from proteins and other contaminants by passing the solution through a filter, leaving the DNA bound in the matrix. Nucleic acids are subsequently released from the filter using an elution buffer (Purdy, Embley, Takii, & Nedwell, 1996). Phenol, especially combined with chloroform effectively separates proteins and lipids from DNA in alkaline solutions by absorbing these into the heavier organic phase produced after centrifugation, leaving DNA in the aqueous partition on top (Green & Sambrook, 2017).

Kunath et. al. (2017) describes a protocol for manually extracting high molecular weight DNA suitable for long-read sequencing from plant biomass using chemical lysis and purification. Cells are first dissociated from biomass by suspension in acidic solution (pH 2) and lysed by incubation at 70°C with a SDS-containing lysis buffer and cetyltrimethylammonium bromide (CTAB) in a saline solution. Purification is then achieved by first adding an equal volume of chloroform followed by phase-separation by centrifugation. This step is then repeated as necessary, before the aqueous phase is transferred to a new tube and mixed with an equal volume phenol:chloroform:isoamylalcohol (25:24:1). Phases are again separated by centrifugation, and the supernatant is transferred to an ethanol solution for precipitation. DNA is then pelleted by centrifugation and briefly airdried after removal of ethanol, before resuspension in an adequate volume of chosen storage buffer.

DNA extraction is an essential step in any metagenomic project, and the choice of method has been shown to have a significant impact on downstream analysis. No method is deemed superior to others for all purposes, and the best method will vary from project to project, depending on the application and the specific nature of the samples (Gerasimidis et al., 2016; Henderson et al., 2013).

1.4.2 DNA quality control

When applying metagenomic approaches, it's important to examine the quality of the DNA to make sure that samples have sufficiently high yields, as well as meeting a certain standard for the overall quality. Some common methods of determining yield, purity and fragment length is discussed below.

Nucleic acids can be directly detected by UV spectrometry due to their ability to absorb UV-radiation. Concentrations can be determined based on standard curves, however presence of common contaminants such as phenol, as well as the combined presence of both DNA and RNA can skew results (Nielsen et al., 2008). In addition, the sensitivity of this method is not as high as that achieved by fluorometric quantification methods (Rengarajan, Cristol, Mehta, & Nickerson, 2002), and is therefore more commonly used for measuring absorbance ratios to determine the purity of a given sample. In UV spectrometry, nucleic acids have a maximum absorbance at 260 nm (A_{260}), and the ratio between this and absorbance at 280 nm (A_{280}) and 230 nm (A_{230}) respectively are of specific interest (Boesenberg-Smith, Pessarakli, & Wolk, 2012). The $A_{260/280}$-ratio can reveal whether a DNA sample is contaminated with RNA as well as proteins: for pure DNA, the ratio should be approximately 1.8 – protein contamination might lower this ratio, whereas RNA contamination will increase it up towards 2. A wide range of common contaminants such as proteins and substances used in DNA extraction and can also be revealed by the $A_{260/230}$-ratio, which for pure samples should be somewhere in the range of 1.8-2.4, with lower values indicating contamination (Koetsier & Cantor, 2019).

Although historically nucleic acid yields have been measured using spectrometry, due to the ability of contaminants to inflate measured yields and the inability to measure DNA and RNA concentrations independently, quantitation is now most commonly achieved by fluorometry. This method uses fluorescent dyes which bind selectively to certain materials such as dsDNA. These dyes are then exited at a certain wavelength and emits another, allowing the intensity of emitted light to be measured and thus determine concentration of the desired molecule (Boesenberg-Smith et al., 2012). Fluorometry is the most sensitive method of quantifying DNA yield and is also highly specific when measuring one nucleic acid in the presence of another (Invitrogen, 2018).

Agarose gel electrophoresis is commonly used in molecular biology to separate large molecules such as proteins, DNA and RNA based on their size (Drabik, Bodzoń-Kułakowska, & Silberring, 2016). In the case of nucleic acids such as DNA, the negatively charged phosphate backbone of the molecule leads to migration in the direction of the positive pole when an electric field is applied. The velocity of migration for a linear polynucleotide in an agarose gel is determined by its size, and the voltage applied. Larger molecules meet more resistance from the gel matrix, and therefore migrate at a slower rate than those of lower molecular weight (Voytas, 2000). Agarose concentrations vary depending on desired resolution, ranging from 0.7% to 2% depending on the expected fragment length, with higher concentrations needed to separate shorter fragments, and lower concentration to allow adequate mobility for longer molecules (Yılmaz, Ozic, & Gok, 2012).

1.4.3 Marker gene/amplicon sequencing:

In marker gene sequencing, a specific phylogenetically conserved region of DNA is examined to determine the taxonomic composition of the microbial community. The chosen gene is typically amplified through the polymerase chain reaction using target specific primers and then sequenced (Knight et al., 2018). The 16S rRNA gene is present at least once in all prokaryotic genomes and is the most widely targeted region in microbial studies for identifying bacterial and archaeal strains (Yong Wang & Qian, 2009). With a total length of approximately 1600 base pairs, the 16S rRNA gene contains nine hypervariable regions with varying degrees of conservation, making it ideal for revealing composition both at species- and higher taxa-level based on choice of region (Bukin et al., 2019).

The polymerase chain reaction (PCR) was developed in the 1980s by Kary Mullis for the synthesis of specific DNA sequences using only a few simple reagents and repetitive cycles of denaturation, hybridization and polymerase extension (Mullis et al., 1986). In essence, PCR amplifies a specific

region of DNA through the annealing of two oligonucleotides (primers) on either side of the target sequence and extending these primers using a DNA polymerase. This is achieved by cycling through different temperatures appropriate for each step of the reaction: first denaturing double stranded DNA, then allowing the primers to hybridize to each strand and finally extending said primers, thus doubling the amount of target DNA per cycle (Erlich, Gelfand, & Sninsky, 1991).

Several factors can impact the results of an amplicon analysis. As mentioned, for prokaryotic communities, the 16S rRNA gene tends to be the marker gene of choice, largely due to its highly conserved structure, its widespread use and the availability of comprehensive reference databases. One weakness of this gene however is its variable copy number in different species, which can cause biases when estimating relative abundances within a community (Kunin, Copeland, Lapidus, Mavromatis, & Hugenholtz, 2008). The choice of primer has also been shown to have considerable impact on measured relative abundances, with the potential of introducing biases by differential amplification of the same template, an effect that increases with the number of cycles (Suzuki & Giovannoni, 1996). Primer-introduced biases have been thought to be caused by differences in primer binding energy and the reannealing of templates inhibiting further amplification in later cycles (Acinas, Sarma-Rupavtarm, Klepac-Ceraj, & Polz, 2005). Furthermore, Sze and Schloss (2019) found that sequencing errors vary both by the number of cycles, and to a lesser extent the polymerase used, and thus recommend using as few cycles as possible, along with a high-fidelity polymerase to limit potential biases. Finally, as next-generation sequencing is limited to only a few hundred base pairs, researchers must generally choose a limited portion of the gene for sequencing, and this choice has also been shown to affect results (Bukin et al., 2019). With the advent of third generation sequencing technologies, this may change, as high throughput sequencing of the complete 16S rRNA gene is becoming increasingly feasible, allowing for better taxonomic resolution by not having to choose a shorter region of the gene (Johnson et al., 2019).

Until recently, the processing pipelines for the output generated by high-throughput amplicon sequencing has generally used clustering to generate operational taxonomic units (OTUs) based on sequence identity, with identities above 95% typically used as the threshold for genus-classification, and 97% identity between sequences interpreted as belonging to the same species. This is done in part to reduce the impact of artifactual sequences that can arise from amplification and sequencing. (Johnson et al., 2019). One weakness of this method however, is that the OTUs are inherently dependent on each dataset, thus one cannot make comparisons of *de novo* OTUs across data sets. (Callahan, McMurdie, & Holmes, 2017) Higher sensitivity methods to determine the exact sequence variants (also called amplicon sequence variants) such as DADA2 are now often recommended to gain a higher resolution view of the community, as long as the sequences analyzed have been generated using the same primer pair and sequencing platform (Knight et al., 2018). DADA2, an open-source R-package uses a model of separating sequencing errors from genuine biological diversity based on their frequency in the data set, assuming that if a sequence is observed at higher frequencies, it is less likely to have originated from sequencing errors. Traditional OTU-clustering is thus not as necessary to minimize the effects of sequencing errors on taxonomic classification, and allows for the use of actual biological sequences as the atomic unit in analysis (Callahan et al., 2017; Callahan et al., 2016).

Amplicon sequencing has been a powerful tool for microbial ecology, despite the potential for biases, and will likely remain so as it is the most cost-effective way of gaining insight into community composition. Even as methods for taxonomic profiling based on whole metagenome sequencing allows for the amplification-bias free identification of species independently of domain of life, marker gene analysis of the 16S rRNA gene benefits from access to comprehensive databases with

information from millions of species, making it an invaluable resource for profiling communities with limited reference genomes in the databases (Breitwieser, Lu, & Salzberg, 2019).

1.4.4 Whole genome/shotgun sequencing

As aforementioned, "shotgun" metagenomics involves randomly sequencing DNA fragments that in theory represent all microbial constituents from a given sample, once samples have been sequenced, two types of computational analyses form the basis for further investigations: alignment and assembly. If the reference genome is known, alignment offers fast confirmation of the success of the sequencing, however, where reference sequences are not known, these must be assembled from the raw reads (de novo assembly) (Flicek & Birney, 2009).

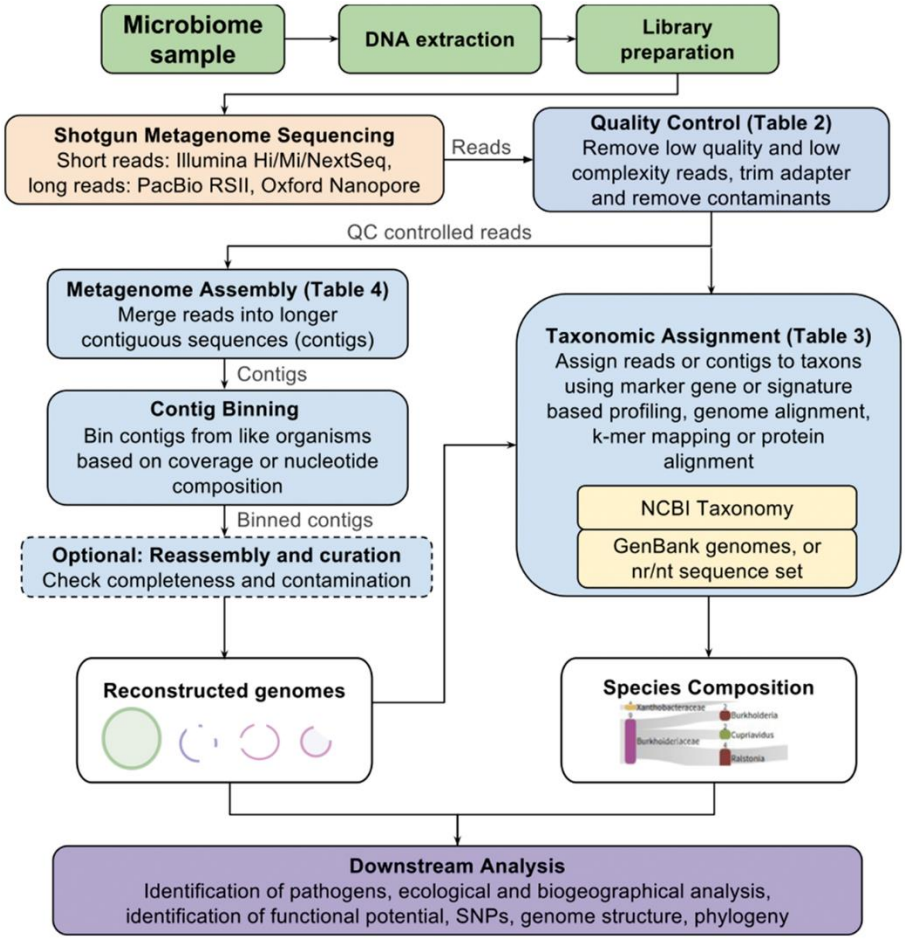


Figure 1.4.4.1: Overview of workflow for whole metagenome sequencing analysis. Figure obtained from Breitwieser et al. (2019)

Taxonomic profiling of raw or quality-filtered reads generally relies on aligning these to existing databases of known genomes and offers an overview of the species present in a sample as well as their relative abundances, much like in amplicon analysis. One key difference between taxonomic analysis of shotgun sequencing and that of marker gene sequencing is the ability to capture sequences across all domains of life, including eukaryotes and viruses, thus gaining a deeper view of the community structure. It is however limited by the availability of reference genomes in databases, meaning that highly complex communities where low-abundance species remain mostly uncharacterized cannot be completely successfully profiled (Quince, Walker, Simpson, Loman, & Segata, 2017).

Assembly is the process of transforming raw sequencing reads into a reconstruction of the target genome. This is achieved by aligning overlapping regions of reads to each other to generate contiguous sequences (contigs), which in turn are constructed into larger scaffolds with information on the position of each contig within the genome (Miller, Koren, & Sutton, 2010). Several methods for de novo assembly exist, but two types of algorithms are more commonly used: the overlap layout consensus (OLC), and the deBruijn graph. Both algorithms operate by generating a graph, the path through which infers the consensus sequence (Z. Li et al., 2012). The deBruijn graph splits reads into even shorter fragments of a certain length, called k-mers, which becomes nodes in the graph. Nodes are then connected based on adjacent sequences from the original reads, ideally forming a path through the graph including all edges that represent the consensus sequence (Pop & Salzberg, 2008). In contrast, overlap layout consensus finds pair-wise overlaps between all reads and creates an overlap graph in which the whole read becomes a node, and where overlapping bases in the reads leads to connected nodes. Finally, a consensus sequence is determined by the arrangement of overlapping reads (Pop, 2009). Due to the shorter k-mers of the deBruijn graph, it has become a popular algorithm for assembling short next generation sequencing reads, but this type of algorithm is particularly vulnerable to repeats and sequencing errors, making it less ideal for error-prone third generation assembly. The OLC algorithm is less sensitive to errors because the information of each read is kept until the consensus step (Miller et al., 2010)

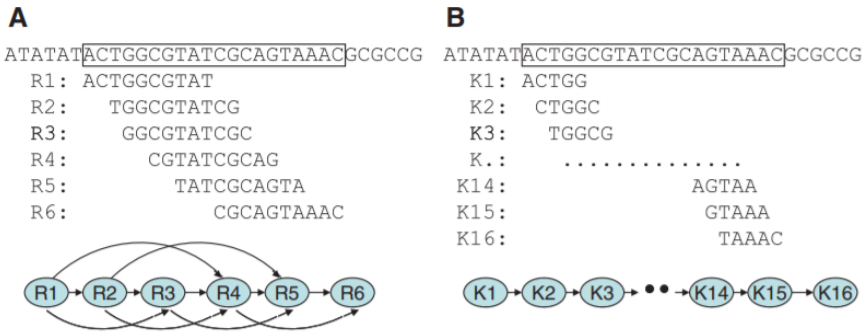


Figure 1.4.4.2: Algorithm for assembly by A) Overlap consensus layout B) DeBruijn graph. Figure obtained from: Z. Li et al. (2012)

Metagenomic assembly is similar to that of genomic assembly but faces unique challenges due to reads belonging to one of many genomes present in the sample. The abundance of a given species varies, leading to uneven sequencing depth (Charuvaka & Rangwala, 2011). Regular assemblers generally assume even sequencing depth across the sequenced genome, which is not the case between species in a metagenomic sample, thus mechanisms that rely on coverage information such as those for resolving repeats, identifying allelic variation and sequencing error can no longer function as intended (Quince et al., 2017). Repeats and other conserved sequences also cause additional difficulties for metagenomic assembly due to the similarity of these regions both between strains and within a single genome, making it even harder to determine the origin of a given read. Another issue when attempting to assemble genomes from a metagenomic sample is getting sufficient sequencing depth for more than just the dominant species, as coverage is proportional to the abundance of a given species, with increasing complexity of the community lowering the sequencing depth per genome, often leading to incomplete assemblies of low-abundance genomes (Breitwieser et al., 2019).

The assemblies generated in metagenomics tend to be highly fragmented, as the origin genome of each contig is unknown. Binning is the process of sorting these contigs into groups or “metagenome assembled genomes (MAGs) that correspond to individual organisms, thus making a scaffold for each

individual genome that has been assembled (Alneberg et al., 2014). This can generally be achieved in one of two ways: supervised binning, where databases of already sequenced genomes are used to sort contigs based on taxonomy, or unsupervised binning, in which contigs are clustered in an attempt to find the natural groups in the data (Quince et al., 2017). Clustering generally uses information about characteristics such as coverage and nucleotide composition to separate contigs into bins representing different species (Sangwan, Xia, & Gilbert, 2016).

An alternative method of binning without the need for assembly assigns taxonomy directly to raw reads. This type of community profiling can be used as an alternative to traditional marker gene analysis, and holds the advantage of detecting sequences from all types of organisms present in the sample, and circumvents the issue of primer- and amplification biases. This method is somewhat limited however by the short reads typically generated by next-generation sequencing (Breitwieser et al., 2019).

Once assembly and binning has successfully yielded one or more MAGs, protein- and RNA-coding genes can be identified from the sequence by gene prediction algorithms, a process commonly called gene calling. By identifying these regions and narrowing down the dataset, the amount of computational strain on further downstream analysis is significantly reduced (Trimble et al., 2012). Gene calling can be performed at any point after sequencing, on unassembled reads, shorter contigs or fully assembled MAGs, and have two main modes of predicting genes. One approach uses sequence similarity to search databases for previously documented genes that match up with those found in the dataset, whereas the second, the "*ab initio*" method, uses features of a sequence such as nucleotide composition and codon frequencies to separate coding and non-coding regions (Kunin et al., 2008). The "*ab initio*" approach is generally preferred for metagenomic projects, as these can have higher frequencies of fragmented or partially sequenced genes, and the organisms present in the sample may from complex communities with less exhaustive databases, thus preventing the successful detection of both homologs and novel genes (Kunath et al., 2017).

The predicted genes can be further used to annotate the genome and predict metabolic functions and/or pathways. Sequences are compared to existing databases in an attempt to find orthologs and predict function of the called genes (Stothard & Wishart, 2006). Several approaches are available, including those for recognizing protein families and domains, and Enzyme Commission (E.C.) numbers that classifies enzymes based on the chemical reactions they catalyze. The Kyoto Encyclopedia of Genes and Genomes (KEGG) database is one of those most widely used and allows users to quickly link genes to function. This is achieved by assigning genes to entries in the KEGG Orthology database, and as entries are defined in functional context, these entries can be used to easily reconstruct metabolic pathways via KEGG Pathway maps (Kanehisa, Sato, Kawashima, Furumichi, & Tanabe, 2016). In the case of complex carbohydrate-utilizing consortia, specialized databases exist to identify carbohydrate active enzymes based on significant amino acid similarity with at least one biochemically characterized founding member (Kunath et al., 2017; Lombard et al., 2014).

While metagenomics and the comparisons of genome sequences makes it possible to determine the physiological potential within a community, combining this with other omics-methods such as transcriptomics and proteomics offers a deeper understanding of the proteins involved in processes such as lignocellulosic degradation (Baldrian & López-Mondéjar, 2014). Using such a multi-omics approach can help shed light on gene regulation and other possible changes in response to certain factors such as dietary adjustments and medicines, as well as gaining a wider functional and mechanistic understanding of microbial communities (Knight et al., 2018).

1.5 Thesis objectives

Culture-independent techniques such as metagenomic sequencing allows for detailed analysis of the composition and potential functions of microbial communities. Two types of sequencing are commonly used in the profiling of microbial communities: marker gene amplicon sequencing and whole metagenome sequencing. Marker gene amplicon sequencing offers a powerful tool for determining the community structure of microbial samples, whereas shotgun sequencing is useful for identifying genes and examining microbial metabolic pathways.

The short-read sequencing technologies that dominate the field of genomics today tends to struggle with resolving larger structural variation and repeated regions, leading to fractured genomes. This can cause the failure to identify genes of interest by cutting open reading frames in two, or missing the part containing the gene completely. New long-read technology such as that developed by Oxford Nanopore may potentially solve assembly-difficulties related to short read lengths by producing reads that cover more of difficult regions such as repeats.

The main objectives of this thesis are to explore various steps of the current metagenomic workflow, both in the lab and bioinformatically, and to examine the potentials of long-read sequencing technology in the form of Oxford Nanopore sequencing. To achieve this, samples from two different studies have been subjected to amplicon analysis through 16S rRNA gene sequencing and whole metagenome sequencing respectively. The samples used for amplicon sequencing originated from sheep rumen and were part of a larger scale project by Foods of Norway and is detailed in section 3.1. The other set of samples were derived from human fecal enrichments and were sequenced using Oxford Nanopore MinION devices.

