Norwegian University
of Life Sciences

Master's Thesis 2020 60 ECTS
Faculty of Biosciences

Exploring different aspects of a metagenomic study using third-generation sequencing

Alexsander Lysberg
Master Biology, specialization in molecular genomics and evolution

Exploring different aspects of a metagenomic study using third-generation sequencing

Master's Thesis

Alexsander Lysberg

Protein engineering and Proteomics Group
Faculty of Chemistry, Biotechnology, and Food Science
Norwegian University of Life Sciences
2020

Acknowledgments

This thesis was performed at the faculty of Chemistry, Biotechnology and Food Sciences at the Norwegian University of Life Sciences (NMBU) under the supervision of Assoc. Prof. Phillip B. Pope and Dr. Live Heldal Hagen.

Firstly, I would like to thank my main supervisor Prof. Phillip B. Pope, for introducing me to this project and guiding me through the obstacles this project had in store. Thank you for being a calm anchor point to rely on in a troublesome period and thank you for encouraging me to follow my own ideas.

Dr. Live Heldal Hagen, thank you for all the help you have provided me with in the lab and for encouraging critical thinking regarding the protocols we used. I have learned a ton thanks to you. Thank you for your patience regarding the plethora of questions you have had to endure.
Thank you, Dr. Sabina Leanti La Rosa, for sharing your data with me and answering all my questions. Your help was essential for this thesis.
I would also like to thank the Protein engineering and Proteomics group (PEP), for allowing me to write my thesis in the group and a thanks for everyone working there. Thank you all for always having time to help me around the lab and for introducing me to a friendly and skillful environment.
I would like to thank all my friends and fellow master students for accompanying me along this journey.
Your company and encouragement have managed to keep me sane throughout this process.
Special thanks to Tina Johannessen for all your help and company, having someone work with in the lab and discuss our experience with has been vital. Also, a special thanks to Morten Nilsen for reviewing my work, giving feedback and encouraging me keep going.
Lastly, I would like to thank my lovely family. Thank you for all your help and support throughout my studies and thank you for believing in me.

Ås, 2020

_____

Alexsander Lysberg

Abstract

The microbial communities found in the gastrointestinal tract of mammals, such as within the rumen of herbivores or in the gut of humans exert significant influence on their host via their vast array of metabolic functions. Understanding the composition and function of these communities can help combat some of the greatest challenges modern society faces. By understanding their function and what biochemical properties they instigate, they can be used to combat famine, reduce greenhouse gas emissions, increase nutrient absorption, and increase overall human health.

These microbial communities have been challenging to explore for decades due to limitations in technology, but through the rise of second and now third generation sequencing platforms, the generation of genomic information via Metagenomic Assembled Genomes (MAGs) has become faster, cheaper and more accurate. This has allowed scientists to explore a multitude of communities previously deemed too expensive and too complex to analyze. Despite this, the number of high-quality MAGs, used to determine biochemical function in online databases is far from optimal to this day.

In this study, we explore the different methodological steps that are required to perform metagenomic analysis of complex communities, with a particular focus on recovering MAGs that represent microbial populations. We applied these approaches to both rumen samples from sheep and gut samples from humans, which were also subjected to different sequencing platforms, in order to determine the strengths and weaknesses of each alternative. Differences in sampling method, DNA extraction method, sequencing platform and analyzing tools were explored to determine which were better equipped for the task of generating high-quality MAGs. Finally, we explored the applicability of long read sequencing and how it will advance metagenomic studies in the coming years.

Sammendrag

De mikrobielle samfunnene som finnes i mage-tarmkanalen hos pattedyr, som i vommen til planteetere eller i tarmen til mennesker, har betydelig innflytelse på verten deres gjennom deres ulike metabolske funksjoner. Å forstå sammensetningen og funksjonen til disse samfunnene kan bidra til å bekjempe noen av de største utfordringene det moderne samfunnet står ovenfor. Ved å forstå deres funksjon og hvilke biokjemiske egenskaper de har, kan man benytte de til å bekjempe hungersnød, redusere klimagassutslipp, øke næringsopptaket og bedre allmenhelsen til mennesker.