2 Materials

2.1 Lab equipment

Product	**Supplier**
913 pH Meter, laboratory version	Metrohm Nordic AS, Bærum, Norway
FastPrep-24™ Classic Grinder	MP Biochemicals, Ohio, USA
Gel Doc™ EZ System	Bio-Rad, California, USA
Labcycler Gradient, Thermoblock 96, silver	SensoQuest GmbH, Göttingen, Germany
Mastercycler® Gradient	Eppendorf, Hamburg, Germany
MinION Flow Cell (R9.4.1)	Oxford Nanopore Technologies, Oxford, Great Britain
MinION Sequencer	Oxford Nanopore Technologies, Oxford, Great Britain
MiSeq® system	Illumina, San Diego, California, USA
Multi RS-60 , Programmable rotator	Biosan, Riga, Latvia
NanoDrop ND-1000 Spectrophotometer	Thermo Fisher Scientific, Massachusetts, USA

PowerPac™ Basic Power Supply	Bio-Rad, California, USA
Qubit ™ 1 Fluorometer	Invitrogen, Carlsbad, California, USA
ThermoMixer® C	Eppendorf, Hamburg, Germany

2.1.2 General lab equipment

Product	**Supplier**
Nitrile gloves	VWR, Pennsylvania, USA
Automatic pipettes, single channel	Thermo Fisher Scientific, Massachusetts, USA
Automatic pipettes, multichannel	Thermo Fisher Scientific, Massachusetts, USA
Axygen® 2.0 mL MaxyClear Snaplock Microcentrifuge Tube	Axygen
ddH2O, Milli-Q® Reference Water Purification System (0,22 µm filter)	Merch-Millipore, Massachusetts, USA
Duran® Glass flasks	Shcott, Wertheim, Germany
Eppendorf® Centrifuge 5418R	Eppendorf, Hamburg, Germany
Falcon tubes, 10 ml	Greiner tubes, SigmaAldrich, Missori, USA
Freezer (-20°C)	Bosch, Stuttgart, Germany
Freezer (-80°C), Innova® C585 Chest Freezer, New Brunswick MG	MG Scientific, Wisconsin, USA
Magnetic stirrer, IKA® RCT basic IKAMAG™ Safety Control	Thermo Fisher Scientific, Massachusetts, USA
IKA Mixer Vortex Shaker Model MS 2	Thermo Fisher Scientific, Massachusetts, USA
Sartorius Quintix 124-1s	VWR, Pennsylvania, USA
Refridgerator (4°C)	Bosch
Galaxy 14D Micro Centrifuge	VWR, Pennsylvania, USA
Tisch-Autoclave	CertoClav
16-Tube SureBeads™ Magnetic Rack	Bio-Rad
Mini-Sub Cell GT Cell	Bio-Rad
Mini-Gel Caster	Bio-Rad
Pasture pipette 5 mL non-sterile graduated up to 1 mL	VWR, Pennsylvania, USA
Biospehere filter tips (volume ranges 0.1-20µL, 2.0-20µL, 20-300µl, 200µl, 1250µl)	VWR, Pennsylvania, USA
ART™ Barrier Hinged Rack Pipette Tips	Thermo Fisher Scientific, Massachusetts, USA
Ultra fine pipette tip	VWR, Pennsylvania, USA

Finntip™ Pipette Specific Pipette Tips, 10mL	Thermo Fisher Scientific, Massachusetts, USA
Axygen® 1.5 mL MaxyClear Snaplock Microcentrifuge Tube	Axygen
Axygen® 0.2 mL Thin Wall PCR Tubes with Flat Cap	Axygen
Axygen® 0.2 mL Thin Wall PCR 8-strip tubes and flat strip caps	Axygen

2.2 Chemicals, manufactured reagents and kits

2.2.1 Chemicals

Chemical	**Supplier**
Seakem LE Agarose	Lonza
Chloroform, EMSURE® ACS, ISO, Reag. Ph. Eur. for analysis	Merck Millipore, Burlington, Massachusetts, USA
Titriplex® II	Sigma-Aldrich, Saint-Louis, Missouri, USA
Ethanol absolute	Merck Millipore, Burlington, Massachusetts, USA
2-Propanol, EMSURE® ACS, ISO, Reag. Ph. Eur. for analysis	Merck Millipore, Burlington, Massachusetts, USA
AnalaR NORMAPUR® Sodium Chloride	
Sodium hydroxide reagent grade, ≥98%, pellets (anhydrous)	Sigma-Aldrich, Saint-Louis, Missouri, USA
Phenol:Chloroform:Isoamyl Alcohol 25:24:1 Saturated with 10 mM Tris, pH 8.0, 1 mM EDTA	Sigma-Aldrich, Saint-Louis, Missouri, USA
Sodium dodecyl sulfate ACS reagent, ≥99.0%	
Trizma® base	Sigma-Aldrich, Saint-Louis, Missouri, USA
Methanol, EMSURE® ACS,ISO,Reag. Ph. Eur. For analysis	Merck Millipore, Burlington, Massachusetts, USA
TWEEN® 80	Sigma-Aldrich, Saint-Louis, Missouri, USA
Tert-butanol	

2.2.2 Manufactured buffers, reagents and kits

Reagent	**Supplier**
50x TAE Electrophoresis Bufffer	Thermo Fisher Scientific, Massachusetts, USA
AMPure XP	Beckman-Coulter
Blunt/TA Ligase Master Mix	New England Biolabs, Ipswich, Massachusetts, USA
DNeasy PowerSoil Kit	QIAGEN, Hilden, Germany
Flow Cell Priming Kit	Oxford Nanopore Technologies, Oxford, Great Britain

Flow Cell Wash Kit	Oxford Nanopore Technologies, Oxford, Great Britain
Gel loading dye blue (6x)	New England Biolabs, Ipswich, Massachusetts, USA
Iproof HF MasterMix	BioRad
Ligation Sequencing Kit	Oxford Nanopore Technologies, Oxford, Great Britain
MiSeq Reagent Kit v3	Illumina
NEBNext® FFPE DNA Repair Mix	New England Biolabs, Ipswich, Massachusetts, USA
NEBNext® Ultra™ II End Repair/dA-Tailing Module	New England Biolabs, Ipswich, Massachusetts, USA
Nextera XT Index Kit	Illumina
PeQGreen DNA/RNA binding dye	PeQlab
PhiX control v3	Illumina
Pro341F PCR primer	Eurofins genomics
Pro805R PCR primer	Eurofins genomics
Qubit dsDNA BR Assay Kit	Invitrogen
Quick-load®, Purple 1 kb DNA ladder	New England Biolabs, Ipswich, Massachusetts, USA

2.3 Buffers

Tris-HCl 1M pH 8

60,57 g Trizma® Base was weighed and dissolved in 400 mL Milli-Q.

pH adjusted to 8 with 37% HCl, and Milli-Q was added to a final volume of 500 mL.

Sterilized by autoclaving.

Tris-HCl 10mM pH 8,5

200 µL of Tris-HCl 1M pH 8 was diluted with sterile water to a volume of 10 mL

pH was adjusted to 8,5 using 5M NaOH, and final volume was adjusted to 20 mL with sterile water.

NaCl 5 M

29,209 g AnalaR NORMAPUR® Sodium Chloride was weighed and dissolved in Milli-Q to a total volume of 100 mL using heated magnetic stirrer.

Sterilized by autoclaving.

EDTA 0,5 M pH 8

11,159 g Titriplex® II weighed and dissolved in 40 mL Milli-Q

pH adjusted to 8 using NaOH pellets and 5M NaOH

Milli-Q added to total volume of 60 mL

Sterilized by autoclaving.

NaOH 0,2 M

0,2372 g anhydrous sodium hydroxide was weighed and dissolved in autoclaved Milli-Q to a total volume of 29,5 mL

Cell wash buffer

500 µL 1 M Tris-HCl
10 mL 5 M NaCl
Milli-Q added to total volume of 50 mL
Sterilized by autoclaving.

Dissosiation/DSS buffer pH 2

1 mL methanol
100 µL Tween 80
1 mL tert-butanol
Sterile water added to a total volume of 100 mL
pH adjusted to 2 using 37% HCl

RBB + C/Lysis buffer

30 mL Milli-Q
10 mL 5 M NaCl
5 mL 1 M Tric-HCl pH 8
10 mL 0,5 M EDTA
Sterilized by autoclaving.

4 g SDS added while solution is still warm and dissolved, autoclaved water added to total volume of 100 mL

CTAB buffer

14 mL 5M NaCl

10 g Cetyl trimethylammonium bromide
Sterile water to a total volume of 100 mL

1 x TAE buffer

100 mL 50x TAE Electrophoresis Buffer
Milli-Q added to total volume of 5 L

2.4 Software tools

Rstudio (DADA2, Biostrings, Phyloseq and ggplot2 packages)
EPI2ME Desktop Agent
Metagenemark (http://exon.gatech.edu/meta_gmhmmp.cgi)
dbCAN (http://bcb.unl.edu/dbCAN2/blast.php)
Ghostkoala (https://www.kegg.jp/ghostkoala/)

3 Methods

3.1 Sampling

Samples for the amplicon analysis were part of a project by Foods of Norway. In an effort to find new ways of feeding livestock using Norwegian bioresources and thus improving food security in years of poor grass crops, 24 lambs at Ås gård were subjected to one of three diets with variable amounts of the seaweed *Saccharina latissimi* (sugar kelp) for an experimental period of one month. Beyond drying and chopping, the seaweed was not processed, and was served as a replacement for some of the roughage fed to these animals. The purpose of the study was to determine the effect of the sugar kelp on the health of the animals and the taste of the meat. The feeding groups, as shown in table 3.1.1, all consisted of 8 biological replicates, and were given 0%, 5% and 2.5% sugar kelp respectively.

Table 3.1.1: Design of feed groups

Feeding group	Seaweed inclusion level	Fluid samples	Particle samples
A	0 %	8	8
B	5 %	8	8
C	2.5 %	8	8

Temporal samples were collected through tubing throughout the month, however the samples discussed here were all from the final sampling, taken at the slaughterhouse. Each sample was separated into a fluid (i.e. lumen) - and particle- (i.e. fibre attached) phase by using sterile stomacher bags with approximately 500-micron pore-sized filter cloth. A total of 48 samples, one fluid- and one particle-phase from each animal were stored at -80°C prior to extraction.

3.2 Extraction

3.2.1 Extraction from dummy samples

3.2.1.1 Manual CTAB + Phenol: chloroform method

Manual HMW DNA extraction was performed as described by Kunath et.al. (2017) with some adjustments. One fluid phase and one particle phase sample were thawed on ice, and homogenized by vortexing, before 0,6 g of biomass from each sample was transferred into new 1,5 mL Eppendorf tubes. Samples were resuspended in 500 µL dissociation buffer and centrifuged for 30 seconds at 100 rcf before transferring cell-containing supernatant to a new tube. Dissociated cells were pelleted by centrifugation at 14 000 rcf for two minutes, and cell-free supernatant was discarded. Resuspension of biomass and supernatant transfer were repeated until cell pellet was easily spotted, a total of three repetitions for the fluid phase sample, and six repetitions for the particle sample. Cell-

containing collection tube was kept on ice and only centrifuged at 14 000 rcf after every second transfer of supernatant to avoid over-compacting of the cell pellet. Cell wash was performed by resuspending the cell pellets in 1 mL of cell wash buffer, followed by centrifugation for 30 seconds at 100 rcf and subsequent transfer of cell-containing supernatant to a new tube. Cells were then pelleted by 2 minutes of centrifugation at 14 000 rcf and supernatant was once again discarded. The pellet was resuspended in 1 mL of cell wash buffer and centrifuged again at 14 000 rcf for two minutes, supernatant was discarded, and one cell pellet per sample was brought forward for DNA extraction.

Cell lysis was accomplished chemically by first resuspending pellets in 1 mL lysis buffer and incubating at 70°C, mixing tube by inversion every 5 minutes for a total of 20 minutes. Then each sample was split equally into two tubes and 80 µL 5M NaCl and 60 µL cetrimonium bromide (CTAB) was added to each tube, which were then incubated at 70°C for 10 minutes. Purification was accomplished by first adding 680 µL of chloroform to each tube and mixing by inversion. Organic and aqueous phase was then separated by 15 minutes of centrifugation at 14 000 rcf, and the clear aqueous phase was transferred into a new tube. An equal volume of Phenol:Chloroform:Isoamylalcohol (25:24:1) was then added, mixed and centrifuged in the same manner as the chloroform and the aqueous phase was again transferred into a new tube. Purified DNA was stored in 1x volume of isopropanol and kept in freezer overnight. A pellet was formed by 20 minutes of centrifugation at 10°C and 14 000 rcf and supernatant was removed. DNA was resuspended by careful pipetting in 200 µL 70% ethanol, before re-pelleting by 2 minutes of centrifugation at 10°C and 14 000 rcf. Ethanol was removed and DNA pellets were allowed to dry briefly before resuspension in 30 µL sterile water. Quality and yield were determined by gel electrophoresis and Qubit fluorometer respectively as described in detail in section 3.5.

3.2.1.2 Qiagen DNeasy PowerSoil Kit

One fluid and one particle phase sample were thawed on ice and homogenized by vortexing prior to extraction. The recommended 0,25 g of biomass was weighed into provided PowerBead tubes, 750 µL PowerBead solution and 60 µL of buffer C1 was added to the tubes and mixed by vortexing briefly. Bead beating was performed using a FastPrep 24 at 4 m/s for 45 seconds and samples were immediately transferred to ice. The post-lysis centrifugation was first attempted at 9 900 g for 30 seconds but was increased to a total of one minute to fully pellet biomass. Supernatant was then transferred into new tubes and mixed with 250 µL of buffer C2 by briefly vortexing, followed by 5 minutes of incubation at 4 °C. Tubes were centrifuged for one minute at 9 900 g and supernatant was again transferred to new tubes, which was then mixed with 200 µL of buffer C3 by vortexing and incubated for another 5 minutes at 4 °C. Another 1-minute centrifugation at 9 900 g was performed to pellet remaining particles, and supernatant was transferred once more. A total of 1200 µL of buffer C4 was added to the supernatant-containing tubes and mixed by vortexing. The solution was then sequentially passed through a spin column in three rounds: 675 µL was loaded onto the column at a time and centrifuged at 9 900 g for one minute, discarding the flow through after each round. The spin column was then washed by adding 500 µL of buffer C5, followed by two rounds of centrifugation at 9 900 g for 30 seconds and 1 minute respectively, discarding flow through after each round. Finally, DNA was eluted by moving the spin filter into a new tube before adding 100 µL of buffer C6 and centrifuging for 30 seconds at 9 900g. Concentration was determined using the Qubit dsDNA BR Assay Kit with a Qubit fluorometer, and quality was assessed by running DNA on a 1% agarose gel as described in detail in section 3.5.

3.2.2 Analytical samples

DNA was extracted from all 48 analytical samples over 6 days using the Qiagen DNeasy PowerSoil Kit. Two samples were used for the first run to verify the protocol, before the number of samples processed together was gradually increased to 12 as the protocol became familiar. For each round of extraction, the samples were thawed on ice and homogenized by vortexing, before processing according to protocol as described in section 3.2.1.2. To determine the yield and quality of the extraction, concentration was measured using a Qubit Fluorometer with the Qubit dsDNA BR Assay Kit, and all samples were run on a 1% agarose gel to determine fragment lengths and quality. In addition, absorbance was measured for half of the samples using a NanoDrop spectrophotometer to check for contamination. Methods for quality control is described in detail in section 3.5. Genomic DNA was stored at -20°C.

3.3 Amplicon sequencing

Genomic DNA was thawed on ice and homogenized by carefully flicking the tube. 2 µL of genomic DNA was transferred to new tubes and diluted to 5 ng/µL with sterile water. The amplification targeted the V3-V4 region of the 16S rRNA gene using the Pro341F/Pro805R primers pair. Samples were processed on ice 4 at a time with one negative control using additional water instead of DNA per run, using the reagent mixture as described in table 3.3.1.

Table 3.3.1: Reaction volumes of each reagent used in initial PCR amplification.

	Volume
iProof HF Master Mix	12,5 µL
Pro341F (10 µM)	1 µL
Pro805R (10 µM)	1 µL
DMSO	1µL
Genomic DNA 5ng/µL	2,5 µL
Sterile water	7 µL

The reaction was run on SensoQuest Labcycler Gradient under the conditions described in table 3.3.2.

Table 3.3.2: Thermal cycling program used in initial PCR amplification

98°C for 3 minutes	
	25 cycles of
	98°C for 30 seconds
	53°C for 30 seconds
	72°C for 30 seconds
72°C for 5 minutes	
Hold at 4°C °	

Quality was assessed by gel electrophoresis of products and negative controls for each run as described in section 3.5. Samples were stored at -20°C until all were ready for clean-up and indexing.

PCR product was purified by mixing with 0,8x volume AMPure XP beads, which were then pelleted on a magnet and washed twice with 200 µL fresh 80% ethanol. The library was resuspended in 52.5 µL of 10 mM Tris pH 8.5, before the beads were pelleted on a magnet and 50 µL from each well was transferred onto a new PCR plate as described in detail in the Library preparation protocol provided

by Illumina (https://support.illumina.com/downloads/16s_metagenomic_sequencing_library_preparation.html). To determine yield, three samples were chosen along a diagonal of the plate for concentration measurement, and three different samples were chosen at random to determine quality and confirm removal of primer-dimers by gel electrophoresis according to protocols described in section 3.5.

Due to low yields, index PCR was performed with some adjustments: the final reaction volume was decreased to 25 µL, and the amplicon DNA remained undiluted, resulting in the reaction solution described in table 3.3.3.

Table 3.3.3: Reaction volumes of each reagent used in index PCR

	Volume
iProof HF Master Mix	12,5 µL
Nextera XT Index 1 Primers	2,5 µL
Nextera XT Index 2 Primers	2,5 µL
DNA	7,5 µL

To increase yield, the number of cycles was slightly increased from 8 to 10 cycles, and index PCR was achieved using the program described in table 3.3.4.

Table 3.3.4: Thermal cycling program used for index PCR

98°C for 3 minutes	
	10 cycles of
	98°C for 30 seconds
	55°C for 30 seconds
	72°C for 30 seconds
72°C for 5 minutes	
Hold at 4°C °	

The second round of clean up was performed using 1,12x volume AMPure XP beads and washed twice in fresh 80% like in the first round. After washing, beads were resuspended in 27.5 µL of 10 mM Tris pH 8.5, pelleted by magnet and 25 µL from each well was transferred onto a new PCR plate.

Concentration was fluorometrically measured for all samples after the second clean-up as described in section 3.5 and three samples with low yields were indexed and cleaned up a second time to ensure sufficient DNA before normalization. All DNA was diluted with sterile water to 4 nM in a total volume of 50 µL per sample. Pooling of samples for sequencing was accomplished by combining 40 µL from each sample into a single 10 mL falcon tube and concentration was controlled with Qubit™ fluorometer.

Library DNA and PhiX Control were both denatured using freshly diluted 0.2 M NaOH according to the protocol and diluted to a final concentration of 6 pM using pre-chilled HT1 from the MiSeq Reagent Kit v3. The final library was combined by adding 120 µL PhiX control to 480 µL amplicon library, for a total of 600 µL sequencing library with 20% PhiX control spike-in. Heat denaturing was performed according to protocol, and all 600 µL were loaded onto the MiSeq reagent cartridge shortly after the recommended 5 minutes on ice.

3.4 Troubleshooting PCR

PCR amplification was first attempted following the Library preparation protocol available at from Illumina the support website, however, the 2x KAPA HiFi HotStart ReadyMix was exchanged with the iProof HF Master Mix, which has a higher activation temperature, and therefore all cycles of 95°C in the protocol were adjusted to 98°C. After low yields in the first round of PCR, several adjustments were tested to maximize the amount of DNA available for sequencing.

The polymerase was replaced with a new one, and new dilutions of the primers were made, while an order was placed for new ones. During the wait for new reagents, several adjustments to the protocol were tested. This included a concentration assay with different starting concentrations ranging from 5 µL to 20 µL, as well as the inclusion of positive controls. An assay adding different stabilizing compounds with variations in additive concentration, as well as DNA concentrations was run as illustrated by table 3.4.1.

Table 3.4.1: Reaction solution for 12 different samples using varying concentration of both template DNA and stabilizing compounds

	Genomic DNA (5ng/µL)	BSA	MgCl2	DMSO	Polymerase	Pro341F (10µM)	Pro805R (10µM)	Water
1	1	1	-	-	12,5	1	1	8,5
2	1	-	1	-	12,5	1	1	8,5
3	1	-	-	1	12,5	1	1	8,5
4	1	2	-	-	12,5	1	1	7,5
5	1	-	2	-	12,5	1	1	7,5
6	1	-	-	2	12,5	1	1	7,5
7	2,5	1	-	-	12,5	1	1	7
8	2,5	-	1	-	12,5	1	1	7
9	2,5	-	-	1	12,5	1	1	7
10	2,5	2	-	-	12,5	1	1	6
11	2,5	-	2	-	12,5	1	1	6
12	2,5	-	-	2	12,5	1	1	6

The sample solution shown to give the highest yield as assessed by gel electrophoresis was then brought forward for further optimization. Optimal annealing temperature for the new solution was determined using a gradient assay of temperatures at 50°C, 53°C, 55°C, 57°C and 60°C.

Initially, once reaction conditions were deemed satisfactory, samples were processed using PCR strips of 8 tubes, 6 of which were used for samples, as well as one for a negative control, for a total of 12 samples per run of the thermal cycler. It was discovered however, that by limiting processing time and only running 4 samples per run in individual PCR tubes resulted in better yields, and thus, all samples was processed this way.

3.5 Assessing yield and quality of DNA

To determine the yield of DNA in all steps described earlier in this chapter, concentration was measured utilizing a Qubit™ 1 Fluorometer with the Qubit® dsDNA BR Assay Kit, which is highly specific for double stranded DNA over RNA and single stranded DNA. Working solutions were prepared by diluting the Qubit® dsDNA BR Reagent using Qubit® dsDNA BR Buffer in 1:200 ratio. The fluorometer was calibrated at the start of the first assay of any given day, and standards were prepared by adding 10 µL of each standard into Qubit® assay tubes containing 190 µL working

solution. Sample concentration was measured using 2 µL of the sample combined with 198 µL working solution, and each assay contained at least one negative control.

Fragment length and relative yield was assessed by gel electrophoresis using the Bio-Rad PowerPac™ Basic Power Supply together with Mini Sub Cell GT system. Genomic DNA was run on 1% agarose gel, made by combining 0,5 g Seakem LE Agarose with 50 mL TAE buffer and heating carefully in a microwave, before 2 µL PeQgreen DNA/RNA dye was added to the mixture and then cast with a 8-well comb. Samples were diluted using 5 µL DNA and 4 µL sterile water and dyed with 1 µL Gel Loading Dye, Blue (6X). The Quick-load®, Purple 1 kb DNA ladder was loaded into the far-left well for all gels, and the gels were run at 90 V for 20 minutes. PCR product was assessed using 1,5% agarose gels, prepared in the same manner as those used for genomic DNA, except for the increased amount of 0,75 g agarose and cast using a 15-well comb. For each sample, 1 µL of product was diluted using 4 µL sterile water and dyed using 1 µL loading dye. The 1 kb DNA ladder was loaded into the left-most well and gels were run at 70 V for 40 minutes.

UV spectrophotometry with NanoDrop ND-1000 Spectrophotometer was used to determine purity of DNA prior to sequencing. Baselines were determined by blanking with the same buffer as DNA was dissolved in, and 1,5 µL of each sample was used to measure absorbance for 260/280 and 260/230 ratios. Measurement pedestals were cleansed with nuclease free water between each measurement.

3.6 16S analysis using DADA2 and Phyloseq in R

Analysis was first attempted by following the DADA2 pipeline tutorial (1.12) on github (http://benjjneb.github.io/dada2/tutorial.html). Due to computational considerations, a subset of six samples was used to determine the parameters best suited to the dataset. Filtering was performed using standard parameters and truncated using *truncLen = c(285, 260)* to balance quality and still conserve overlap between forward and reverse reads, as well as *trimLeft c(17, 21)* to remove primers. After these parameters produced insufficient merging, several adjustments were tested, both by increasing read length after truncation, as well as increasing the number of expected errors, *maxEE*, for reverse reads and both forward and reverse reads respectively. The minimum overlap for the *mergePairs* function was adjusted to 8 bp, as well as relaxed to allow 1 mismatch. As neither of these adjustments resulted in merge rates above 40%, it was ultimately decided to use only forward reads in the analysis. Due to computational restrictions, the Big Data pipeline described on github (http://benjjneb.github.io/dada2/bigdata.html), was applied to the complete dataset, separating fluid and particle phase samples into separate "runs" and subsequently merging the sequence tables prior to assigning taxonomy. Filtering was done using *maxEE = 2* and *truncQ = 2*, and reads were truncated at 285 bp. The Phyloseq package was applied to visualize results. The relative diversity between samples was assessed by non-metric multidimensional scaling, a method which when used to generate a two-dimensional scatterplot, causes more similar data points to cluster together. This was achieved by using the function *ordinate()*, inputting the dataset with ASV-proportions and their assigned taxonomy for each sample, with arguments *method = "NMDS"* and *distance = "bray"* to create an ordination object based on a Bray-Curtis distance matrix and then generating a plot using *plot_ordination()*. A bar plot was created by sorting the 50 most common taxa using function *names(sort(taxa_sums(data), decreasing=TRUE))[1:50]*, the output of which was used in the function *prune_taxa()* to generate a subset containing only the top 50 most abundant ASVs, which was then plotted using *plot_bar()*.

To determine which samples would be sent to the Norwegian Sequencing Centre in Oslo, the relative taxonomic diversity of the most abundant sequences was examined in each sample individually. Those samples that displayed higher diversity within abundant ASVs were preferentially selected,

and in order to achieve a holistic view of the total communities in downstream analysis, it was decided that both the fluid- and particle phase from the same sheep would be used and that an equal number of animals from each feed group would be included. In the end, half of all the 48 the samples were selected for shotgun sequencing and sent to Oslo.

3.7 Nanopore sample preparation and sequencing

Given the time constraints of this M.Sc. candidature as well as the waiting times at the Norwegian Sequencing Centre, biological samples were switched for the remainder of the project; from rumen samples to a human fecal consortia that was already sequenced using the Illumina HiSeq3000. For shotgun metagenomics using long read technology (Oxford Nanopore sequencing), high quality DNA extracted from two human stool enrichment samples using the method outlined above in Section 3.2.1.1 was utilized. Library preparation and loading was performed for the samples XDC03 and XDCOriginal in parallel following the Genomic DNA by Ligation (SQK-LSK109) protocol available from the Nanopore Community. Both samples and all reagents were thawed and kept on ice during processing. Samples were homogenized by flicking and briefly spun down, before appropriate volumes of each sample were transferred into Lo-bind tubes such that each tube contained 1000 ng of HMW DNA, and total volume was adjusted to 49 µL using sterile water. The DNA repair and end-prep reaction solution was combined into new LoBind tubes in place of PCR-tubes and incubated at 20°C for 5 minutes and 65°C for 5 minutes using two separate Eppendorf Thermomixer C's. Clean-up was performed by adding 60 µL resuspended AMPure XP beads and homogenizing by flicking. Samples were incubated on Biosan Multi RS-60 programmable rotator at 11 rpm for 5 minutes, and briefly spun down before pelleting on magnetic stand. Once clear, supernatant was removed, and beads were washed twice in 200 µL fresh 70% ethanol without disturbing the pellet. The beads were resuspended in 61 µL sterile water and incubated at room temperature for 5 minutes before pelleting on magnet and transferring eluate into a new LoBind tube. Adapter ligation mixture was combined as described in protocol into eluate-containing tubes and incubated at room temperature for 10 minutes. Purification was performed using 40 µL AMPure XP beads and incubating on the programmable rotator at 11 rpm for 5 minutes. The mixture was briefly spun down and pelleted on the magnetic rack. Supernatant was removed and beads were washed twice by resuspending in 250 µL Long Fragment Buffer and re-pelleting on the magnet, then removing buffer. Finally, beads were resuspended in Elution Buffer and incubated at 37°C with a Thermomixer for 10 minutes, before pelleting and transfer of eluate into new LoBind-tubes.

Flow cells were brought to room temperature and inserted into each MinION sequencer, nicknamed "Rosalind" and "Esther" respectively. Priming was completed by removing a few microliters of buffer through the priming port by carefully turning the wheel of a P1000 pipette set to 200 µL. The priming mix was prepared and 800 µL were loaded into the priming port and left for 5 minutes, while the final library mix was prepared. An additional 200 µL of priming mix was added through the priming port while sample port was open, before final library was mixed by careful pipetting and loaded dropwise into sample port. Both ports and the MinION Mk 1B lid were closed and sequencing was initiated. "Rosalind" and "Esther" sequenced for a predetermined 6 and 7 hours respectively.

3.8 EPI2ME workflows for Nanopore reads

3.8.1 Taxonomy

The output fastq-files from each sequencing run were processed with the FASTQ WIMP workflow using the EPI2ME desktop agent available from the nanopore community. Filtering parameters allowed reads of all lengths, with a Q-score cutoff for reads where $Q < 7$.

3.8.2 Alignment to MAGs

A selection of the most complete MAGs (>95% completeness) available from a previous Illumina sequencing run using the same DNA as those sequenced by the MinION were uploaded as refence sequences using the EPI2ME desktop agent with the Fasta Reference Upload workflow. Nanopore reads were uploaded from each sequencing run and aligned to each MAG individually with the FastQ Custom Alignment workflow. Filtering parameters allowed reads of all lengths, with a Q-score cutoff for reads where Q < 7.

3.9 Gene calling and annotation

3.9.1 Gene calling

Three MAGs were chosen for gene calling and annotation based on nanopore alignment results. Gene calling was performed using the web-based platform MetaGeneMark (http://exon.gatech.edu/meta_gmhmmp.cgi) by uploading fasta files directly in browser, with output settings to produce protein sequences in LST format.

3.9.2 Annotation

Protein sequence fasta files from MetaGeneMark were edited to fit format requirements, and all three MAGs were annotated with web-based services. CAZymes were predicted using dbCAN (http://bcb.unl.edu/dbCAN2/blast.php), and GhostKOALA (https://www.kegg.jp/ghostkoala/) was used to reconstruct metabolic pathways for glycolysis/gluconeogenesis and propanoate metabolism and assign E.C. numbers.

4 Results

4.1 Extraction

4.1.1 Manual method vs Kit – Dummy sample comparison

In order to determine which extraction method should be used for the analytical samples, two approaches were tested: one in which cells were lysed chemically and purified with phenol:chloroform, and one kit extraction in which cells were lysed by bead beating and purified using a spin column. The respective results were compared on the basis of total yield as determined by fluorometry, and fragment length as determined by gel electrophoresis. Table 4.1.1.1 shows that the total yield of DNA was comparable for the two extraction methods, with the highest overall yield achieved by kit extraction from the particle-phase sample, however more biomass was needed to achieve this. The agarose gels (figure 4.1.1.1) show that the fragment length of the DNA is generally higher for the manual extraction that uses phenol: chloroform-purification, as the HMW DNA bands are fully located above the Quick-Load® Purple 1 kb DNA Ladder, indicating that fragment length is <10 kb. However, additional bands are also shown beneath the ladder in all samples extracted manually, suggesting the presence of some kind of degraded fragments as well. The bands from the kit extraction has an upper limit slightly higher than that of the ladder, but the bands are more “smeared” suggesting less uniform fragment length in the range of 5-10 kb.

Sample	Biomass used (g)	Total DNA yield (µg)
Manual-Particle	0,6	12,99
Manual-Fluid	0,6	10,77
Kit-Particle	0,25	13,80

Kit-Fluid	0,25	9,72

Table 4.1.1.1: Biomass used, and total DNA yields obtained from particle- and fluid-phase samples using two different extraction methods.

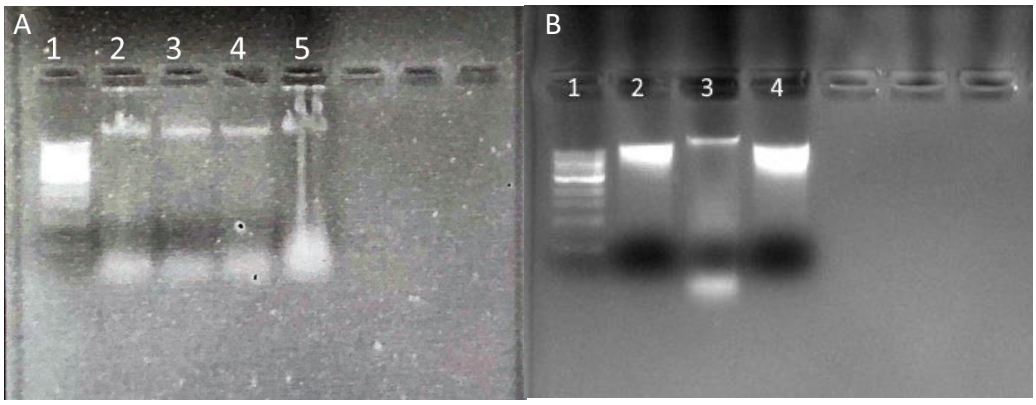


Figure 4.1.1.1: 1% agarose gel with 1 kb ladder showing A) DNA extracted by the manual phenol:chloroform method from particle phase- (lanes 2-3) and fluid phase samples (lanes 4-5). HMW DNA bands are located above the top of the ladder, signifying high molecular weight, however additional bands below the ladder suggests the presence of degraded nucleic acids as well. B) DNA extracted by kit from particle phase- (lane 2) and fluid phase samples (lane 4) as well as one particle phase sample extracted using the phenol:chloroform method (lane 3). DNA from kit extractions show slightly lower and less uniform fragment lengths as compared to the manual phenol:chloroform, but did not result in any additional bands formed by presence of highly degraded nucleic acids.

4.1.2 Extraction of analytical samples using the Qiagen DNeasy PowerSoil Kit

The analytical samples were extracted using the Qiagen DNeasy PowerSoil Kit. To determine the yields of each extraction, concentration of all samples was measured fluorometrically and an overview of results is shown in Table 4.1.2.1. Although all samples were handled in the same way, the results were quite varied, with the most significant differences found in the particle phase samples, with almost 10 µg separating the highest and lowest total yields. Full details of extraction results is available in Appendix A.

Table 4.1.2.1: Statistic summary of concentration (ng/µL)-yields from DNA extraction of all 48 analytical samples

Sample type	Average	Median	Standard deviation	Maximum	Minimum
Fluid phase (24 samples)	65,44	64,75	11,87	91,10	45,10
Particle phase (24 samples)	92,17	88,05	23,97	136,00	37,40
Total (48 samples)	78,80	75,20	23,08	136,00	37,40

To assess the purity of the extracted DNA, UV spectrophotometry was utilized to check for the presence RNA, proteins and other chemical contamination in half of the samples: with an equal number of samples from fluid- and particle phase, each with four representative samples from the three diets. The absorbance ratios measured by spectrophotometry is summarized in table 4.1.2.2. Most samples had A260/280-ratios close to the average, as illustrated by a relatively small standard deviation, and thus most samples were within the desired range. The A260/230-ratios varied slightly

more within a few samples, with four samples slightly below the desired 1,8 threshold, although most were close to the average value.

Table 4.1.2.2: Summary of absorbance-ratios as measured by NanoDrop spectrophotometry from 24 samples: 12 fluid phase- and 12 particle phase samples evenly distributed across the three diets.

Measurement	Average	Median	Standard deviation	Maximum	Minimum
A260/280	1,88	1,86	0,07	2,14	1,83
A260/230	1,89	1,92	0,12	2,13	1,58

4.2 Marker gene sequencing

4.2.1 Amplicon PCR

In order to analyze the microbial community of both the liquid- and particle phase of the sheep rumen samples when fed different diets, 16S rRNA gene amplicon sequencing was used to assess the community structure. This required the PCR amplification of the 16S rRNA gene which was accomplished using the Pro341F/Pro805 primer pair, targeting the V3-V4 region and reaction conditions outlined in section 3.3.

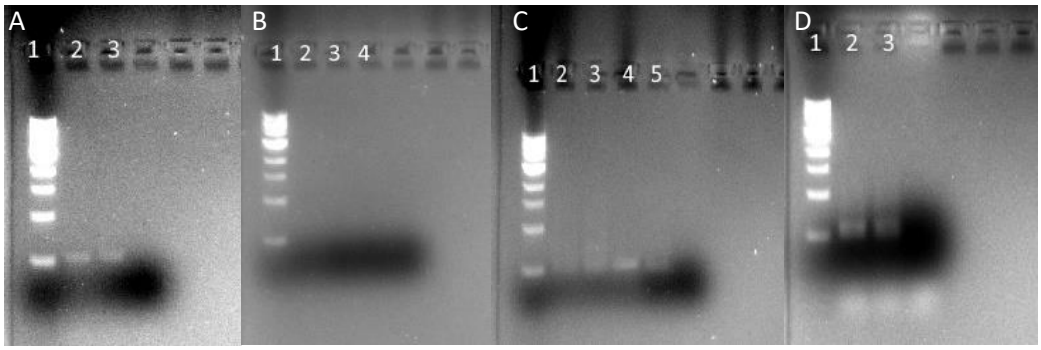


Figure 4.2.1.1 Gel showing yield after A: following protocol B: Concentration gradient assay (5, 10 and 20 ng/µL): C: Positive controls (lanes 4-5) D: New primers

As shown by Figure 4.2.1.1, after the initial round of 16S rRNA gene amplification, PCR products were recovered with much lower yields than expected (A), which resulted in several adjustments to reaction conditions being made. The concentration assay of template DNA that was used in the PCR reaction with starting concentration at 5, 10 and 20 ng/µL (respectively) showed no visible yield, and the addition of positive controls (C) showed minimal amplification regardless of DNA origin, however, a slightly more intense band (lane 4) may indicate a somewhat more successful amplification of one of the controls. Once new primers arrived, new dilutions of these were also made and tested, but did not significantly improve the results (D).

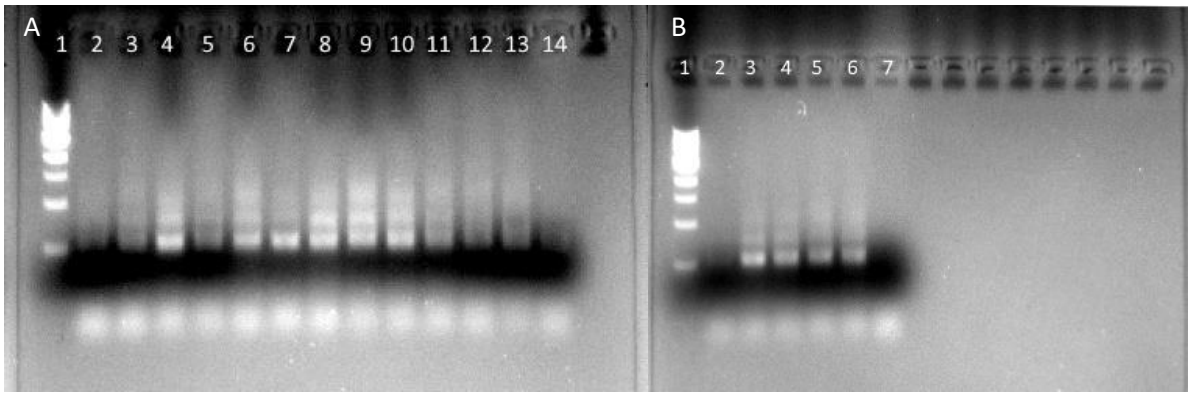


Figure 4.2.1.2 Gel showing A: Additive assay with 1 µL (lanes 2-7) and 2,5 µL (lanes 8-13) of template DNA (5 ng/ µL), 1 µL of BSA (lanes 2, 8), $MgCl_2$ (lanes 3, 9) and DMSO (lanes 4, 10) and 2 µL BSA (lanes 5, 11), $MgCl_2$ (lanes 6, 12) and DMSO (lanes 7, 13) B: Annealing temperature gradient assay with annealing at 50°C (lane 2), 53°C (lane 3), 55°C (lane 4), 57°C (lane 5) and 60°C (lane 6)

As the first set of adjustments made to the reaction conditions made little difference in terms of increasing the yield of PCR product, the effects of a variety of additives known to aid in stabilizing PCR were compared in the assay depicted in Figure 4.2.1.2A. Bovine serum albumin (BSA), magnesium chloride ($MgCl_2$) and dimethyl sulfoxide (DMSO) were added in 1 µL and 2 µL respectively to reaction mixtures with 1 (lanes 2-7) and 2,5 (lanes 8-13) µL of 5 ng/ µL genomic DNA in same order as described in Table 3.4.1. DMSO (lanes 4, 7, 10, 13) was determined to give best yield increase overall, and the reaction mixture with 1 µL DMSO and 2,5 µL genomic DNA was chosen for further optimization. Figure 4.2.1.2B shows annealing temperature gradient assay where 53°C (lane 3) was determined to be the optimal annealing temperature.

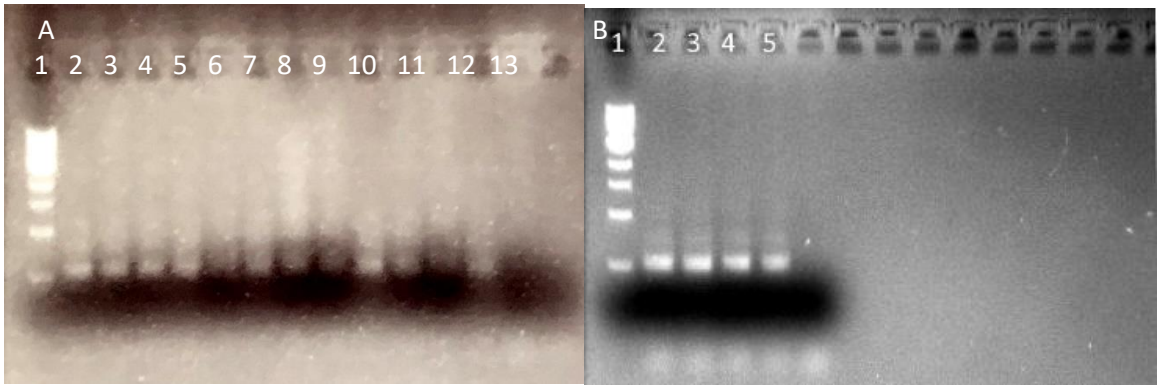


Figure 4.2.1.3 Gel showing yield when samples were run A: 12 at a time using two PCR-strips B: 4 at a time using individual PCR-tubes

After optimizing the reaction conditions through several rounds of troubleshooting, the rumen samples were processed 12 at a time, using PCR strips with six samples and one negative control in each strip. It was discovered however, that by using single PCR-tubes and only processing 4 samples and one positive control, thereby cutting the preparation time by more than half, resulted in higher yields of desired PCR product and decreased the amounts of observed non-specific product. Figure 4.2.1.4 illustrates the differences observed when amplifications were performed using exactly identical reaction conditions except for one: samples in A were processed using PCR-strips with a total of 12 samples per run, whereas samples in B were processed in individual PCR-tubes with a total of 4 samples per run. After this discovery, the 48 samples were processed 4 at a time, sequencing adapters and indexes were added to the cleaned PCR products through another round of

PCR. All cleaned PCR products were normalized to 4 nM and utilized for generating a 6 pM sequencing library with 20% PhiX control spike-in and which was then sequenced using an Illumina MiSeq system with a run time of 48 hours.