Disse mikrobielle samfunnene har vært utfordrende å utforske i flere tiår på grunn av begrensninger i teknologien, men gjennom forbedringer og utviklingen av andre og nå tredje generasjons sekvenseringsplattformer er utforskningen av genomisk informasjon via Metagenomic Assembled Genomes (MAGs) blitt raskere, billigere og mer nøyaktig. Dette har gjort det mulig for forskere å utforske et mangfold av samfunn, som tidligere ble ansett for for dyre og for kompliserte til å utforske. Til tross for dette er antallet MAGs av høy kvalitet, brukt til å bestemme biokjemisk funksjon, i nettbaserte databaser langt fra optimalt, selv i dag.

I denne studien utforsker vi de forskjellige metodologiske trinnene som er nødvendige for å utføre metagenomisk analyse av komplekse samfunn, med særlig fokus på å uthente MAGs som representerer mikrobielle populasjoner. Vi praktiserte disse metodene på både vom prøver fra sauer og tarmsprøver fra mennesker, som igjen ble sekvensert på ulike sekvenseringsplattformer, for å utforske fordeler og ulemper ved hvert alternativ. Forskjeller i prøvetakingsmetode, DNA-ekstraksjonsmetode, sekvenseringsplattform og analyseverktøy ble undersøkt for å bestemme hvilke som var bedre rustet til oppgaven med å generere MAG-er av høy kvalitet. Til slutt undersøkte vi anvendeligheten av tredjegenerasjons sekvensering og hvordan det vil fremme metagenomiske studier de kommende årene.

Abbreviations

OTU	Operational Taxonomic Unit
PCR	Polymerase Chain Reaction
rRNA	Ribosomal ribonucleic acid
MAGs	Metagenomic Assembeled Genomes
VFA	Volatile Fatty Acids
CAZymes	Carbohydrate Active enzymes
E.C Number	Enzyme Comission Number
KEGG	Kyoto Encyclopedia of Genes and Genomes

Contents

1 Background

Microorganisms surround us daily; they exist in complex communities and represent the largest genetic diversity on Earth. They are estimated to be responsible for 50-78% of the worlds' biomass and manage the worlds biogeochemical cycles (Kallmeyer, Pockalny, Adhikari, Smith, & D'Hondt, 2012). They recycle essential elements, forms soil and breaks down both natural and anthropogenic organic material (Heyer et al., 2017; Rodríguez-Valera, 2004). Certain microorganisms produces bioactive products that promotes health and can be utilized in various societal, scientific and industrial fields (Garbeva, Veen, & Elsas, 2004). Society has long benefitted from the bioactive properties of microbial communities. Ever since the discovery of bread-baking and brewing have these communities been utilized to our benefit.

Although there is a vast potential in microbial communities, little is understood about them. The study of microbiology can be considered to have started alongside the invention of the microscope in the 16th century. Nevertheless, researchers have only recently started to study the genomic composition of diverse microbial communities.

The study of microbiology is founded on the exploration of microbes through cultivation. However, this approach is not optimal to study communities. The amount of bacteria actually suited for cultivation under standard conditions is estimated to be roughly 0.1-1.0% (Staley & Konopka, 1985) and this 1% is not the most abundant in an environment, and rarely the ones of biochemical interest, but rather the most adaptable to alterations in environment.
According to (Hugenholtz, 2002) the majority of microbial research conducted in the period 1991-1997 only studied the same eight bacterial genera, due to their ability to outcompete other microbes on agar-cultures. These "microbial weeds" make the traditional culture-based approach for community study unreliable (Hugenholtz, 2002).

Fortunately, through the development of 2nd and 3rd generation sequencing machines, this problem has been solved. The newer technology doesn't rely on bacterial cultivation and as a result the amount of Metagenomic Assembled Genomes (MAGs), which are genomes assembled from a community sample, have grown exponentially the last decade (see figure 1.1). By the end of 2016 there were 2,866 Single-Cell Assembled Genomes (SAGs) and 4,622 Metagenomic Assembled Genomes (MAGs) (Robert M Bowers et al., 2017), but when compared to the number of genomes assembled by 2019, these numbers seem inconsequential. Studies such as Almeida et al., 2019 and Pasolli et al., 2019 managed to sequence and assemble

roughly 250,000 MAGs combined. This illustrates how rapid the field of metagenomics are evolving, and what more to expect from it in the future.