4.2.2 Amplicon analysis in DADA2

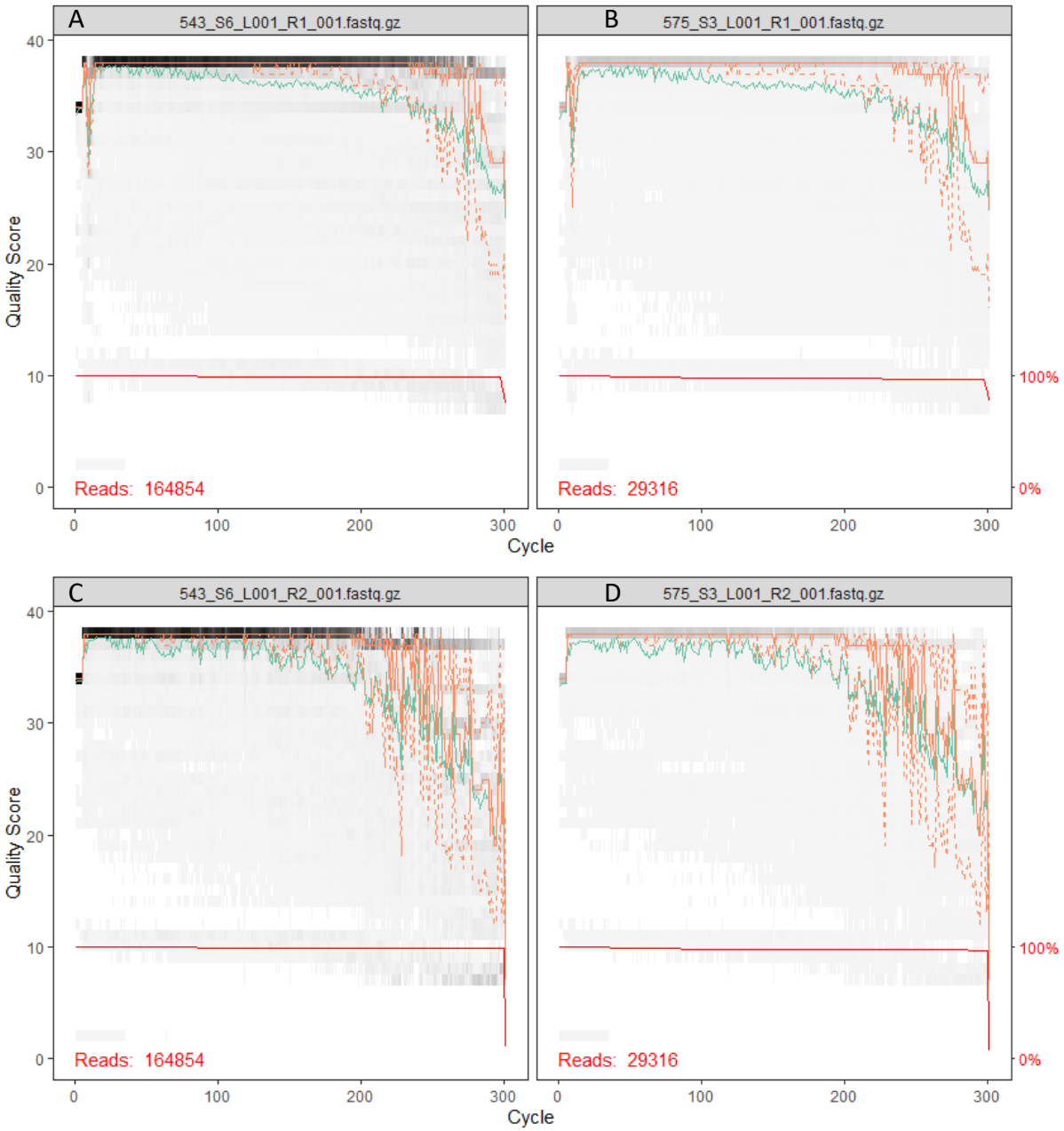


Figure 4.2.2.1: Quality profile of A: Forward reads from fluid sample B: Forward reads from particle sample C: Reverse reads from fluid sample D: Reverse reads from particle sample

The Illumina MiSeq sequencing of the 48 samples yielded a total of 4 363 677 paired end reads (2x300) for downstream analysis in DADA2. Figure 4.2.2.1 shows quality plots for the forward and reverse reads from one fluid and one particle sample. The average quality of the forward reads is relatively stable and only drops somewhat towards the end of the reads. For the reverse reads

quality is more variable and the average quality shows significant decrease from around the 200th cycle.

The analysis yielded a total of 32 530 amplicon sequence variants (ASVs), across a variety of taxa as summarized in table 4.2.2.1. Although a few ASVs were assigned to Eukaryota and Archaea, >99% were determined bacterial. Phyla Bacteroidetes and Firmicutes represented the largest number of ASVs, with 42% and 35% of the total number respectively. Only 34% of reads were determined at the genus level, representing 220 different genera, whereas 65% of all sequence variants were assigned at the family level, distributed across 99 different families.

Table 4.2.2.1: Summary of the diversity of taxa observed in amplicon sequence variants

	Kingdom	Phylum	Class	Order	Family	Genus
Number of taxa	3	23	35	65	99	220
ASVs n/a at this level	9	3473	4511	6209	11594	21442

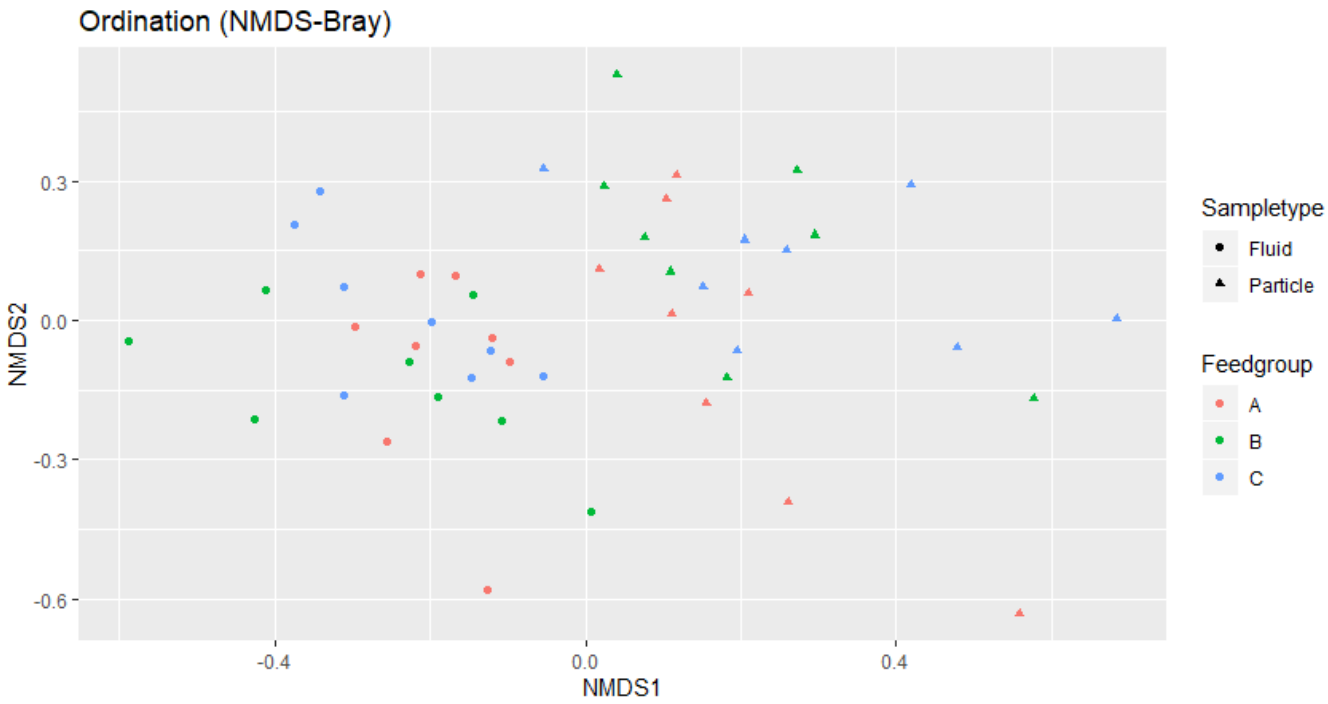


Figure 4.2.2.2: NMDS-plot using Bray-Curtis distances to visualize relative distance between samples in two dimensions. Round data points belong to samples extracted from the fluid phase of the rumen, and triangular data points belong to particle phase samples. The different diets, denoted as A, B and C (0, 5, 2.5% of seaweed respectively) are represented by the color of the data point.

To visualize the potential differences between samples, non-metric multidimensional scaling (NMDS) was used to generate an ordination plot. This technique translates information about pairwise dissimilarity in multivariable data as described by a distance matrix into points in a low-dimensional space, such as the two dimensions shown in Figure 4.2.2.2. As the length of distances between each data point represents the level of similarity between each sample, any clustering of samples from similar origin might indicate a systematic difference in microbial composition between these and other samples. No clear clustering can be observed between samples recovered from sheep digesting the different diets, including the control diet (A) and those that contained 2,5 (C) and 5% (B) of added

sugar kelp. However, clear separation was observed between samples analyzed from the fluid phase and those recovered from the particle phase.

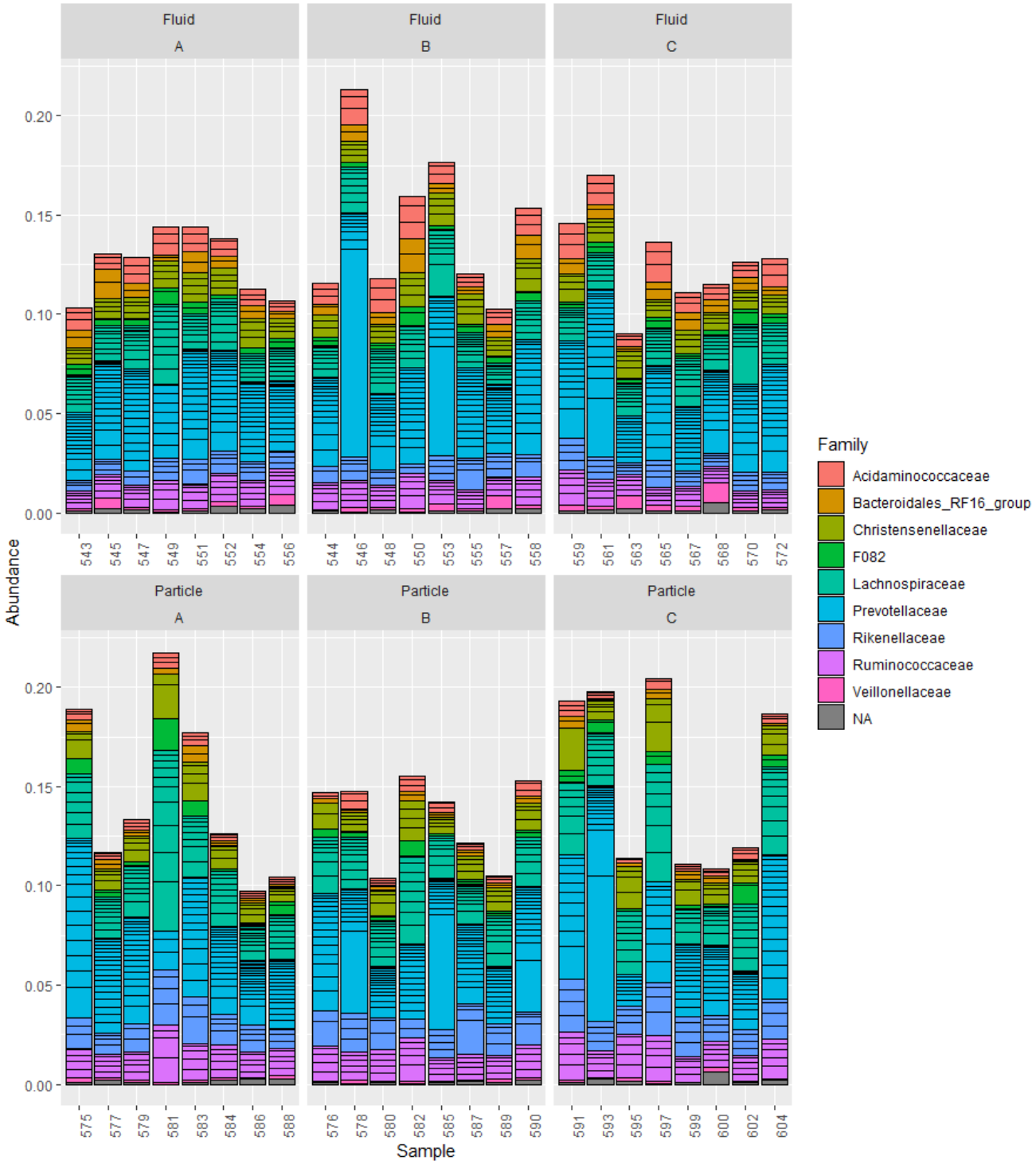


Figure 4.2.2.3: Taxonomic distribution across samples of the 50 most abundant amplicon sequence variants (ASVs) observed in the analysis, divided by sample type and feed groups.

Looking at the taxonomic distribution across the rumen samples analyzed in this study, Figure 4.2.2.3 shows that the abundance of the 50 most common amplicon sequence variants (ASVs) tend to cover between 10-20% of the total abundance in the samples. The distribution of families is mostly similar across all samples, with Prevotellaceae and Lachnospiraceae as the most prominent. Small differences can be observed between fluid- and particle phase samples, such as a somewhat higher prevalence of Acidaminococcaceae in fluid phase samples. A small number of the most abundant ASVs were not assigned to any family. As these distributions displayed only minor differences in community composition, it was ultimately decided that the fluid- and particle phase samples from the same animals would be chosen from four individuals in each feed group for shotgun sequencing at the Norwegian Sequencing Centre in Oslo, amounting to a total of 24 samples.

4.3 Whole metagenome/shotgun sequencing

4.3.1 Metrics

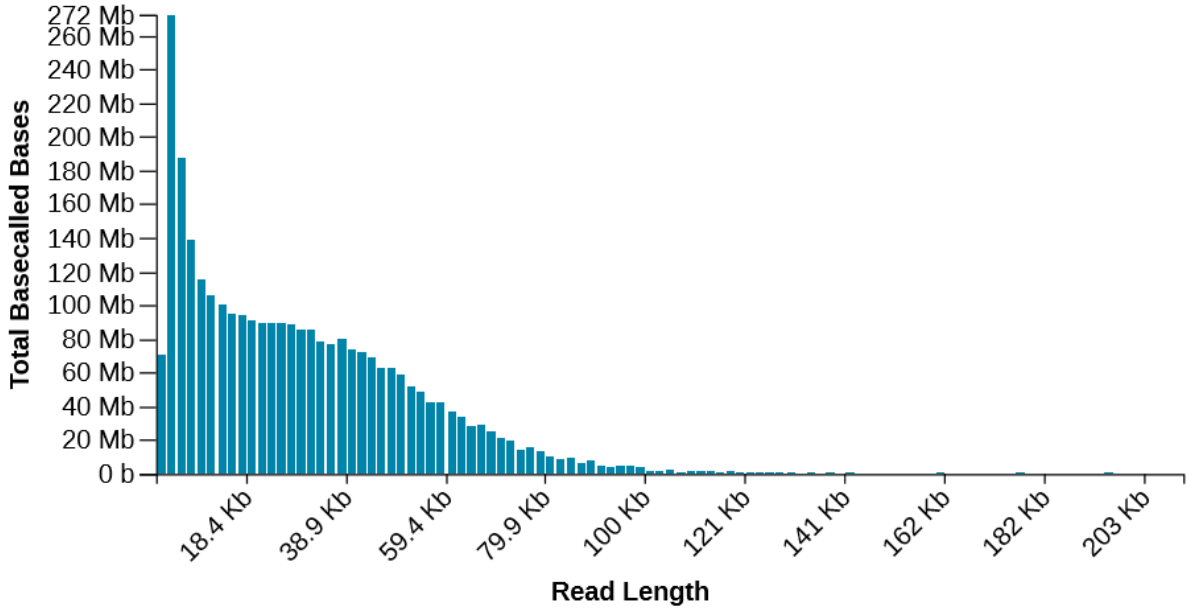
Whilst the first section of this thesis analyzed the community structure in rumen samples, a shotgun metagenomics approach was used to assess community function in gut communities. As outlined in the methods, two alternate DNA samples (XDC03 and XDCOriginal) were used for shotgun sequencing (due to time constraints) and were sequenced using long read technology to supplement the Illumina-generated MAGs that were previously generated from this particular sample set (Table 4.3.1.1). Approximately 350M pair-end reads were utilized in metagenomic assembly of the Illumina data, yielding a total of 29 MAGs from the two samples. Of these, 7 and 6 MAGs from XDC03 and XDCOriginal respectively were estimated to be >90% complete. More detailed assembly statistics for all MAGs are available in Appendix B.

Table 4.3.1.1: Assembly statistics for MAGs generated using HiSeq3000 Illumina data

Sample	Approx. no. reads (per end)	Approx. no. bp	Total assembled MAGs	No. MAGs >90% completeness	Avg. genome size (>90% completeness)
XDC03	350M	90 Gb	14	7	7,7 Mbps
XDCOriginal	350M	90 Gb	15	6	5,3 Mbps

Table 4.3.1.2: Sequencing metrics from two runs on individual Oxford Nanopore MinION sequencers

Sequencing run	Run time (h)	Total generated reads	Pass reads	Avg. read length (bp)	Avg. Q score
Esther (XDC03)	7	325,86 K	300,30 K	9 843	11,20
Rosalind (XDCOriginal)	6	346,05 K	301,66 K	10 743	11,18



The two DNA samples (XDC03 and XDCOriginal) were sequenced on two individual MinION flow cells (Table 4.3.1.2) with the distribution of read lengths of basecalled bases from the "Esther" sequencing run shown in Figure 4.3.1. A total of 325.86 K reads were generated from 7 hours of sequencing with an estimated N50 of 25.13 Kb and the longest read approximately 200 Kb in length.

4.3.2 Taxonomy

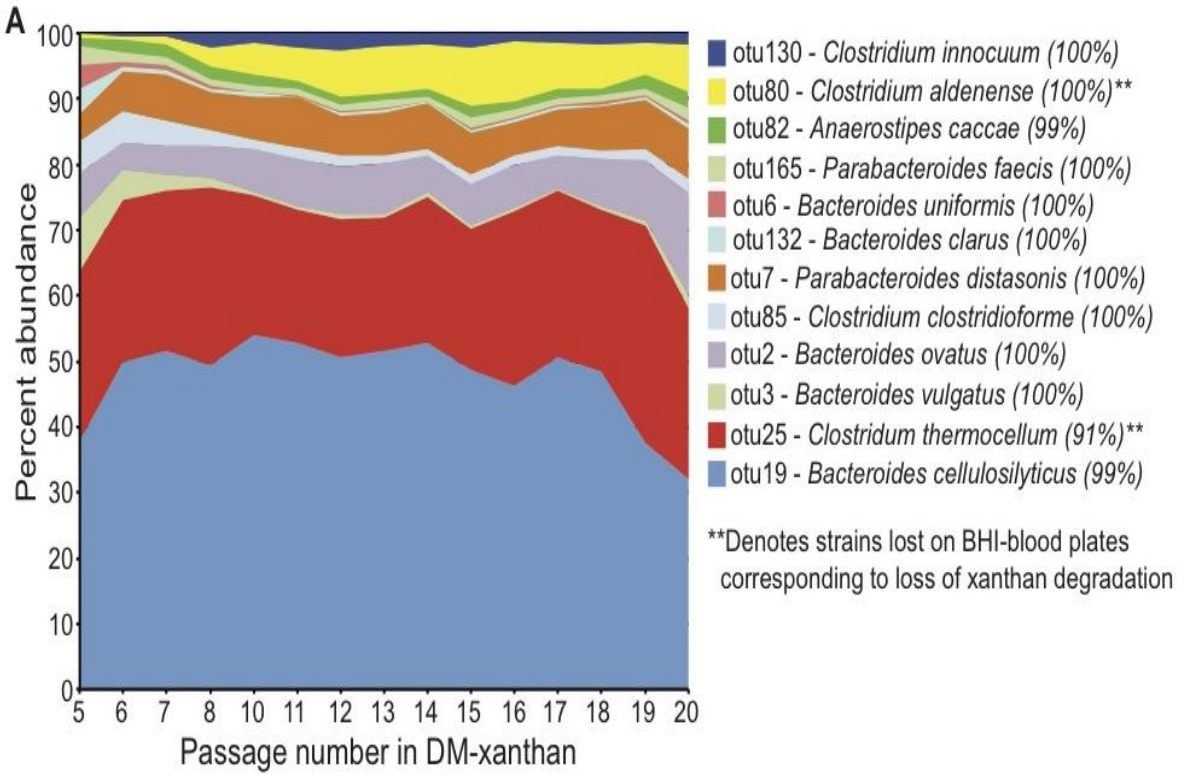


Figure 4.3.2.1: Taxonomic distribution as determined by 16S analysis of original sample. (determined by 16S analysis) of a representative sample from the human fecal enrichment that was used for long read shotgun metagenomes. Figure obtained courtesy of Dr. Sabina Leanti La Rosa (unpublished)

Previous 16S rRNA gene amplicon analysis of XDCOriginal (Figure 4.3.2.1) illustrated that populations affiliated to *Bacteroides cellulosilyticus* were the most abundant, followed by *Clostridium thermocellum*, with these two taxa making up approximately 70% of the sample.

To assess the taxonomy of the unassembled reads from both MinION runs, the data were analyzed using EPI2ME. From each run 60,3% and 69,4 % of all reads from Esther (XDC03) and Rosalind (XDCOriginal) respectively were successfully assigned taxonomy as described by Table 4.3.2.1. Some host contamination was detected, more so in the Esther run than Rosalind, with 993 compared to 394 reads identified as *Homo Sapiens*. The majority of reads were classified as belonging to one of three bacterial phyla: Bacteroidetes, Proteobacteria and Firmicutes, although low levels (<0,5%) of several other phyla such as Actinobacteria, Fusobacteria and Spirochaetes were also detected. Taxonomic diversity of samples with minimum abundance cutoff at 0,5% are shown in figures 4.3.2.2 and 4.3.2.3.

Table 4.3.2.1: Summary of whole metagenome taxonomic classification

Sequencing run	Reads analyzed	Reads assigned taxonomy	Bacteria	Eukaryota	Archaea	Viruses	Contamination
Esther	300 303	60,3 %	173 444	2654	311	113	0,5 %
Rosalind	301 661	69,4 %	195 148	1041	95	151	0,19 %

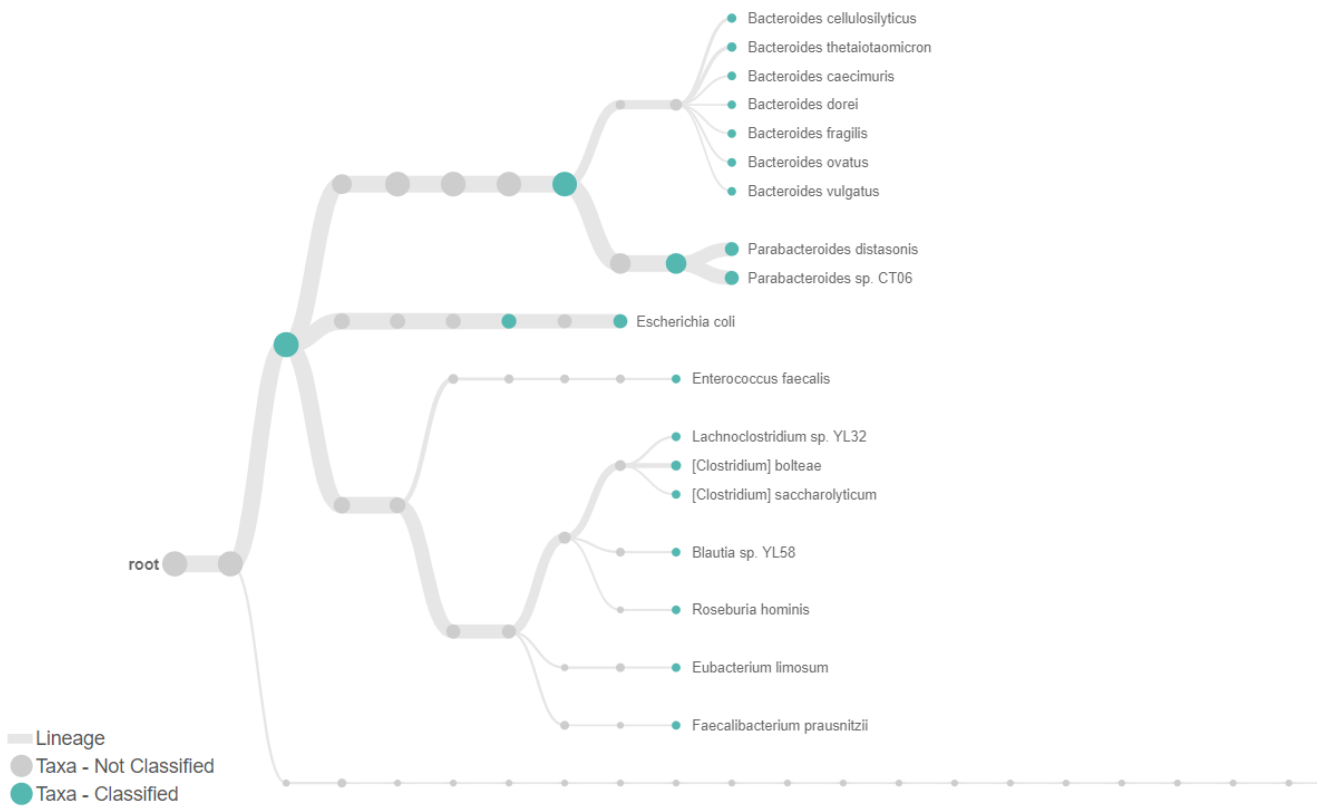


Figure 4.3.2.2: Taxonomic distribution of >0.5% abundant species in Esther sequencing run

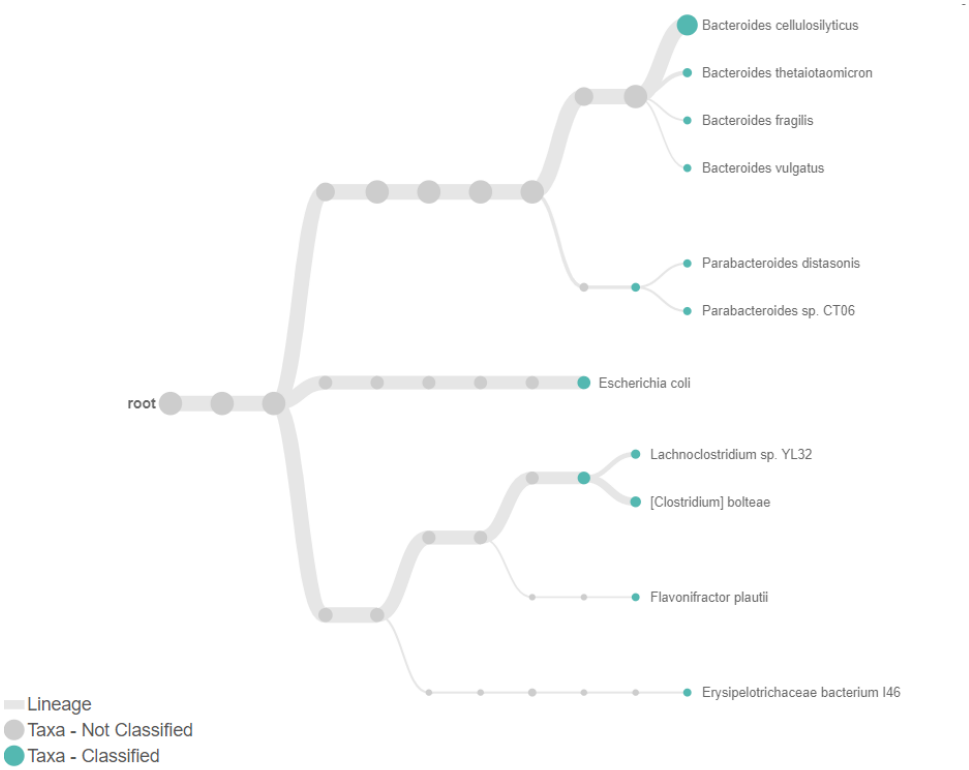


Figure 4.3.2.3: Taxonomic distribution of >0.5% abundant species in Rosalind sequencing run

Although more reads in total were assigned from Rosalind, reads from Esther shows a higher number of sequences belonging to both Eukaryota and Archaea. Most eukaryotic reads were determined to belong to kingdom Fungi with 1460 and 577 reads respectively in each sequencing run, showing a wide variety of species present in the samples, as illustrated by figure 4.3.2.4.

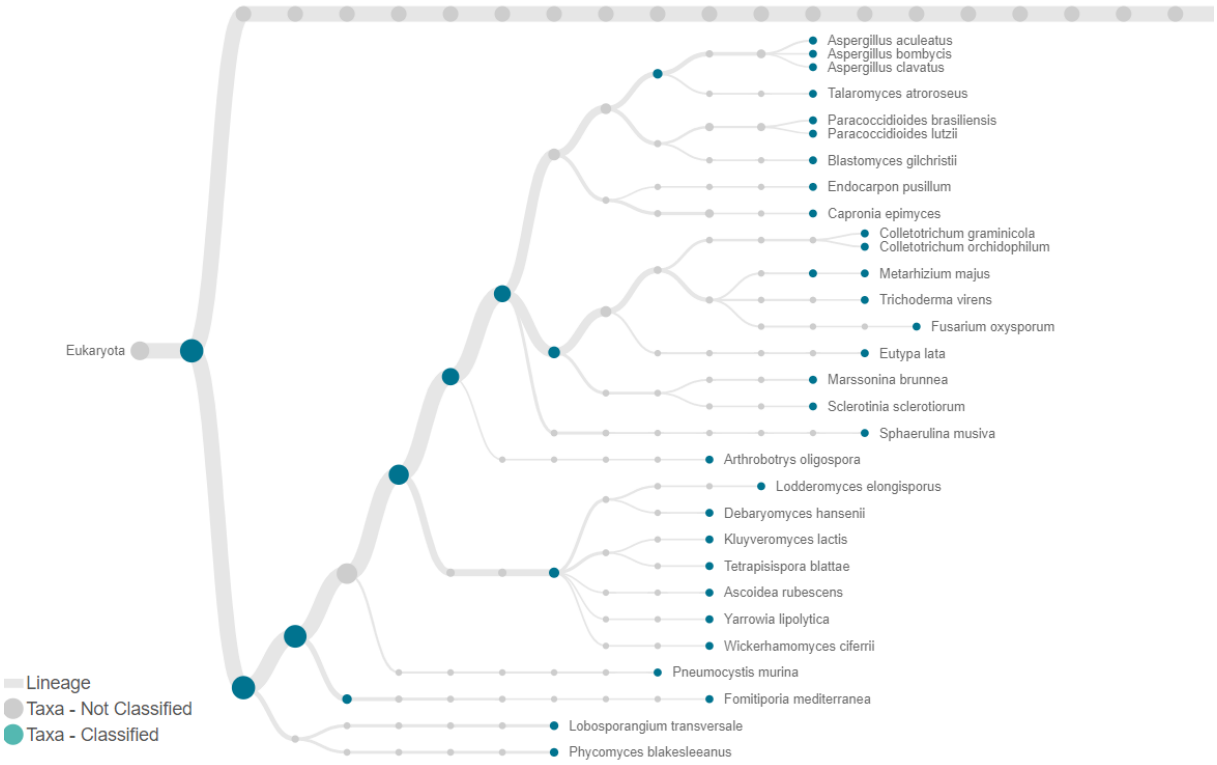


Figure 4.3.2.4 : Taxonomy tree showing diversity of fungi present in the 30 most common eukaryotic taxa in Esther sequencing run with minimum abundance cutoff 0.1%

A variety of archaea previously not detected in 16S analysis were also present in both samples as illustrated by figure 4.3.2.5.

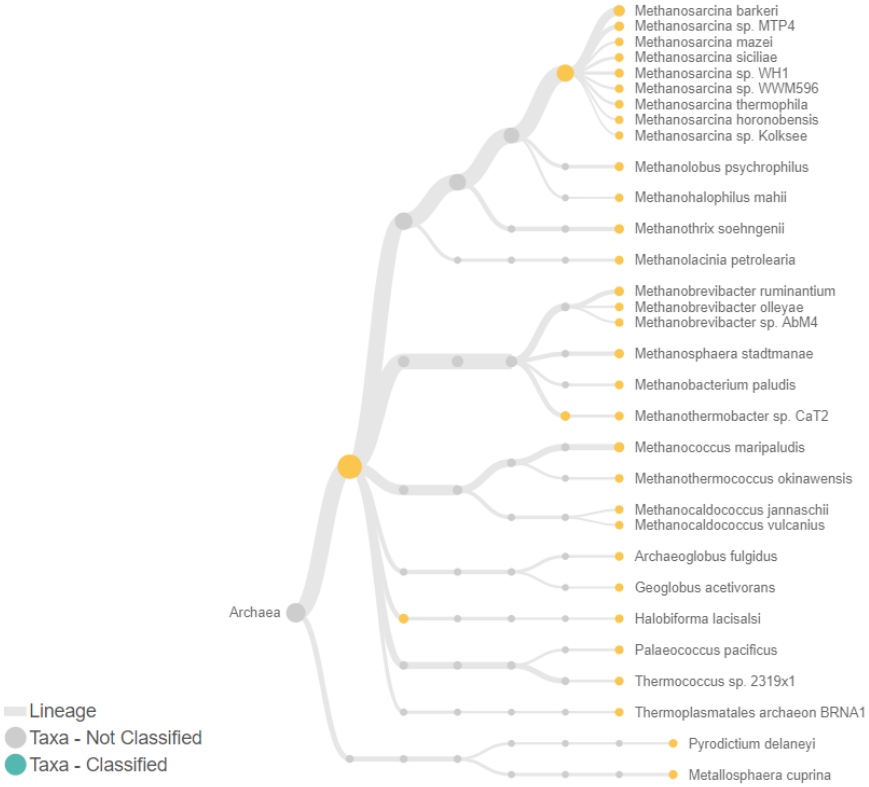


Figure 4.3.2.5 : Taxonomy tree showing diversity of archaea present in the 30 most common archaeic taxa in Esther sequencing run with minimum abundance cutoff 0.1%

4.3.3 Alignment

To compare the MinION data against short read Illumina data generated from the same samples (XDC03 and XDCOriginal long-reads were aligned against selected MAGs (constructed from Illumina). Of the selection of MAGs used for alignment, three yielded high-quality alignments, with average alignment identities >96% and consistent coverage across all or most contigs above 30x. These MAGs were previously assigned species based on closest relative, and will henceforth be referred to as *Bacteroides intestinalis, Parabacteroides distasonis* and *Escherichia coli* based on this. The *P. distasonis*- and *E. coli*-MAGs showed best alignment to reads from Esther, whereas the *B. intestinalis* aligned best to Rosalind reads. Results are summarized in table 4.3.3.1 and figure 4.3.3.1

Table 4.3.3.1: Summary of alignment results for the three best aligned MAGs. MAGs were previously generated from the same samples (XDC03 and XDCOriginal) using HiSeq3000 illumina data (La Rosa 2020, unpublished).

MAG	Reads aligned	Avg. alignment identity (%)	% of bp >200x coverage	% of bp 100-200x coverage	% of bp 30-99x coverage	% bp < 30x coverage
B. intestinalis	150 569	96,8	96	2,8	1,1	0,01
P. distasonis	74 804	96,3	1,7	97	0,7	0,01
E. coli	37 789	96,7	0,25	0	96	3,8

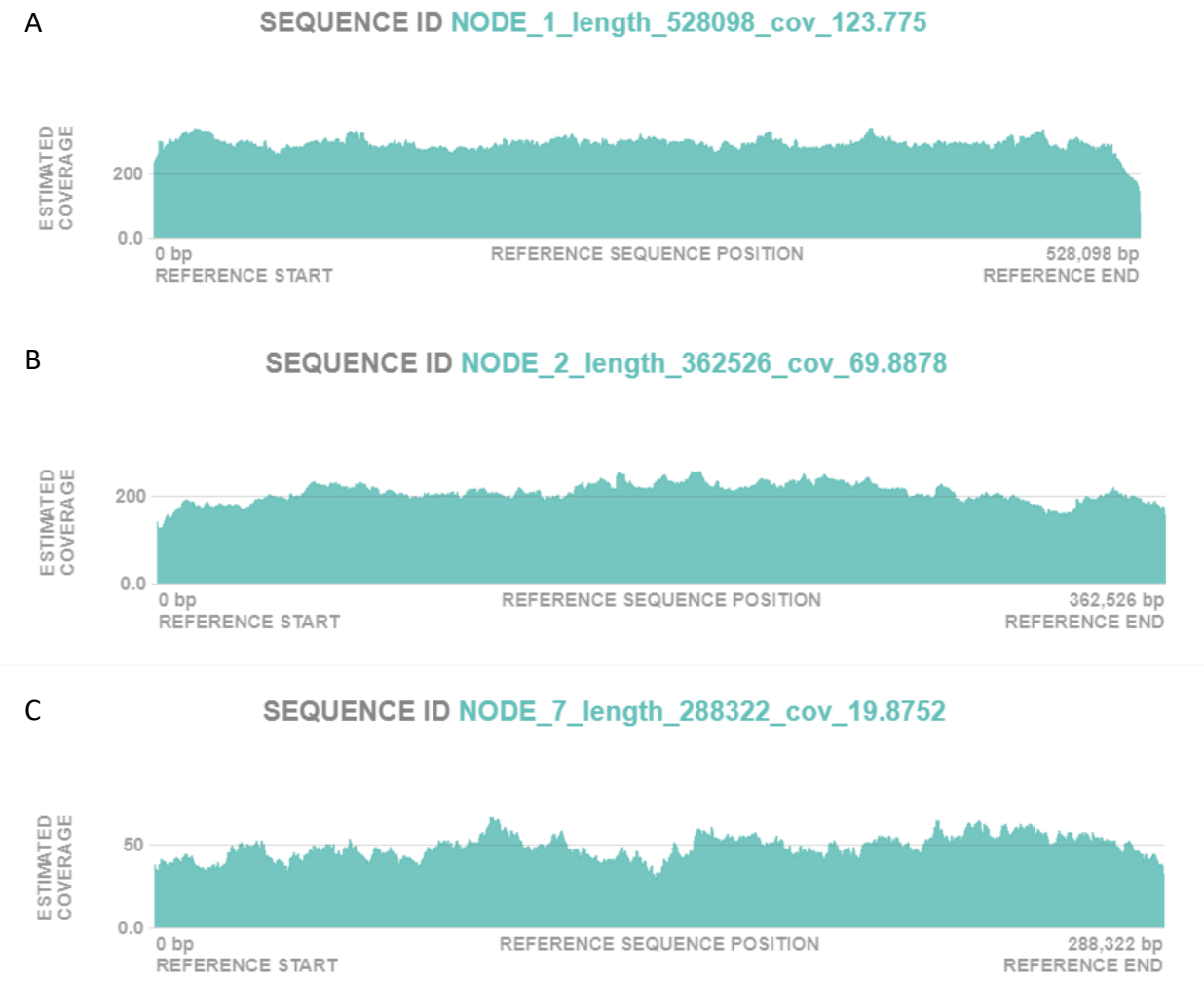


Figure 4.3.3.1: Alignment coverage of longest scaffold for A: B. intestinalis B: P. distasonis C: E. coli

Figure 4.3.2.2.1 illustrates the read coverage along the longest scaffold of each MAG. The alignments show relatively even distribution of reads across the entire reference, ranging from approximately 50x coverage for the *E. coli*-MAG, to well above 200x coverage for the *B. intestinalis*-MAG.

4.3.4 Gene calling and annotation

To further analyse the metagenomic data, gene calling and annotation was performed using web-based platforms to predict functional potential. As the long-read nanopore data was not assembled and polished due to time restrictions, the three MAGs listed in Table 4.3.3.1 were used for subsequent analysis. Gene calling with MetaGeneMark resulted in several thousand predicted genes for each MAG, with the highest number of genes found in the *B. intestinalis*-MAG. The number of CAZymes as annotated with dbCAN showed high variability, with the *P. distasonis*-MAG and *B. intestinalis*-MAG yielding more than two and five times as many predicted CAZymes as the *E. coli*-MAG respectively. Additionally, KEGG-annotation using GhostKOALA resulted in successful orthology-assignment of nearly 75% of the called genes in the *E.* coli-MAG, but only slightly above 40% of called genes from *B. intestinalis*- and *P.* distasonis-MAGs. Results from gene calling and annotation is summarized Table 4.3.4.1.

Table: 4.3.4.1: Summary of gene calling using MetaGeneMark and annotation with dbCAN and GhostKOALA respectively for each MAG

MAG	Genes called	Predicted CAZymes (loci)	KEGG-annotation
B. intestinalis-MAG	4661	482 (439)	40,6%
P. distasonis-MAG	3858	231 (214)	43,7%
E. coli-MAG	4226	94 (89)	74,2%

Annotated pathways for glycolysis and propanoate metabolism in each MAG are illustrated in figure 4.3.4.1 and 4.3.4.2

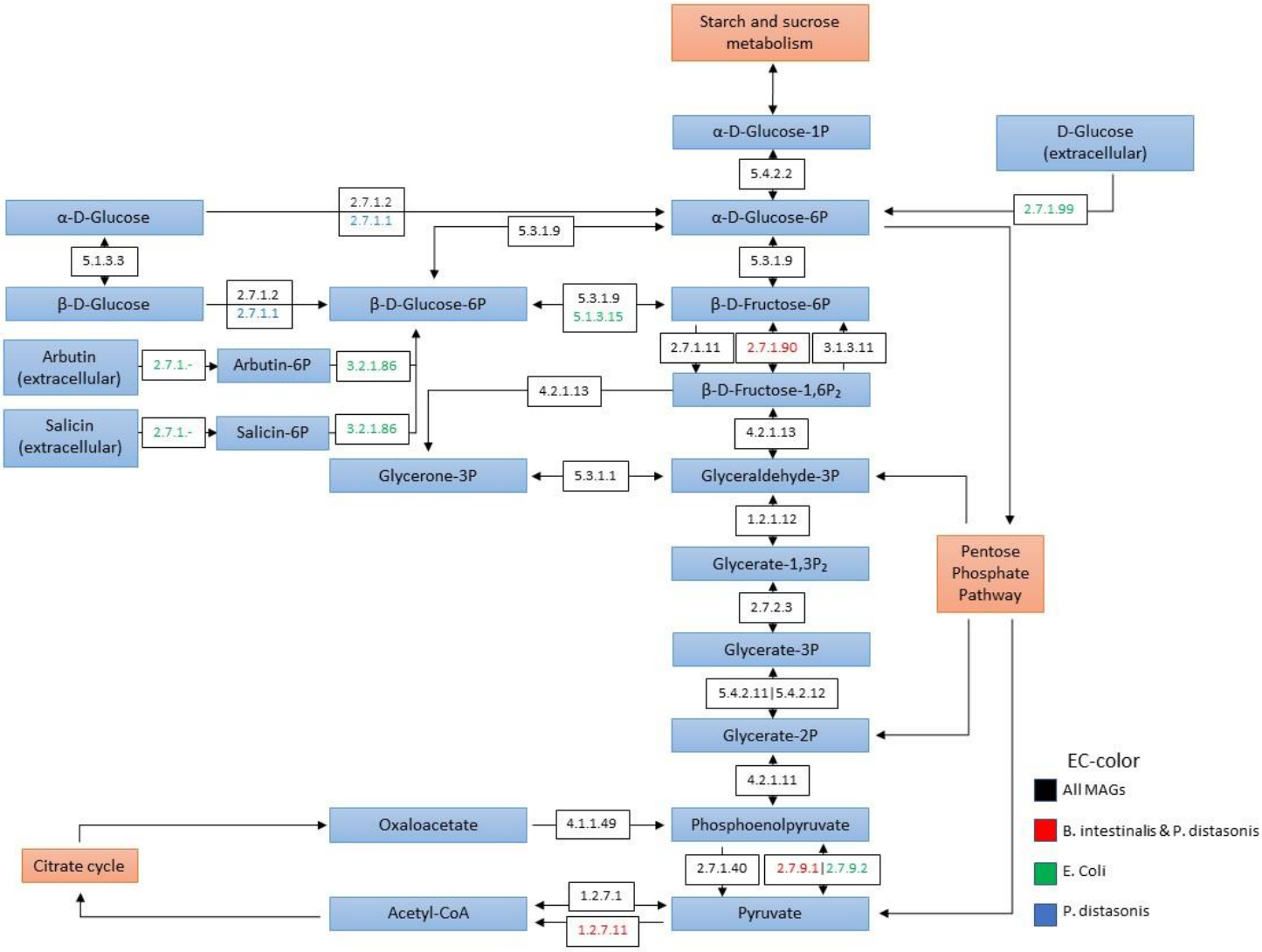


Figure 4.3.4.1: Annotated pathway for glycolysis/gluconeogenesis in all MAGs. Differences in EC-color differentiates which enzymes catalyzes a reaction in the different species.