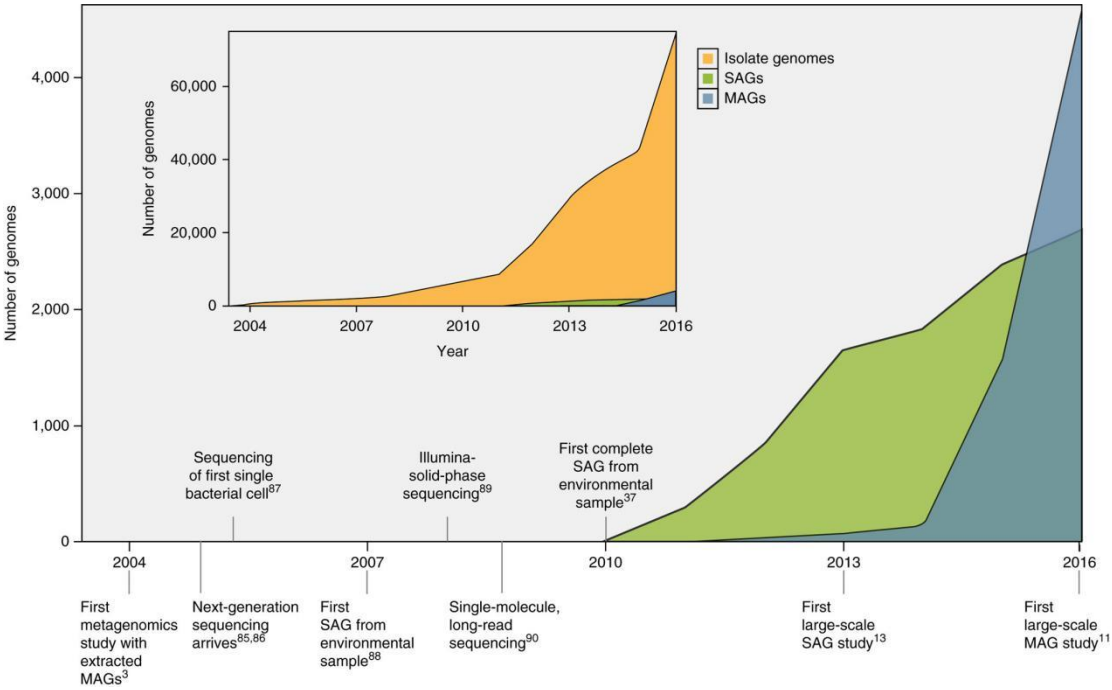


Figure 1.1 Increase in number of Single-cell Assembeled genoms (SAGs) and Metagenomic Assembeled Genomes (MAGs) over the period 2010-2016 (Robert M Bowers et al., 2017). The trend has grown exponentially over the last few years through the development of newer sequencing technologies like Oxfords' Nanopore and PacBio' Single Molecule Real Time Sequencing (SMRT) and increased interest in the field. Databases like JGI Gold (https://gold.jgi.doe.gov/distribution) and EBI metagenomics (https://www.ebi.ac.uk/metagenomics/) contain >130,000 MAGs combined. These numbers dwarf the amount of MAGs in this figure, and illustrated the rapid growth the field is experiencing. Illustration taken from Bowers et al.2017

1.1 The importance of gut and rumen microbiomes

The microbiomes consists of organisms from various taxa across the tree of life, such as fungi, eukaryotes, bacteria, protozoa and viruses (Jose C. Clemente, 2012,). They help their host break down complex molecules like fibers and starch to smaller and more easily digestible components such as volatile fatty acids (Dijkstra, 1994). Through processes like fermentation and hydrolysis they produce nutrients that benefit themselves and their host (Moran, 2005), and are fundamental for their hosts health.