The annotated genes were further analyzed in the larger context of metabolic pathways depicting microbial metabolism. This was achieved by generating two pathways maps based on E.C. numbers obtained through KEGG Orthology for each MAG and the KEGG Pathway database. Figure 4.3.4.1 shows a reconstruction of the pathways involved in glycolysis and gluconeogenesis in relation to

starch and sucrose metabolism, as well as the citrate cycle and the pentose phosphate pathway in all three MAGs. In *E. coli*, enzymes that allow for a number of extracellular substrates to enter glycolysis was found, whereas only enzymes that use byproducts from other pathways were identified in the other two genomes. Other enzymes related to glycolysis and gluconeogenesis were also observed, however, only those with complete pathways was included in the pathway map.

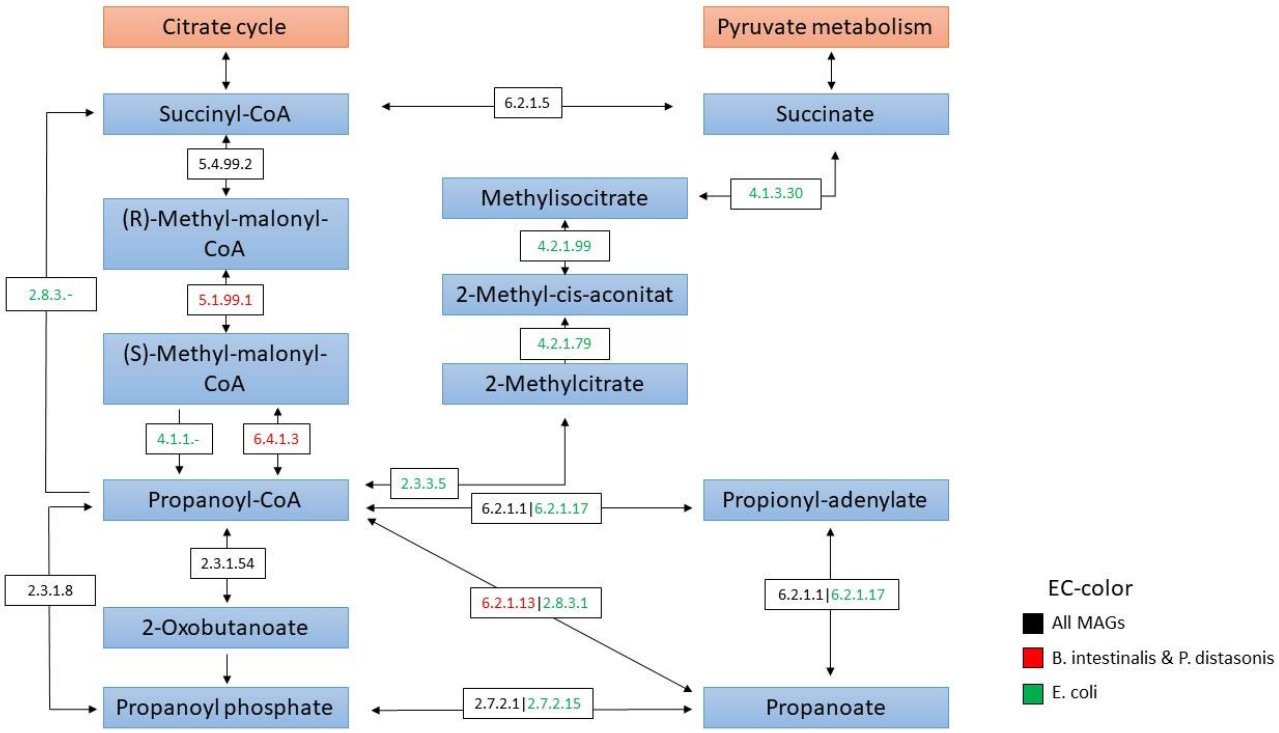


Figure 4.3.4.2: Annotated pathway for propanoate metabolism in all MAGs. Differences in EC-color differentiates which enzymes catalyzes a reaction in the different species.

Figure 4.3.4.2 shows complete reversible metabolism of propanoate in relation to the citrate cycle and pyruvate metabolism in both *B. intestinalis* and *P. distasonis.* Several possible pathways for *E. coli* to utilize propanoate as an energy source were also identified. A number of other enzymes related to propanoate metabolism were also identified, however, one or more were missing from all MAGs to link a substrate or other area of metabolism to propanoate and was thus left out of the pathway map.

5 Discussion

5.1 Metagenomic extraction

Obtaining high quality DNA from an environmental sample is an important part for all metagenomics studies. Aspects such as cost, time and quality requirements must be considered when deciding on which methods to use. Two main approaches exist for isolating microbial DNA from a sample matrix: direct and indirect extraction (Delmont et al., 2011). Direct approaches, where cells are lysed directly in the sample are known to extract higher quantities and to capture more of the total diversity present. As direct approaches often involve vigorous mechanical or physical processing however, the DNA obtained is generally of limited length due to shearing (Robe et al., 2003). In contrast, indirect methods first separate cells from the sample matrix, which allows for gentler lysis approaches and

are subsequently able to recover longer fragments of DNA, but typically produce lower yields and are known to introduce biases by not effectively dissociating all cells from the sample (Williamson et al., 2011).

Both approaches were tested and compared prior to the extraction of analytical rumen samples through two different protocols: the indirect CTAB + phenol:chloroform method, as described by Kunath et al. (2017) and the direct method using the Qiagen DNeasy PowerSoil Kit with bead beating. The results were consistent with literature, as yields for both methods were comparable (Table 4.1.1), but more sample biomass was necessary for the indirect method to obtain these yields. Gel images revealed that DNA was generally of higher molecular weight when using the indirect extraction (Figure 4.1.1.1), but these samples also showed additional bands at the bottom of the gel, indicating the presence of highly degraded polynucleotides. These bands were hypothesized be caused by RNA not being effectively partitioned into the phenolic phase during purification, resulting in additional bands of degraded RNA on the gel. Phenol-purification of nucleic acids displays sensitivity to the pH and salts present in the phenol-solution, where neutral to acidic pH causes RNA to partition into the aqueous phase (Xu et al., 2019). Even though the phenol being used was buffered to pH 8 by the manufacturer, it had also been opened for several months at the time of extraction, which may have affected the pH enough to allow some RNA to partition into the aqueous phase. Although less likely, the bands may also have been caused by severely degraded DNA, in which case the measured yield of HMW DNA could have been overestimated. If this extraction method had been chosen for the analytical samples, an RNase treatment could have been utilized to identify the type of nucleic acid, as this would remove bands caused by RNA. Regardless of origin, the additional bands would potentially require another round of purification, which in turn could negatively impact yields and leave less target DNA for the downstream analysis.

Although the longer fragment lengths generated by the indirect extraction were highly desirable, it was decided that the direct method would be used for the analytical samples. As the number of samples was fairly high, the trade-off of slightly shorter read lengths was deemed acceptable when compared to the amount of time necessary to extract sufficient DNA using the indirect method, which due to the need for precision pipetting was not as suited to parallelization. The direct method was also considered less sensitive to human errors, limiting the possibility of failed extraction or insufficient yields. However, as illustrated in Table 4.1.2.1 the kit extractions from the analytical samples did result in quite varying yields, in spite of identical treatment. The highest standard deviation was found among the particle phase samples, at approximately double that of fluid samples. It seems likely that this variation of especially the particle samples may have been caused by uneven distribution of cells, as these samples were quite compact and therefore difficult to successfully homogenize. This uneven distribution of cells may also be a potential source of extraction bias in these samples and demonstrates the challenges of extracting metagenomic DNA from a biomass sample matrix.

5.2 Marker gene analysis

5.2.1 Amplicon PCR

Although the polymerase chain reaction is routine in most laboratories, it can still pose significant challenges. Reagents such as polymerase and primer can lose their effectiveness with time and might need replacing if amplification fails. In addition, highly stringent conditions may limit yields, and low stringency may have the opposite effect and lead to non-specific products (Lorenz, 2012). As shown in Section 4.2.1, initial amplification of DNA from the rumen samples resulted in low yields which spurred several rounds of troubleshooting. Positive controls were employed to assess whether problems were related to template DNA, and although the controls were old, one of these did result

in a slightly stronger band than those from the samples (Figure 4.2.1.1C). The primers were also replaced: making new dilutions first from the original stock, and later from freshly arrived ones, however, this also resulted in little to no effect (Figure 4.2.1.1D).

A gradient assay was used to determine whether failure to amplify was caused by insufficient concentration of template in the original reaction, comparing both a doubling and a quadrupling of the template concentration to the original reaction. Results of the assay showed little to no yield across all concentrations, and thus it was speculated that the low yields were caused by the presence of an inhibitory contaminant not properly removed in the purifying steps of the extraction instead, as DNA obtained by direct metagenomic extraction methods is susceptible to contamination by for example humic acids (Robe et al., 2003).

The polymerase chain reaction is vulnerable to a number of inhibitory substances, including several compounds typically used in extraction and purification such as phenol and ethanol, as well as other substances present in the original sample matrix, like humic acids (Schrader, Schielke, Ellerbroek, & Johne, 2012). The presence of one or more of these seemed likely to be the main reason for low yields. Templates with high G+C content have also been shown to less efficiently dissociate into single strands, leading to poorer annealing of the primers and consequently reducing the amplification (Suzuki & Giovannoni, 1996). Both of these issues were addressed simultaneously in the assay described in Table 3.4.1. The samples were further diluted, thus diluting any potential contaminant, and three additives known to help stabilize PCR and denature template were tested. The sample mix brought forward contained Dimethyl sulfoxide (DMSO) which can help denature GC-rich templates, but might require changes to the annealing temperature (Bio-Rad), and an annealing temperature assay was performed to address this.

A variety of non-specific products can be produced during PCR, the formation of which can be affected by concentrations of the various reaction components, as well as the time spent on sample preparation. These so-called artifacts can be formed by hybridization and subsequent extension of the primes to each other (primer-dimers), or by two strands of genomic DNA annealing to one another. Although the use of hot-start polymerases should limit the extension activity at room temperatures, residual activity and unspecific binding of primers to genomic DNA may still lead to unintended products (Ruiz-Villalba, van Pelt-Verkuil, Gunst, Ruijter, & van den Hoff, 2017). Due to bulky structure caused by non-complementary regions of the molecules, artifacts formed by the hybridization of heterologous sequences within the template tend to migrate slower through gels in electrophoresis, appearing as additional bands above the desired product (Kanagawa, 2003). The occurrence of artifacts could potentially limit yields of intended products, perhaps particularly in the case of primer-dimers, which can form in master mixes prior to the addition of template DNA. Although these non-specific products can arise at any point, limiting the processing time, and keeping samples on ice, as well as using as few cycles as possible is important to minimize the occurrence of PCR artifacts.

The observed difficulties in amplifying DNA from the rumen samples was likely caused by a combination of factors, mainly a difficult template paired with some form of contamination carried over from the extraction. Although the A260/230-ratios measured were mostly within the desired range (Table 4.1.2.2), it is still possible that some inhibitory substance was present in high enough concentration to hinder PCR, but low enough to not substantially affect the absorbance. Through several rounds of troubleshooting, yields were increased, and the levels of non-specific product was decreased (Figure 4.2.1.3). Artifacts were however, not entirely eliminated and may still have affected downstream analysis.

5.2.2 16S amplicon sequencing analysis:

Amplicon sequencing of the 16S r RNA gene is one of the most widely used methods of determining microbial community composition in an environmental sample. Traditionally, analysis has been built on the identification of operational taxonomic units (OTUs), where sequences are grouped together based on similarity (Johnson et al., 2019). This method is useful for reducing the impact of sequencing errors but offers limited resolution and is highly dependent on the dataset in which it is employed (Tikhonov, Leach, & Wingreen, 2015). To address this, another approach instead utilizes a denoising algorithm to separate real biological sequence from false sequences introduced by amplification and sequencing and then classifies the inferred true biological sequences as amplicon sequence variants (ASVs) (Bharti & Grimm, 2019). ASV-based methods such as DADA2 is now often recommended for 16S r RNA gene analysis based on their ability to successfully assign taxonomy with greater resolution (Knight et al., 2018).

As illustrated in Figure 4.2.2.1, even though the quality of forward reads was relatively high, the reverse reads suffered from longer low quality-ends. This is thought to have caused the issues when trying to merge forward and reverse reads by not allowing for sufficient overlap after quality filtering and denoising. The V3-V4 region is a relatively long amplicon of approximately 470 base pairs, and is therefore vulnerable to low quality ends (Rausch et al., 2019). One major difference between the DADA2 pipeline and other popular tools for amplicon analysis is that merging takes place after denoising, and thus imposes strict requirements for exact overlaps between reads (Callahan et al., 2016). Although these parameters, as well as those used in quality filtering can be significantly relaxed, doing so did not increase the rate of merging beyond 40% of filtered reads. Based on this, it was determined that reverse reads were not of sufficient quality to obtain high-accuracy merged sequences and therefore the DADA2 pipeline was applied to forward reads only.

Of the approximately 32,5k amplicon sequence variants inferred by DADA2, only 34% were successfully assigned a taxon at the genus level. This is likely due at least in part to the short sequences derived from the use of only forward reads in the analysis, as shorter sequences have been shown to impact both the sensitivity and specificity of taxonomic classification. When comparing assigned taxonomy of different regions and the full length 16S rRNA gene using virtual PCR for amplification, Martínez-Porchas, Villalpando-Canchola, and Vargas-Albores (2016) found that shorter regions resulted in shallower taxonomic assignment compared to those attained by unilarge fragments. Additionally, some of the shorter amplicons were subject to misclassifications where amplicons were wrongfully identified as different organisms when using short internal region sequences as compared to their corresponding full-length counterparts. This shows that as the length of sequence variants decreases, it becomes increasingly difficult to correctly assign taxonomy at deeper levels, which may have substantially affected the proportion of reads assigned by the DADA2 pipeline as well as their accuracy.

Less than optimal PCR conditions and a high number of total cycles could also have resulted in a significant portion of PCR artefacts, thus limiting accurate taxonomic assignment by causing small changes to sequences found in the original templates. These artefacts can appear due to a number of reasons such as polymerase error and the amplification of non-specific product as discussed in Section 5.2.1, and become increasingly more abundant as the number of cycles increases (Acinas et al., 2005). Since the DADA2 algorithm depends on abundances to distinguish between biological sequence and artefact (Callahan et al., 2016), it is possible that in the event of artefacts arising early in the amplification process, these may have reached sufficient abundances to be mistaken for actual biological variance.

To examine whether there was any immediately observable impact on sample diversity caused by the inclusion of sugar kelp in the sheep's diet, ordination was plotted using non-metric multidimensional scaling with Bray-Curtis distances (Figure 4.2.2.2). The plot reveals clear clustering of fluid and particle samples, illustrating that these portions of the rumen are home to a distinct variety of the total community. The lack of apparent patterns between samples within fluid and particle-phases suggests that there has been no immediately observable effect in community structure caused by dietary change. As the purpose of this project was to examine the potential health effects of sugar kelp as a substitute for parts of the usual roughage fed to the sheep, the lack of difference in community structure could be considered a good thing, as the diet seems to have had minimal impact. Although community structure may not have changed substantially, there is still the possibility that the proteomic profile of the rumen may have changed due to differential expression of genes.

The distribution of the most common ASVs (Figure 4.2.2.3) further supports the hypothesis that the impact of diet on community structure has been negligible, as the relative abundances of families seems to remain comparable across all samples. The fact that the 50 most common ASVs only accounts for 10-20% of all sequences detected indicates that the rumen is a highly complex environment, with a wide variety of species. However, as the taxonomic profiles assigned in this analysis are subject to several weaknesses as discussed above, results may be significantly skewed, and any observed trend should probably be considered tentative at best. As only minor differences in sample composition were observed between the different diets, the 24 samples sent to the Sequencing Centre in Oslo were selected from a total of 12 sheep, with both fluid- and particle phase samples from each individual, with each diet equally represented by four animals. The particular animals were chosen based on assessment of relative diversity of the most abundant ASVs within each sample. More diverse community compositions were deemed more preferable than those highly dominated by only one or two genera, as the diversity would hopefully allow for the successful assembly of a higher number of genomes in downstream analysis by providing more even sequencing depth for each species.

5.3 Whole metagenome/shotgun sequencing

To investigate potential microbial function within a microbial community, access to genes and genomes is required. The original intent in this project was to perform shotgun metagenomic sequencing on the aforementioned rumen samples using both short read (Illumina) and long read (Oxford Nanopore) technology. However, due to COVID-19, long waiting times at the sequencing centre in Oslo and the uncertainty of how the nanopore technology would respond to the potentially troublesome rumen sample DNA, it was instead decided that left-over DNA extracted two enrichments (XDC03 and XDCOriginal) derived from human stool (Ostrowski et al., 2020 (in preparation)) would be used for nanopore sequencing. These samples had the added benefit of longer DNA fragment length and had been previously profiled with 16S rRNA gene analysis, shotgun Illumina assemblies and MAG reconstruction, thus providing the basis for alignment analysis.

5.3.1 Whole metagenome taxonomic profiling

There are several advantages to taxonomic profiling by shotgun sequencing. Perhaps the most obvious of these is the ability to identify organisms and their relative abundance regardless of which domain of life they belong to. Another benefit of shotgun methods is the removal of bias potentially introduced by the repeated amplification of template DNA in amplicon sequencing (Rausch et al., 2019). Taxonomic classification using 16S rRNA sequencing typically also doesn't classify most OTUs beyond the genus level, whereas shotgun sequencing has been shown to successfully resolve species and strain-level profiles more accurately (Hillmann et al., 2018). This higher resolution can be

illustrated by the results of the taxonomic profiling of the XDC03 and XDCOriginal samples using Oxford Nanopore (MinION) sequencing runs. The distribution of reads generated from the MinION flow cells (referred to “Esther” for XDC03 and “Rosalind” for XDCOriginal) showed high abundances of reads classified as Bacteroidales and Clostridales (Figure 4.3.2.2 and 4.3.2.3), as is consistent with initial 16S rRNA gene classification (Figure 4.3.2.1). At lower abundances however, a previously undetected diversity of both archaeal and fungal reads were identified (Figure 4.3.2.4 and 4.3.2.5), as well as viruses and a host of less abundant bacterial strains, many of which have been assigned at the species-level. Although most of these species are present at such low levels that functional profiling still remains a significant challenge, from an ecological viewpoint, it is still desirable to gain an as complete view of the total complexity of the community as possible.

One major drawback of using whole metagenome sequencing reads for taxonomic profiling is the availability of suitable references, as it is only effective in determining the presence of already known species (Milanese et al., 2019). In addition, the relatively high error rates of nanopore sequencing, commonly between 5% and 15% (Rang, Kloosterman, & de Ridder, 2018) poses a potential challenge for accurately assigning taxonomy based on homology-searches. A total of 30-40% of all reads generated by the Esther and Rosalind flow cells were not successfully assigned to any taxonomic level, which could possibly be explained by the factors mentioned above.

Firstly, as the average base call accuracy of reads was only approximately 92%, with quality cutoff for the analysis well below this, at 80% (Q-score 7) some sequences may significantly differ from those found in the NCBI database used in the analysis, and therefore disrupt the alignment of reads to references. As some species may also be separated by only a few distinctions in a given read, this may also potentially lead to mis-assignment between closely related species. The potential for sequencing errors to cause large enough differences in sequences to wrongfully or not identify homologs with sufficient identities could be considered a considerable bottleneck for meaningful taxonomic profiling using whole metagenome nanopore reads, although the long reads should be able to limit this to a certain degree, as longer sequences generally increases resolution (Bleidorn, 2016).

Even though the human gut it a relatively well characterized microbiome (Quince et al., 2017), the lack of relevant sequences in databases may still be a contributing factor, as robust refence genomes for many of the strains present in the sample may not be readily available. This hypothesis is further supported by the ability of many more reads to align with high identity to metagenomes assembled from the same sample than those taxonomically assigned to the corresponding species using reference genomes. *Parabacteroides distasonis* was assigned 29,6k Nanopore reads in the taxonomic profile, yet 74,8k reads successfully aligned to the corresponding Illumina-generated MAG (Table 4.3.3.1), signifying that only 39,5% of reads used in alignment were identified by the classifier. *Escherichia coli*, a model organism shows significantly lower discrepancies in this area, with approximately 30,7k reads assigned by the taxonomic classifier and 37,8k reads successfully aligned to the corresponding Illumina-generated MAG. This substantial gap in percentage of identified reads between the two species underlines the importance of having a comprehensive database to successfully use whole metagenome shotgun sequencing for taxonomic profiling.

5.3.2 Alignment and assembly

Alignment of raw nanopore reads to already available Illumina-generated MAGs (from the same samples) was performed to assess the relative success of the sequencing runs and to consider possible implications for future assembly. Of those tested, alignments to three MAGs were determined particularly successful, with high sequence identity (>96%) and mostly evenly distributed coverage across the entire reference above 30x.

The alignments of *P. distasonis*- and *E. coli*-affiliated MAGs to the Nanopore data represented a total of 24,6% and 12,5% of all reads generated by the Esther (XDC03) sequencing run respectively, whereas the *B. intestinalis*-affiliated MAG alignment alone represented 49,9% of all reads generated by the Rosalind (XDCOriginal) run. This read recruitment is reflected by the coverages achieved in alignments, with the *B. intestinalis*-affiliated MAG achieving the greatest coverage (>200x) and the *E. coli*-affiliated MAG with the lowest (≈ 50x). Based on the total of aligned reads and comparisons to taxonomic profiling using both 16S rRNA gene and whole metagenome analysis, it seems likely that *B.* intestinalis-affiliated MAG uses reads previously identified as *Bacteroides cellulosilyticus*. As these species are closely related, further investigation might be necessary to confidently assign taxonomy.

As time was a limiting factor, assembly of the reads generated by Nanopore sequencing was not attempted. The incorporation of long reads in assembly, however, holds great potential for improving the overall quality of metagenomically-assembled genomes by resolving repeats and reducing the number of contigs/scaffolds in the draft genomes. The higher error-rates currently observed by third generation technology remains the greatest obstacle for robust assemblies, and a number of different approaches to correct these errors exists. Algorithms for error correction generally falls into one of two classes: self-correction, and hybrid-correction (Amarasinghe et al., 2020).