Two of the most explored microbiomes are the ones of ruminants and humans. These microbiomes' production of bioactive products impacts their hosts' health and are important to understand in order to utilize them. Through the exploration of the ruminant's microbiome we

might enhance meat and dairy production, and simultaneously reduce their greenhouse gas emissions. While the exploration of the human microbiome has been closely linked to the maturation and function of our immune system and has been found to have a major impact on our general health (Czerkawski, 1986; Lloyd-Price, Abu-Ali, & Huttenhower, 2016). Solving these problems could help combat global problems like greenhouse-gas emissions and famine.

1.2 The Rumen

The ruminant's gastrointestinal tract is comprised of four compartments, the rumen, reticulum, omasum and abomasum. These four compartments combined is responsible for digestion of consumed biomass and absorption of nutrients in ruminants (Moran, 2005). When the ruminants are fed, the biomass is broken into smaller pieces through rumination (cud-chewing). The rumination process makes the biomass more susceptible to carbohydrate-hydrolysis and bacterial fermentation. The rumination is needed to extract the nutrients found in lignocellulose, which comprises most of the ruminants' diet.

Lignocellulose consists mainly of cellulose, hemicellulose and pectin, and is found in the plant cell-wall where the different polymers interact to create a rigid recalcitrant structure (Moraïs et al., 2012). Lignocellulose is a complex material and requires a wide array of enzymes to utilize, which ruminants themselves cannot encode for, and are therefore dependent on their microbiome.

The enzymes needed for lignocellulose degradation are called carbohydrate active enzymes (CAZymes) can be divided into 5 groups depending on their function (http://www.cazy.org/). Glycosyl Transferases (GTs) are transferases responsible for carbohydrates assemblage, as they introduce glycoside linkages. Polysaccharide Lysases (PLs) cleave activated glyosidic linkages, Glycoside hydrolysis (GHs) catalyzes the hydrolysis of glyosidic bonds between carbohydrates and Carbohydrate Esterases (CEs)catalyzes the acylation of the Oxygen or Nitrogen of substituted saccharides. PLs, GHs and CEs are all involved in carbohydrate degradation. Axilliary Activities (AAs) are redox enzymes that act in conjunction with the other CAZymes.

The rumen microbiota comprises a large variety of bacteria, which aid in the degradation of complex polysaccharides. The more important bacteria are the ones involved in cellulose, pectin, lactate proteolytic and lipolytic degradation. All of which play a major role in digestion, pH regulation and provides energy for the host.

The host receives energy from bacterial fermentation and carbohydrate-hydrolysis. The desired end-product of these processes are volatile fatty acids (VFA) (Dijkstra, 1994). VFAs consists of 1-6 carbon atoms, and molecules like propionate, acetate and butyrate can transverse the hosts epithelium and be used in its energy metabolism. Roughly 70% of ruminants caloric requirements come from VFA (Bergman, 1990).

Another important member of the microbiota is the protozoa, which contributes to 40-80% of the rumen biomass. The majority of protozoa (90%) is involved in hydrolysis and fermentation of cellulose and (Castillo-González, 2014). Bacteria and protozoa together degrades together 70-80% of the ingested dry-matter (Moran, 2005).

In addition, fungi represents 8 % of the biomass in the rumen and aid in hydrolyzing cellulose and hemicellulose by producing enzymes capable of breaking down plant-cell wall components, and plays an important role in digestion as other microbes gain access to the plant material ingested (Castillo-González, 2014).

1.3 The Human Gut

The human microbiome differs quite a bit from the one found in ruminants. While the ruminants heavily depend on their microbiome for energy absorption, the human gut operates more independently. Roughly 85% of carbohydrates, 66%-95% of proteins and all the fats are absorbed before the food enters the large intestine where fermentation takes place (Krajmalnik-Brown, Ilhan, Kang, & DiBaise, 2012). Approximately ~10% of our energy intake comes in the form of VFA as a result of fermentation and carbohydrate hydrolysis (Bergman, 1990).

Despite this, the microbes still play a critical role in numerous physiological and microbiological processes which aids both our health and metabolism. The microbiota is fundamental in influencing host-cell proliferation (Ijssennagger et al., 2015) regulate abnormal/excessive blood vessel formation (Reinhardt et al., 2012), regulate intestinal endocrine functions by interacting with hosts' hormone production (Neuman, Debelius, Knight, & Koren, 2015), neurologic signaling through microbial serotonin production (Yano et al., 2015), influencing bone density by bodyweight regulation (I. Cho et al., 2012), micronutrient synthesis and drug metabolism (Ijssennagger et al., 2015). Furthermore, the microbiome have been found to be fundamental in e healthy immune-system, by aiding in maturation and

continued education of the hosts' immune response (Fulde & Hornef, 2014) and suppresses pathogen overgrowth (N. Kamada, Chen, Inohara, & Núñez, 2013).

The structure of the human gut microbiome is heavily altered by factors like, age, host genetics, diet and local environment (Browne, Neville, Forster, & Lawley, 2017). This makes it hard to determine what a healthy baseline for all humans should be and how to optimize it. Correlations between poor diet and malnourishment indicate that both obesity and starvation have detrimental effect on our microbiome. Paradoxically this malnourishment leads to a weakened immune system, which in turn reduces the body's ability to absorb nutrients, creating a negative feedback loop (Kau, Ahern, Griffin, Goodman, & Gordon, 2011) (See figure 1.2 below).

Undernutrition is responsible for ~45% of the death of children under the age of 5, illustrating the massive global problem malnourishment, and poor microbiome composition present (Bryce, Boschi-Pinto, Shibuya, Black, & Group, 2005; Robertson, Manges, Finlay, & Prendergast, 2019). Other disease, such as Crohn's disease and ulcerative colitis are also linked to the microbial communities in the human gut (Morgan et al., 2012). To help combat problems and diseases such as these, we need to further our knowledge about our microbiome, and its functions.

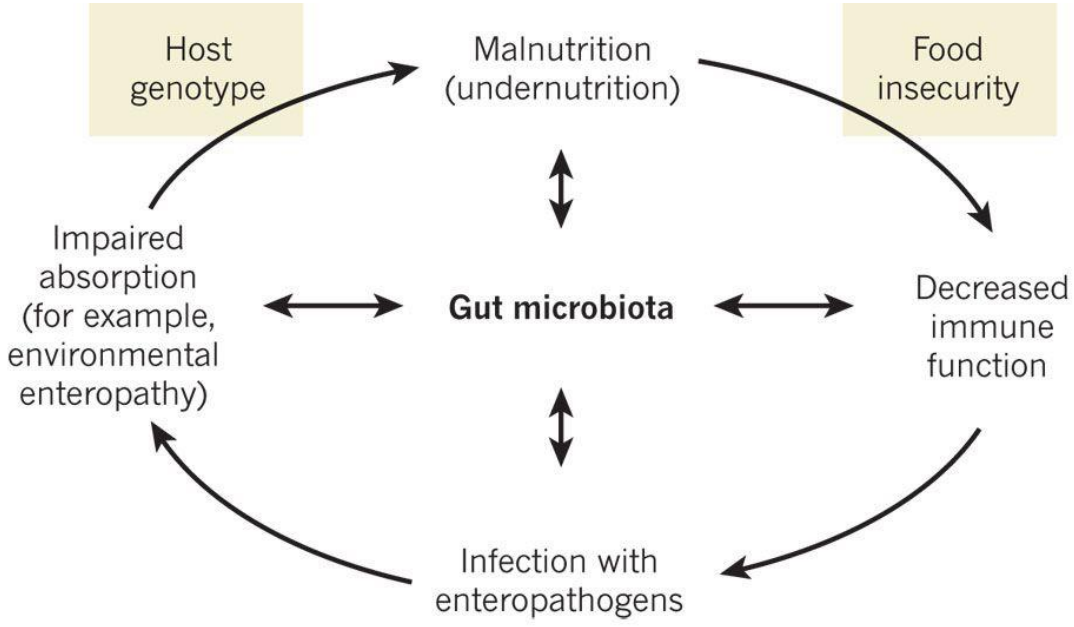


Figure 1.2 **Proposed relationship between gut microbiome, nutrition and immune system.** Illustrates how poor microbiome as a result of malnourishment and poor immune function could result in an increase of infections, which in turn results in a reduced ability to absorb nutrients,

resulting in a negative feedback loop.
Illustration gathered from https://www.nature.com/articles/nature10213/figures/1.

1.4 How to study microbiomes: metagenomics

The study of microbial communities is a relatively new field of science. The collective interest in genomes spiked after projects like Human Genome Project took place (Boeke et al., 2016). Further development in sequencing machines, techniques and data processing have also made it more of a prominent method to examine both macro and micro-organisms. Notably it was the development of second-generation sequencing, or high-throughput sequencing, that allowed researchers to properly study complex microbial communities with unprecedented resolution and throughput.

Metagenomics is a powerful research technique that help us explore microbe's species-richness, distribution and relationship to each other in samples (Barzon, Lavezzo, Militello, Toppo, & Palù, 2011). It is a culture-independent approach and analyzes the collective set of genomes found in a sample taken directly from a community of interest. The most used techniques to analyze metagenomes is 16S rRNA sequencing and shotgun metagenomic sequencing.

Microbial community analysis using 16S rRNA sequencing utilizes Polymerase Chain Reactions (PCR) to amplify the ribosomal RNA in prokaryotes. The ribosomal RNA contains several hypervariable regions (V1-V9) which are used to determine phylogenetic rank. 16S rRNA gene analysis is well suited towards exploring the taxonomic diversity of prokaryotes in communities. It has a high bacteria coverage through online reference databases, has a low risk of false positives and it is cheap. However, care must still be taken to avoid biases that can arise through PCR, depending on how many cycles the PCR runs, what primes are being used and what analyzing pipeline is being applied.

Shotgun metagenomic sequencing, unlike 16S rRNA gene analysis, sequences all the given genomic DNA from a given sample, instead of just the ribosomal RNA in prokaryotes. As such, shotgun sequencing captures a much broader range of information from a community with a higher level of resolution, meaning we can study genes and their predicted function.
Amplicon and shotgun methods use different processes for this. 16S rRNA uses a method called clustering, while shotgun-based methods use a process called 'binning'. While these two

methods function similarly, by grouping contigs together based on similarities and turning them into operational taxonomic units (OTUs) they operate with different criteria for grouping.

Shortly summarized, 'clustering' groups the reads based on similarity, meaning the reads in each cluster is more similar to each other as opposed the ones in a different cluster, these groups then represent different OTUs. 'Binning' utilizes both previously available information and the intrinsic information from the sample to create its OTUs.

These OTUs represent an algorithms best effort to group the microorganisms together, based on the similarities of their genomic data. Metagenomic binning entail the creation of metagenome assembled genomes (MAGs) that represent as-yet uncultured microorganisms that are in your sample of interest. The usefulness of MAGs depends greatly on their completeness, quality and their degree of novelty. For example, trying to assign taxonomy to MAGs that are less completed is only reliable at more general taxonomic ranks, like domain, kingdom or phylum. In contrast, completed/near-completed genomes can provide more precise proximations.

Which metagenomic approach (16S rRNA gene analysis vs shotgun metagenomics) depends on the aim of one's study. If you're only after taxonomic profiling of a sample, then 16S is cheaper and require less data processing. However, shotgun sequencing provides more data and can be connected to the other `omics`, such as proteomics and transcriptomics, which is best suited for determining biochemical function and metabolic potential of microbes.

1.5 Sequencing Technology

Different DNA sequencing machines provide different output (i.e. sequence reads), with some providing long, but few, while other produce massive number of shorter reads. Each of the sequencing machines provide us with some unique information the others cannot.

1.5.1 First Generation Sequencing

When researchers first started to study the metagenomes, Sanger-sequencing (first generation-sequencing) was used. However, it wasn't very suited towards it. While it creates long and high accuracy, which is beneficial when sequencing novel reference genomes for individual species, it relies on bacterial cloning for sequencing.

The vector-based cloning and *Escherichia coli*-based amplification can implement biases when sequencing. Certain regions in the transferred genome can be cloned less often than others when amplifying due to biological factors, like vector preferences and palindromic sequences. This bias manifests in the form of a lower expected coverage of affected regions (Mardis, 2008). Sanger sequencing also produces a low amount of sequencing reactions, and relies on electrophoresis to detect the sequencing output, all of which make first-generation sequencing too time consuming and ineffective to study samples that contains a multitude of genomes.

1.5.2 Second Generation Sequencing

Second-generation sequencing, or high-throughput sequencing, operates in principle the same way as Sanger-sequencing. It uses DNA polymerase to add fluorescent nucleotides, one by one, to a DNA template, where each nucleotide us identified by its fluorescent tag. The main difference between these technologies is while Sanger sequences only one nucleotide at a time second-generation can sequence a multitude simultaneously. Therefore it can produce massive amounts of reads, they are however shorter (35-250 base-pairs) compared to sanger-sequencing (650-800 base-pairs) (Mardis, 2008). While first generation sequencing can produce hundreds of sequence reactions, second generation can produce thousands-millions, and the sequence output can be detected without the need of electrophoresis. In addition, the samples being sequenced can be taken directly from the gut, without the need of cultivation (Huttenhower, Kostic, & Xavier, 2014) removing potential biases when looking at community composition.

In addition to analyzing population diversity, high-throughput sequencing can also help determine microbe functions through gene annotation and comparative metagenomics (Meyer et al., 2008). By comparing sequence composition, taxonomic diversity, or meta-transcriptomes through online reference databases, high-throughput sequencing can determine the chemical pathways in communities. This can help explore the metabolic potential in microorganisms and how the various microorganisms co-evolve with each other and their host (Cardona et al., 2012). Moreover, understanding the metabolic potential of microbes would be beneficial when examining how the various interactions can benefit us.

Although high-throughput short-read sequencing generates massive amounts of data for genome-recreation, it falls short when trying to finalize and polish genomes. The short length of the reads makes it difficult for bioinformatic software to analyze repeating regions, and ambiguities in alignment of contigs often occur. Since repetitive regions can cover large

portions of a genome, like in humans where it is responsible for nearly half of our gene-material, major difficulties will occur when it is not analyzed properly (Treangen & Salzberg, 2012).

Given the length of sequence reads, first-generation sequencing does not have the same problems with this, and it is still being used to this day on small scale projects, despite being over 40 years old. It is however still too slow to be reliably used to study large scale metagenomic samples. Fortunately, through recent development in sequence technology a multitude of new sequencing machines fills the gaps high-throughput sequencing leaves and is a lot more efficient than traditional Sanger-sequencing.

1.5.3 Third Generation Sequencing: Long read

Third generation (i.e. Long-read) sequencing has gotten a lot of attention recently, and for good reason. 3rd-generation sequencing-technology can analyze single molecules of DNA in real-time without the use PCR amplification, which eliminates potential biases that can arise through amplification or cultivation. This can make it better suited for *de novo* sequencing than 2nd generation and has an increased conscious accuracy for base-calling, if the same DNA strand is sequenced multiple time, which enables rare variant detection (Wick, Judd, & Holt, 2019). However, DNA-strands only sequenced once will have potential faults in its base-calling which can complicate *de novo* assembly as well (Amarasinghe et al., 2020).

Arguably the most important aspect of long-read sequencing is the pore technology itself. While second-generation sequencing needs to 'reassemble' the reads with DNA templates and free nucleotides during sequencing, third generation can allow complete strands of DNA to be analyzed at once. The pores measure the electrical resistance of the nucleotides as they pass through the nanopore, and since each nucleotide has a different resistance to electricity, it can identify them (Jain, Olsen, Paten, & Akeson, 2016). Third-generation library preparations can also produce longer fragments than its predecessors (Amarasinghe et al., 2020). This is due to its potential in exploiting various DNA-polymerases or avoid chemical and biological processes all together. This reduction in chemical handling has to potential to massively increase read-length (Schadt, Turner, & Kasarskis, 2010).

Despite this, 3rd generation sequencing is not without its challenges. For example, it produces fewer reads than 2nd generation, and is not as accurate as 1st and 2nd generation when recognizing nucleotides (Ye, Hill, Wu, Ruan, & Ma, 2016) and DNA-extraction protocols for Third-

generation sequencing needs improvements