Self-correction can be accomplished by finding consensus between overlapping reads, but is limited by the overall coverage, and is therefore not as useful when coverage is low such as for reads from less abundant species in metagenomic samples. In contrast, hybrid-correction uses more accurate next-generation short reads to identify and correct errors within long read sequences, thus conserving more information from long reads of low coverage (Fu, Wang, & Au, 2019).

Hybrid approaches utilizing complementary short-read data are useful in gaining higher resolution and are commonly incorporated in one of three ways. 1) Error correction of raw long reads prior to assembly by alignment of short sequences to generate a high-accuracy consensus sequence (Koren et al., 2012). 2) Post-assembly “polishing”, using short reads to error-correct contigs (Schmidt et al., 2017). 3) Assembly using input from both long and short reads, simultaneously taking advantage of both the high accuracy of short reads, and the ability of long reads to resolve repeats and close gaps (Wick, Judd, Gorrie, & Holt, 2017).

A comparative analysis of available de novo metagenomic assemblers using only self-correction found that often used tools for short read data sets did not perform well on data generated by nanopore sequencing, resulting in highly fragmented draft genomes. Of software specifically designed for long-read assembly, two were noted to perform particularly well: Canu and Flye, which demonstrated assembly accuracies of up to 99.87% and 99.67%, respectively. Although Canu generated slightly more accurate assemblies and more accurately resolved insertion/deletion errors, Flye managed to assemble complete bacterial draft genomes with only 2-21 contigs, and was several times faster (Latorre-Pérez, Villalba-Bermell, Pascual, Porcar, & Vilanova, 2019).

Based on results of alignment and the comparisons of relative read counts as determined by taxonomic profiling, it seems likely that assembly of the nanopore datasets generated for this project would result in several robust draft genomes using only the raw nanopore reads. However as illustrated by alignment, a significant portion of reads aligned to only a few highly abundant species, with nearly half of all reads from Rosalind aligning to the *B. intestinalis*-affiliated MAG . Although Flye has been reported to assemble genomes with coverage as low as approximately 10x (Wick & Holt, 2019), it is likely that hybrid approaches would be preferable to aid in successfully recovering more draft genomes for the less abundant strains present in samples.

5.3.3 Gene calling and annotation

In the case of functional annotation, it is important to have as completely assembled genomes as possible (Fraser, Eisen, Nelson, Paulsen, & Salzberg, 2002). Gaps between contigs could possibly cut genes in half and thus not allow them to be identified by gene callers, and if they are, annotation remains less likely to characterize the gene, particularly when the gene in question does not easily allow for assignment through homology. Depending on intended application, choice of gene callers can vary, however when it comes to metagenomes, *ab initio* algorithms hold several advantages. When mining for novel genes in often fragmented metagenome-assembled genomes, *ab initio* gene callers such as MetaGeneMark are integral to finding gene sequences of previously uncharacterized proteins by identifying open reading frames (ORFs) based on patterns within the genome. Metagenomes in general may contain protein-coding regions that are difficult to detect using sequence similarity due to evolutionary distance resulting in low homology when compared to reference databases (Besemer & Borodovsky, 2005; Zhu, Lomsadze, & Borodovsky, 2010).

Annotation can be used to determine whether particular enzymes or metabolic pathways of interest is detected within a sequenced genome. Algorithms for assigning a function to genes tends to rely on homology between genome and references in databases and is therefore somewhat limited by the availability of suitable reference sequences (Richardson & Watson, 2013). Carbohydrate-active enzymes hold high economical interest due to the potential of finding particular strains or enzymes that could aid in the production of efficient biofuels from lignocellulosic biomass, as well as providing further insights into the global carbon cycle (Baldrian & López-Mondéjar, 2014). Tools that incorporate domain-specific identification of CAZymes such as dbCAN allows for fast annotation of possible loci and categorizes each gene based on the classification scheme of the CAZy database (Yin et al., 2012). Both the *B. Intestinalis* and *P distasonis*-affiliated MAGs analyzed in this study revealed a relatively large number of potential CAZymes using dbCAN, with 482 and 231 total domains identified respectively, whereas a significantly lower 94 CAZyme domains were identified in the *E. coli*-affiliated MAG, indicating a potentially much more comprehensive machinery for degrading more complex carbohydrates in the first two.

GhostKOALA, an automated annotation pipeline for metagenomes using KEGG Orthology has the benefit of directly linking annotated genes to high-level function (Kanehisa, Sato, & Morishima, 2016), and thus eases the process of reconstructing known metabolic pathways within a genome. A significant disparity in the percentage of annotated genes using GhostKOALA can be observed (Figure 4.3.4.1), with 74,2% of the genes identified in the *E. coli*-affiliated MAG assigned a function through homology. In contrast, for the *B. intestinalis* and *P. distasonis*-affiliated MAGs, only slightly above 40% of genes called were successfully annotated. As *E. coli* is among the most studied organisms in the world (Blount, 2015), it seems likely that this difference arises at least in part from the availability of relevant reference sequences in the database. This further highlights the need for continued efforts in expanding databases to help gain a deeper understanding of the metabolic processes in microbial communities. The reconstruction of glycolysis and gluconeogenesis in each MAG as illustrated in Figure 4.3.4.1 shows that a set of core genes is shared by all MAGs, as most enzymes capable of catalyzing a reaction have been identified in all three metagenomes. Notably, many of the enzymes identified only in the *E. coli*-affiliated MAG were involved in the delivery of extracellular substrates (D-glucose, arbutin and salicin) to glycolysis/gluconeogenesis. Of the enzymes found in only the *B. intestinalis*- and/or the *P. distasonis*-affiliated MAGs most seemingly served as additional alternatives for reactions already catalyzed by other enzymes in the shared set of core genes. The reconstructed propanoate metabolism map (Figure 4.3.4.2) shows a completely reversible metabolic pathway linked to both the citrate cycle and pyruvate metabolism for *B. intestinalis* and *P. distasonis*-affiliated MAGS. In the *E. coli*-affiliated MAG, however, several one-way pathways linking propanoate

to either the citrate cycle or pyruvate metabolism may suggest that propanoate only serves as an energy/carbon source in this MAG. The identification of specific enzymes and the reconstruction of metabolic pathways can be used to predict interactions between species in a sample. For example, comparisons between MAGs indicates the potential for both *B. intestinalis* and *P. distasonis*-affiliated populations to act as primary degraders, with a host of CAZymes to aid digestion of complex carbohydrates. In contrast, a low number of observed CAZymes, along with the ability to use several extracellular substrates and potential metabolites from primary degraders such as propanoate in energy metabolism suggests that *E. coli*-affiliated population primarily utilizes less complex carbohydrates and benefits from cross-feeding interactions, as is consistent with literature (Conway & Cohen, 2015; Flint et al., 2008).

5.4 Nanopore sequencing: challenges and potential

Sequencing technologies have come a long way since Frederick Sanger and colleagues first applied their new method of chain-terminating inhibitors to sequence the bacteriophage φX174 (Frederick Sanger et al., 1977). The development of next generation high-throughput platforms led to a revolution in genomics, generating massive amounts of data; however, these technologies produced limited read lengths, causing difficulties with downstream bioinformatic processing (Koren & Phillippy, 2015). A new generation of sequencing technology is now on the rise, promising ultra-long reads and real time output (Lu, Giordano, & Ning, 2016). The concept of nanopore sequencing was proposed as early as the 1980s, and after three decades of development, it culminated in the release of the Oxford Nanopore MinION in 2014, the very first commercially available nanopore sequencer (D. Deamer et al., 2016). Within the first few years, the MinION was among other things reported to have successfully sequenced viral (Jing Wang, Moore, Deng, Eccles, & Hall, 2015), bacterial (Loman, Quick, & Simpson, 2015), and even eukaryotic (Giordano et al., 2017) genomes, and was also used for field-laboratory diagnostics during an Ebola virus outbreak (Hoenen et al., 2016), which sparked significant interest and illustrated the inherent potential of nanopore technology.

The biggest caveat for nanopore sequencing has been the error rate, and in early stages, these were sometimes reported as high as 40%. Through an impressive amount of development in both chemistry and base-calling however, this rate is now routinely reported in the much lower range of 5% to 15% (Rang et al., 2018). A top contributor to these errors is the inability to achieve single base resolution with the pore used for sequencing. As the current detected at any point is a function of the combined impact of a k-mer interacting with the recognition sites of the pore, these signals must be accurately translated into a sequence of nucleotides (Y. Wang et al., 2015). Although current technology uses biological nanopores, Oxford Nanopore is actively working to develop a future generation of solid-state pores, which is intended to help with the both mechanical and chemical stability (ONT, 2019b). Particularly graphene is seen as a promising prospect for achieving maximum resolution, as it is considered to be the thinnest membrane capable of separating to liquid compartments, and has demonstrated sensitivity to ionic conductance during DNA translocation (Garaj et al., 2010).

In spite of the inherent limitations posed by the current error rates, several studies have indicated the ability to obtain higher resolution taxonomic assignment in marker gene sequencing when utilizing long-read technology. As mentioned in Section 5.2.2, the length of the amplicon significantly influences both sensitivity and specificity of assigned taxa in 16S rRNA gene analysis. Nanopore sequencing, as opposed to short-read platforms such as the Illumina MiSeq system, is not limited to choosing only small regions of the desired marker gene, and routinely generates reads well above the approximately 1500 base pairs required to cover the entirety of the 16S rRNA gene (Johnson et al., 2019). Nanopore sequencing may also help in identifying PCR artefacts in a dataset, as J. Wang et

al. (2015) noted that when amplifying parts of the influenza genome, some sequences generated by the MinION were significantly longer than the expected fragments generated by amplicons. They hypothesized that these sequences may have been derived from chimeric amplicons that would have otherwise gone undetected if sequencing was accomplished using short read technology. Although this was only speculation, it remains possible that the presence of PCR artifacts of unexpected lengths could be more easily detected and subsequently filtered out in amplicon analysis when using nanopore sequencing.

A comparative analysis of the mouse gut microbiome as characterized by short-read Illumina sequencing of the V3-V4 regions and near full gene sequencing on the MinION found that the relative community structure did not significantly differ when using the different platforms, and that the longer nanopore sequences did in fact more accurately assign taxonomy, providing higher resolution that allowed for the separation of more species from their respective genera. However, only a small proportion of the total reads generated by MinION sequencing passed all quality filters and mapped successfully to a reference (J. Shin et al., 2016). This demonstrates that the MinION is capable of accurately resolving community structure in marker gene analysis but requires strict quality filtering, which could somewhat skew results and lead to the potential loss of sequence variants of interest. It therefore seems likely that the higher-accuracy Illumina platforms will continue to dominate the field of marker gene sequencing as long as the error rate of nanopore sequencing remains high, particularly for projects involving a large amount of samples where statistical comparisons between certain groups in the data set is desirable.

The perhaps greatest potential of long-read technologies is their ability to improve genomic assemblies by closing gaps and resolving repeats by having single reads span significantly larger parts of the genome. Short read sequencing notoriously struggles to assemble into finished genomes, instead producing fragmented draft assemblies which can affect downstream analysis (Koren & Phillippy, 2015). High-quality assemblies with as few gaps as possible are of great importance for applications such as annotation and comparative genomics, but traditional short read assemblers were quickly deemed unsuitable for efficient incorporation of long erroneous reads, and thus development of new tools was necessary to unlock the full potential of long reads (Koren et al., 2012). Although the error rate still poses a challenge, several assemblers now exist capable of producing high quality draft assemblies using only long reads (Latorre-Pérez et al., 2019; Wick & Holt, 2019), and hybrid approaches can be applied to error-correct the long reads (Fu et al., 2019). S. C. Shin et al. (2019) demonstrated that the inclusion of nanopore reads vastly improved a previous Illumina-only assembly of the winged midge *Parochlus steinenii*, reducing the number of contigs from 9132 to 162 and increasing completeness from 87,8 to 98,7%, which in turn helped with further annotation of the genome. Similarly, a hybrid nanopore approach has been reported to have doubled the number of bacterial and archaeal metagenome-assembled genomes from a complex aquifer system compared to assembly using only short reads (Overholt et al., 2019). This clearly demonstrates the potential of nanopore sequencing to aid in the assembly and annotation of thus far inaccessible genomes from other complex environments such as soil and herbivore rumen, which in turn could lead to the discovery of novel genes of high economical interest like those encoding different CAZymes and novel pharmaceuticals (Baldrian & López-Mondéjar, 2014; Belknap, Park, Barth, & Andam, 2020).

Although the MinION holds great promise for smaller scale projects, it does however offer limited throughput. The Esther and Rosalind flow cells generated 5.56 and 6.19 gigabytes of data over 6 and 7 hours of sequencing respectively, resulting in approximately 300k pass reads in each run, of which Rosalind averaged slightly higher in terms of read lengths. Even though sequencing times could be

increased significantly, obtaining the extreme depth necessary to assemble more than the most abundant species in complex communities is most likely well beyond the scope of the small MinION. To address the need for higher-throughput, ONT introduced the PromethION, a modular benchtop system containing up to 48 flow cells, each with the sequencing capacity of approximately six MinIONs. The PromethION promises to produce up to 8 Tb of sequencing data from a single run (ONT, 2020) and might even be able to generate sufficient coverage for accurate nanopore-only assembly of several lower abundance species in complex environments, especially if current development continues to lower error rates as well as increase read lengths.

Extraction protocols may also need to be optimized to fully utilize the long-read potential of nanopore platforms. Figure 4.3.1.1 shows that although most reads generated by Esther are on the shorter end of the spectrum, a significant number of reads were well above the approximately 10 kb average, with several reads surpassing 100kb. To ensure the success of a nanopore sequencing run, it seems prudent to use extraction methods that yield fragments of as high molecular weight as possible. Based on the obtained average in this analysis, aiming for a minimum fragment length of 10 kb is suggested as a tentative guideline for extracting DNA suitable for nanopore application. As such, common mechanical lysis methods like bead beating that typically result in shearing (Robe et al., 2003) are less applicable in this type of sequencing, as the fragment length is likely going to a limit the potential benefits of using long-read technology.

Nanopore sequencing holds massive potential for a host of different applications. Cost and ease make particularly the MinION accessible for smaller laboratories and is perhaps an especially promising tool for real time diagnostic purposes such as determining the presence of pathogenic strains in a sample (Quick et al., 2015). The technology has already been shown to significantly increase the quality of fragmented assemblies, both in complex eukaryotic genomes and microbial metagenomes. For the sequencing of complex genomes or metagenomes, the higher throughput offered by the PromethION could drastically increase the availability of robust reference genomes and provide a deeper understanding of both structural variants and improvement in functional profiling. The biggest obstacle for nanopore sequencing remains that of the error rate, however, with further development of chemistry and base-calling software, this is expected to continue to decrease. Additionally, as illustrated by several studies as mentioned above, the error rate can in many cases be negated by sufficient coverage, or by the implementation of hybrid approaches. It seems unlikely that short read platforms such those by Illumina will fall out of favor in the near future, particularly for analyses where high accuracy is key; however, in the areas where these approaches fall short, nanopore sequencing is likely to play an important role in the coming years.

References

Acinas, S. G., Sarma-Rupavtarm, R., Klepac-Ceraj, V., & Polz, M. F. (2005). PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Appl. Environ. Microbiol., 71*(12), 8966-8969.

Akeson, M., Branton, D., Kasianowicz, J. J., Brandin, E., & Deamer, D. W. (1999). Microsecond time-scale discrimination among polycytidylic acid, polyadenylic acid, and polyuridylic acid as homopolymers or as segments within single RNA molecules. *Biophysical journal, 77*(6), 3227-3233.

Alneberg, J., Bjarnason, B. S., De Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., . . . Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nature methods, 11*(11), 1144-1146.

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome biology, 21*(1), 1-16.

Ardui, S., Ameur, A., Vermeesch, J. R., & Hestand, M. S. (2018). Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic acids research, 46*(5), 2159-2168.

Baldrian, P., & López-Mondéjar, R. (2014). Microbial genomics, transcriptomics and proteomics: new discoveries in decomposition research using complementary methods. *Applied microbiology and biotechnology, 98*(4), 1531-1537.

Belknap, K. C., Park, C. J., Barth, B. M., & Andam, C. P. (2020). Genome mining of biosynthetic and chemotherapeutic gene clusters in Streptomyces bacteria. *Scientific reports, 10*.

Besemer, J., & Borodovsky, M. (2005). GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic acids research, 33*(suppl_2), W451-W454.

Bharti, R., & Grimm, D. G. (2019). Current challenges and best-practice protocols for microbiome analysis. *Briefings in bioinformatics*.

Bio-Rad. iProof™ High-Fidelity PCR Master Mix In.

Bleidorn, C. (2016). Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. *Systematics and biodiversity, 14*(1), 1-8.

Blount, Z. D. (2015). The natural history of model organisms: The unexhausted potential of E. coli. *Elife, 4*, e05826.

Boesenberg-Smith, K. A., Pessarakli, M. M., & Wolk, D. M. (2012). Assessment of DNA yield and purity: an overlooked detail of PCR troubleshooting. *Clinical Microbiology Newsletter, 34*(1), 1-6.

Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T., . . . Huang, X. (2010). The potential and challenges of nanopore sequencing. In *Nanoscience and technology: A collection of reviews from Nature Journals* (pp. 261-268): World Scientific.

Breitwieser, F. P., Lu, J., & Salzberg, S. L. (2019). A review of methods and databases for metagenomic classification and assembly. *Briefings in bioinformatics, 20*(4), 1125-1136.

Bukin, Y. S., Galachyants, Y. P., Morozov, I., Bukin, S., Zakharenko, A., & Zemskaya, T. (2019). The effect of 16S rRNA region choice on bacterial community metabarcoding results. *Scientific data, 6*, 190007.

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME journal, 11*(12), 2639.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nature methods, 13*(7), 581.

Charuvaka, A., & Rangwala, H. (2011). *Evaluation of short read metagenomic assembly.* Paper presented at the BMC genomics.

Conway, T., & Cohen, P. S. (2015). Commensal and pathogenic Escherichia coli metabolism in the gut. *Metabolism and bacterial pathogenesis*, 343-362.

Courtois, S., Frostegård, Å., Göransson, P., Depret, G., Jeannin, P., & Simonet, P. (2001). Quantification of bacterial subgroups in soil: comparison of DNA extracted directly from soil or from cells previously released by density gradient centrifugation. *Environmental Microbiology, 3*(7), 431-439.

Deamer, D., Akeson, M., & Branton, D. (2016). Three decades of nanopore sequencing. *Nature biotechnology, 34*(5), 518.

Deamer, D. W., & Akeson, M. (2000). Nanopores and nucleic acids: prospects for ultrarapid sequencing. *Trends in biotechnology, 18*(4), 147-151.

Delmont, T. O., Robe, P., Clark, I., Simonet, P., & Vogel, T. M. (2011). Metagenomic comparison of direct and indirect soil DNA extraction approaches. *Journal of microbiological methods, 86*(3), 397-400.

Drabik, A., Bodzoń-Kułakowska, A., & Silberring, J. (2016). 7 - Gel Electrophoresis. In *Proteomic Profiling and Analytical Chemistry*

(2nd ed.): Elsevier.
Du, R., Guo, L., Li, S., Xie, D., & Yan, J. (2018). Metagenomic DNA Extraction of Natural Cellulose-Degrading Consortia. *BioEnergy Research, 11*(1), 115-122.
Erlich, H. A., Gelfand, D., & Sninsky, J. J. (1991). Recent advances in the polymerase chain reaction. *Science, 252*(5013), 1643-1651.
Flicek, P., & Birney, E. (2009). Sense from sequence reads: methods for alignment and assembly. *Nature methods, 6*(11), S6-S12.
Flint, H. J., Bayer, E. A., Rincon, M. T., Lamed, R., & White, B. A. (2008). Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis. *Nature Reviews Microbiology, 6*(2), 121-131.
Fraser, C. M., Eisen, J. A., Nelson, K. E., Paulsen, I. T., & Salzberg, S. L. (2002). The value of complete microbial genome sequencing (you get what you pay for). *Journal of bacteriology, 184*(23), 6403-6405.
Fu, S., Wang, A., & Au, K. F. (2019). A comparative evaluation of hybrid error correction methods for error-prone long reads. *Genome biology, 20*(1), 26.
Garaj, S., Hubbard, W., Reina, A., Kong, J., Branton, D., & Golovchenko, J. (2010). Graphene as a subnanometre trans-electrode membrane. *Nature, 467*(7312), 190-193.
Gerasimidis, K., Bertz, M., Quince, C., Brunner, K., Bruce, A., Combet, E., . . . Ijaz, U. Z. (2016). The effect of DNA extraction methodology on gut microbiota research applications. *BMC research notes, 9*(1), 365.
Gibson, L. J. (2012). The hierarchical structure and mechanics of plant materials. *Journal of the royal society interface, 9*(76), 2749-2766.
Giordano, F., Aigrain, L., Quail, M. A., Coupland, P., Bonfield, J. K., Davies, R. M., . . . Li, J. (2017). De novo yeast genome assemblies from MinION, PacBio and MiSeq platforms. *Scientific reports, 7*(1), 1-10.
Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature reviews genetics, 17*(6), 333.
Green, M. R., & Sambrook, J. (2017). Isolation of high-molecular-weight DNA using organic solvents. *Cold Spring Harbor Protocols, 2017*(4), pdb. prot093450.
Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and molecular biology reviews, 68*(4), 669-685.
Henderson, G., Cox, F., Kittelmann, S., Miri, V. H., Zethof, M., Noel, S. J., . . . Janssen, P. H. (2013). Effect of DNA extraction methods and sampling techniques on the apparent structure of cow and sheep rumen microbial communities. *PloS one, 8*(9).
Hert, D. G., Fredlake, C. P., & Barron, A. E. (2008). Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods. *Electrophoresis, 29*(23), 4618-4626.
Hillmann, B., Al-Ghalith, G. A., Shields-Cutler, R. R., Zhu, Q., Gohl, D. M., Beckman, K. B., . . . Knights, D. (2018). Evaluating the information content of shallow shotgun metagenomics. *MSystems, 3*(6).
Hoenen, T., Groseth, A., Rosenke, K., Fischer, R. J., Hoenen, A., Judson, S. D., . . . Squires, R. B. (2016). Nanopore sequencing as a rapidly deployable Ebola outbreak tool. *Emerging infectious diseases, 22*(2), 331.
Hugenholtz, P., & Tyson, G. W. (2008). Metagenomics. *Nature, 455*(7212), 481-483.
Huws, S. A., Creevey, C. J., Oyama, L. B., Mizrahi, I., Denman, S. E., Popova, M., . . . Morgavi, D. P. (2018). Addressing Global Ruminant Agricultural Challenges Through Understanding the Rumen Microbiome: Past, Present, and Future. *Frontiers in Microbiology, 9*(2161). doi:10.3389/fmicb.2018.02161
Illumina. Sequencing by Synthesis (SBS) Technology. Retrieved from https://emea.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html

Invitrogen. (2018). Comparison of fluorescence-based quantitation with UV absorbance measurements. Retrieved from https://assets.thermofisher.com/TFS-Assets/LSG/Technical-Notes/fluorescence-UV-quantitation-comparison-tech-note.pdf

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome biology, 17*(1), 239.

Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., . . . Gerstein, M. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature communications, 10*(1), 1-11.

Jørgensen, H., Kristensen, J. B., & Felby, C. (2007). Enzymatic conversion of lignocellulose into fermentable sugars: challenges and opportunities. *Biofuels, Bioproducts and Biorefining, 1*(2), 119-134.

Kanagawa, T. (2003). Bias and artifacts in multitemplate polymerase chain reactions (PCR). *Journal of bioscience and bioengineering, 96*(4), 317-323.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic acids research, 44*(D1), D457-D462.

Kanehisa, M., Sato, Y., & Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *Journal of molecular biology, 428*(4), 726-731.

Kasianowicz, J. J., Brandin, E., Branton, D., & Deamer, D. W. (1996). Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences, 93*(24), 13770-13773.

Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., . . . McDonald, D. (2018). Best practices for analysing microbiomes. *Nature Reviews Microbiology, 16*(7), 410-422.

Koetsier, G., & Cantor, E. (2019). A Practical Guide to Analyzing Nucleic Acid Concentration and Purity with Microvolume Spectrophotometers. New England BioLabs. *Inc.: Ipswich, MA, USA*.

Koren, S., & Phillippy, A. M. (2015). One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current opinion in microbiology, 23*, 110-120.

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., . . . Jarvis, E. D. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature biotechnology, 30*(7), 693.

Kunath, B. J., Bremges, A., Weimann, A., McHardy, A. C., & Pope, P. B. (2017). Metagenomics and CAZyme discovery. In *Protein-Carbohydrate Interactions* (pp. 255-277): Springer.

Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., & Hugenholtz, P. (2008). A bioinformatician's guide to metagenomics. *Microbiol. Mol. Biol. Rev., 72*(4), 557-578.

Latorre-Pérez, A., Villalba-Bermell, P., Pascual, J., Porcar, M., & Vilanova, C. (2019). Assembly methods for nanopore-based metagenomic sequencing: a comparative study. *bioRxiv*, 722405.

Levasseur, A., Drula, E., Lombard, V., Coutinho, P. M., & Henrissat, B. (2013). Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. *Biotechnology for biofuels, 6*(1), 41.

Li, M., Pu, Y., & Ragauskas, A. J. (2016). Current understanding of the correlation of lignin structure with biomass recalcitrance. *Frontiers in chemistry, 4*, 45.

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., . . . Liu, B. (2012). Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. *Briefings in functional genomics, 11*(1), 25-37.

Loman, N. J., Quick, J., & Simpson, J. T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature methods, 12*(8), 733-735.

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., & Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic acids research, 42*(D1), D490-D495.

Lorenz, T. C. (2012). Polymerase chain reaction: basic protocol plus troubleshooting and optimization strategies. *JoVE (Journal of Visualized Experiments)*(63), e3998.

Lu, H., Giordano, F., & Ning, Z. (2016). Oxford Nanopore MinION sequencing and genome assembly. *Genomics, proteomics & bioinformatics, 14*(5), 265-279.
Lynd, L. R., Weimer, P. J., Van Zyl, W. H., & Pretorius, I. S. (2002). Microbial cellulose utilization: fundamentals and biotechnology. *Microbiol. Mol. Biol. Rev., 66*(3), 506-577.
Malherbe, S., & Cloete, T. E. (2002). Lignocellulose biodegradation: fundamentals and applications. *Reviews in Environmental Science and Biotechnology, 1*(2), 105-114.
Martínez-Porchas, M., Villalpando-Canchola, E., & Vargas-Albores, F. (2016). Significant loss of sensitivity and specificity in the taxonomic classification occurs when short 16S rRNA gene sequences are used. *Heliyon, 2*(9), e00170.
Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences, 74*(2), 560-564.
McKendry, P. (2002). Energy production from biomass (part 1): overview of biomass. *Bioresource technology, 83*(1), 37-46.
Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nature reviews genetics, 11*(1), 31-46.
Mikheyev, A. S., & Tin, M. M. (2014). A first look at the Oxford Nanopore MinION sequencer. *Molecular ecology resources, 14*(6), 1097-1102.
Milanese, A., Mende, D. R., Paoli, L., Salazar, G., Ruscheweyh, H.-J., Cuenca, M., . . . Coelho, L. P. (2019). Microbial abundance, activity and population genomic profiling with mOTUs2. *Nature communications, 10*(1), 1-11.
Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics, 95*(6), 315-327.
Mohnen, D. (2008). Pectin structure and biosynthesis. *Current opinion in plant biology, 11*(3), 266-277.
Moraïs, S., Morag, E., Barak, Y., Goldman, D., Hadar, Y., Lamed, R., . . . Bayer, E. A. (2012). Deconstruction of lignocellulose into soluble sugars by native and designer cellulosomes. *MBio, 3*(6), e00508-00512.
Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., & Erlich, H. (1986). *Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction.* Paper presented at the Cold Spring Harbor symposia on quantitative biology.
Naas, A. E., Mackenzie, A., Mravec, J., Schückel, J., Willats, W. G. T., Eijsink, V., & Pope, P. (2014). Do rumen Bacteroidetes utilize an alternative mechanism for cellulose degradation? *MBio, 5*(4).
Nielsen, K., Mogensen, H. S., Hedman, J., Niederstätter, H., Parson, W., & Morling, N. (2008). Comparison of five DNA quantification methods. *Forensic Science International: Genetics, 2*(3), 226-230.
ONT. (2019a). How nanopore sequencing works animation. Retrieved from https://nanoporetech.com/resource-centre/how-nanopore-sequencing-works-animation
ONT. (2019b). Types of nanopores.
ONT. (2020). PromethION. Retrieved from https://nanoporetech.com/products/promethion
Ostrowski, M., Rosa, S. L. L., Kunath, B. J., Yao, T., Buttner, D., Flint, G., . . . Martens, E. C. (2020 (in preparation)). Digestion of the food additive polysaccharide xanthan gum in the human gut requires a single uncultivated bacterium.
Overholt, W. A., Hölzer, M., Geesink, P., Diezel, C., Marz, M., & Küsel, K. (2019). Inclusion of Oxford Nanopore long reads improves all microbial and phage metagenome-assembled genomes from a complex aquifer system. *bioRxiv*.
PacBio. (2020). SMRT Sequencing. Retrieved from https://www.pacb.com/smrt-science/smrt-sequencing/
Pop, M. (2009). Genome assembly reborn: recent computational challenges. *Briefings in bioinformatics, 10*(4), 354-366.
Pop, M., & Salzberg, S. L. (2008). Bioinformatics challenges of new sequencing technology. *Trends in genetics, 24*(3), 142-149.

Proctor, L. M. (2011). The human microbiome project in 2011 and beyond. *Cell host & microbe, 10*(4), 287-291.

Purdy, K., Embley, T., Takii, S., & Nedwell, D. (1996). Rapid extraction of DNA and rRNA from sediments by a novel hydroxyapatite spin-column method. *Appl. Environ. Microbiol., 62*(10), 3905-3907.

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., . . . Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics, 13*(1), 341.

Quick, J., Ashton, P., Calus, S., Chatt, C., Gossain, S., Hawker, J., . . . Peters, T. (2015). Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. *Genome biology, 16*(1), 114.

Quick, J., & Loman, N. J. (2019). DNA Extraction Strategies for Nanopore Sequencing. *Nanopore Sequencing: An Introduction*, 91.

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature biotechnology, 35*(9), 833.

Rang, F. J., Kloosterman, W. P., & de Ridder, J. (2018). From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome biology, 19*(1), 90.

Rausch, P., Rühlemann, M., Hermes, B. M., Doms, S., Dagan, T., Dierking, K., . . . Hentschel, U. (2019). Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms. *Microbiome, 7*(1), 1-19.

Rengarajan, K., Cristol, S. M., Mehta, M., & Nickerson, J. M. (2002). Technical Brief Quantifying DNA concentrations using fluorometry: A comparison of fluorophores. *Molecular Vision, 8*, 416-421.

Richardson, E. J., & Watson, M. (2013). The automatic annotation of bacterial genomes. *Briefings in bioinformatics, 14*(1), 1-12.

Robe, P., Nalin, R., Capellano, C., Vogel, T. M., & Simonet, P. (2003). Extraction of DNA from soil. *European Journal of Soil Biology, 39*(4), 183-190.

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., . . . Edwards, M. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature, 475*(7356), 348-352.

Ruiz-Villalba, A., van Pelt-Verkuil, E., Gunst, Q. D., Ruijter, J. M., & van den Hoff, M. J. (2017). Amplification of nonspecific products in quantitative polymerase chain reactions (qPCR). *Biomolecular detection and quantification, 14*, 7-18.

Russell, J. B., Muck, R. E., & Weimer, P. J. (2009). Quantitative analysis of cellulose degradation and growth of cellulolytic bacteria in the rumen. *FEMS microbiology ecology, 67*(2), 183-197.

Sanger, F., & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology, 94*(3), 441-448.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences, 74*(12), 5463-5467.

Sangwan, N., Xia, F., & Gilbert, J. A. (2016). Recovering complete and draft population genomes from metagenome datasets. *Microbiome, 4*(1), 8.

Schadt, E. E., Turner, S., & Kasarskis, A. (2010). A window into third-generation sequencing. *Human molecular genetics, 19*(R2), R227-R240.

Schmidt, M. H.-W., Vogel, A., Denton, A. K., Istace, B., Wormit, A., van de Geest, H., . . . Pfaff, C. (2017). De novo assembly of a new Solanum pennellii accession using nanopore sequencing. *The Plant Cell, 29*(10), 2336-2348.

Schrader, C., Schielke, A., Ellerbroek, L., & Johne, R. (2012). PCR inhibitors–occurrence, properties and removal. *Journal of applied microbiology, 113*(5), 1014-1026.

Seshadri, R., Leahy, S. C., Attwood, G. T., Teh, K. H., Lambie, S. C., Cookson, A. L., . . . Varghese, N. J. (2018). Cultivation and sequencing of rumen microbiome members from the Hungate1000 Collection. *Nature biotechnology, 36*(4), 359.

Shin, J., Lee, S., Go, M.-J., Lee, S. Y., Kim, S. C., Lee, C.-H., & Cho, B.-K. (2016). Analysis of the mouse gut microbiome using full-length 16S rRNA amplicon sequencing. *Scientific reports, 6*, 29681.

Shin, S. C., Kim, H., Lee, J. H., Kim, H.-W., Park, J., Choi, B.-S., . . . Kim, S. (2019). Nanopore sequencing reads improve assembly and gene annotation of the Parochlus steinenii genome. *Scientific reports, 9*(1), 1-10.

Sommer, F., & Bäckhed, F. (2013). The gut microbiota—masters of host development and physiology. *Nature Reviews Microbiology, 11*(4), 227-238.

Spanogiannopoulos, P., Bess, E. N., Carmody, R. N., & Turnbaugh, P. J. (2016). The microbial pharmacists within us: a metagenomic view of xenobiotic metabolism. *Nature Reviews Microbiology, 14*(5), 273.

Stewart, R. D., Auffret, M. D., Warr, A., Walker, A. W., Roehe, R., & Watson, M. (2019). Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nature biotechnology, 37*(8), 953.

Stothard, P., & Wishart, D. S. (2006). Automated bacterial genome analysis and annotation. *Current opinion in microbiology, 9*(5), 505-510.

Suzuki, M. T., & Giovannoni, S. J. (1996). Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl. Environ. Microbiol., 62*(2), 625-630.

Sze, M. A., & Schloss, P. D. (2019). The impact of DNA polymerase and number of rounds of amplification in PCR on 16S rRNA gene sequence data. *mSphere, 4*(3), e00163-00119.

Tikhonov, M., Leach, R. W., & Wingreen, N. S. (2015). Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. *The ISME journal, 9*(1), 68-80.

Trimble, W. L., Keegan, K. P., D'Souza, M., Wilke, A., Wilkening, J., Gilbert, J., & Meyer, F. (2012). Short-read reading-frame predictors are not created equal: sequence error causes loss of signal. *BMC bioinformatics, 13*(1), 183.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The human microbiome project. *Nature, 449*(7164), 804-810.

Van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in genetics, 30*(9), 418-426.

Voragen, A. G., Coenen, G.-J., Verhoef, R. P., & Schols, H. A. (2009). Pectin, a versatile polysaccharide present in plant cell walls. *Structural Chemistry, 20*(2), 263.

Voytas, D. (2000). Agarose gel electrophoresis. *Current protocols in molecular biology, 51*(1), 2.5 A. 1-2.5 A. 9.

Wallace, R. J., Rooke, J. A., McKain, N., Duthie, C.-A., Hyslop, J. J., Ross, D. W., . . . Roehe, R. (2015). The rumen microbial metagenome associated with high methane production in cattle. *BMC genomics, 16*(1), 839.

Wang, J., & Jia, H. (2016). Metagenome-wide association studies: fine-mining the microbiome. *Nature Reviews Microbiology, 14*(8), 508-522.

Wang, J., Moore, N. E., Deng, Y.-M., Eccles, D. A., & Hall, R. J. (2015). MinION nanopore sequencing of an influenza genome. *Frontiers in Microbiology, 6*, 766.

Wang, Y., & Qian, P.-Y. (2009). Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. *PloS one, 4*(10).

Wang, Y., Yang, Q., & Wang, Z. (2015). The evolution of nanopore sequencing. *Frontiers in Genetics, 5*(449). doi:10.3389/fgene.2014.00449

Wick, R. R., & Holt, K. E. (2019). Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *F1000Research, 8*(2138), 2138.

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS computational biology, 13*(6), e1005595.

Williamson, K. E., Kan, J., Polson, S. W., & Williamson, S. J. (2011). Optimizing the indirect extraction of prokaryotic DNA from soils. *Soil Biology and Biochemistry, 43*(4), 736-748.

Xu, L., Sun, L., Guan, G., Huang, Q., Lv, J., Yan, L., . . . Zhang, Y. (2019). The effects of pH and salts on nucleic acid partitioning during phenol extraction. *Nucleosides, Nucleotides and Nucleic Acids, 38*(4), 305-320.

Ye, C., Hill, C. M., Wu, S., Ruan, J., & Ma, Z. S. (2016). DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Scientific reports, 6*, 31900.

Yılmaz, M., Ozic, C., & Gok, İ. (2012). Principles of nucleic acid separation by agarose gel electrophoresis. *Gel Electrophoresis–Principles and Basics, 33*.

Yin, Y., Mao, X., Yang, J., Chen, X., Mao, F., & Xu, Y. (2012). dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic acids research, 40*(W1), W445-W451.

Zhu, W., Lomsadze, A., & Borodovsky, M. (2010). Ab initio gene identification in metagenomic sequences. *Nucleic acids research, 38*(12), e132-e132.

Appendices

Appendix A:

Results from extraction of environmental DNA

Sample	Fluid (F)/Particle (P)	Biomass (g)	Qbit (ng/µL)	Yield (µg)	A260/280	A260/230
RUM-543	F	0,2498	66,8	6,68	1,86	1,92
RUM-544	F	0,251	82,6	8,26	1,84	1,99
RUM-545	F	0,2506	61,6	6,16		
RUM-546	F	0,2503	91,1	9,11		
RUM-547	F	0,2511	78,9	7,89		
RUM-548	F	0,2506	81,5	8,15	1,84	1,91
RUM-549	F	0,2501	75,5	7,55	1,86	1,87
RUM-550	F	0,2497	57,1	5,71	1,88	1,98
RUM-551	F	0,2504	63,3	6,33		
RUM-552	F	0,2509	47,8	4,78	2,05	1,66
RUM-553	F	0,2506	66,2	6,62	1,85	1,96
RUM-554	F	0,2509	70,3	7,03		
RUM-555	F	0,2504	59,6	5,96		
RUM-556	F	0,2508	67,8	6,78	2,14	1,58
RUM-557	F	0,2513	53,4	5,34		
RUM-558	F	0,2505	55,6	5,56		
RUM-559	F	0,2497	74,9	7,49	1,83	1,92
RUM-561	F	0,2501	69,5	6,95		
RUM-563	F	0,2501	59,7	5,97	1,88	1,96
RUM-565	F	0,2511	50,4	5,04		
RUM-567	F	0,2514	45,1	4,51		
RUM-568	F	0,2495	57,9	5,79	1,85	1,92
RUM-570	F	0,25	76,9	7,69	1,88	1,94
RUM-572	F	0,2499	57	5,7		
RUM-575	P	0,2506	97,8	9,78	1,85	1,85
RUM-576	P	0,2506	113	11,3	1,85	1,96
RUM-577	P	0,2509	69,8	6,98		
RUM-578	P	0,2517	102	10,2		
RUM-579	P	0,2515	81,6	8,16		
RUM-580	P	0,2513	112	11,2	1,86	1,92
RUM-581	P	0,2502	89,2	8,92	1,83	1,79
RUM-582	P	0,2512	77,7	7,77	1,86	1,89
RUM-583	P	0,2501	37,4	3,74		
RUM-584	P	0,2506	64,3	6,43	1,88	1,73
RUM-585	P	0,2501	73,3	7,33	1,86	1,89
RUM-586	P	0,2497	114	11,4		
RUM-587	P	0,2509	80,8	8,08		
RUM-588	P	0,2513	121	12,1	1,91	1,96
RUM-589	P	0,2498	110	11		

RUM-590	P	0,2498	136	13,6		
RUM-591	P	0,2492	76,1	7,61	1,85	1,81
RUM-593	P	0,2507	83,8	8,38		
RUM-595	P	0,2501	135	13,5	1,87	2,01
RUM-597	P	0,2497	70	7		
RUM-599	P	0,2513	70,2	7,02		
RUM-600	P	0,2504	86,9	8,69	1,92	1,89
RUM-602	P	0,2502	113	11,3	1,9	2,13
RUM-604	P	0,2495	97,1	9,71		

Appendix B

MAG statistics

XDC03

Bin Id	Completeness	Genome size (bp)	# scaffolds	closest relative (MiGA)	% AAI (MiGA)
INDI03.1	98,27	4828551	42	Parabacteroides distasonis NZ AP019729	96,75
INDI03.10	95,75	13057750	2709	Bacteroides caecimuris NZ CP015401	67,92
INDI03.11	98,6	9986514	1386	Clostridium bolteae NZ CP022464	91,56
INDI03.12	0	1293041	359	Bacteroides cellulosilyticus NZ CP012801	46,26
INDI03.13	15,79	436355	197	Phascolarctobacterium faecium NZ AP019004	97,1
INDI03.14	10,34	349379	151	Veillonella atypica CP020566	95,52
INDI03.2	98,85	7434917	76	Blautia producta NZ CP035945	56,31
INDI03.3	99,65	4518942	69	Escherichia coli NZ CP025747	99,57
INDI03.4	97,99	2934651	65	Monoglobus pectinilyticus NZ CP020991	47,28
INDI03.5	98,25	11113103	632	Clostridium saccharolyticum WM1 NC 014376	65,61
INDI03.6	63,57	3461272	162	Parabacteroides sp. CT06 NZ CP022754	72,47
INDI03.7	8,77	642760	38	Desulfitobacterium dehalogenans ATCC 51507 NC 018017	38,27
INDI03.8	0	805359	69	Mycoplasma wenyonii str. Massachusetts NC 018149	35,88

INDI03.9	12,5	2484859	779	Bacteroides cellulosilyticus NZ CP012801	83,31

XDCOriginal

Bin Id	Completeness	Genome size (bp)	scaffolds	closest relative (MIGA)	%AAI
INDIORI.1	99,13	6141190	48	Bacteroides intestinalis NZ CP041379	93,48
INDIORI.10	99,22	4598235	61	Parabacteroides distasonis NZ CP040468	98,73
INDIORI.11	84,16	5807572	476	Parabacteroides sp. CT06 NZ CP022754	71,6
INDIORI.12	1,72	217838	24	Parabacteroides sp. CT06 NZ CP022754	65,82
INDIORI.13	34,27	907155	407	Flavonifractor plautii NZ CP015406	96,69
INDIORI.14	0	247590	80	Dehalobacter sp. CF NC 018867	37,05
INDIORI.15	4,17	324714	156	Anaerostipes rhamnosivorans NZ CP040058	40,2
INDIORI.2	5,26	280303	3	Monoglobus pectinilyticus NZ CP020991	41,27
INDIORI.3	0	204562	8	Lachnoclostridium sp. YL32 NZ CP015399	54,89
INDIORI.4	92,05	5936853	82	Clostridium bolteae NZ CP022464	73,79
INDIORI.5	0	307881	8	Streptococcus sp. 1643 NZ CP040231	34,45
INDIORI.6	90,1	2186472	24	Monoglobus pectinilyticus NZ CP020991	46,95
INDIORI.7	0	401209	53	Butyricimonas faecalis NZ CP032819	43,75
INDIORI.8	99,26	6880578	132	Bacteroides thetaiotaomicron NZ CP012937	95,51
INDIORI.9	95,32	5881521	80	Clostridium bolteae NZ CP022464	98,84

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway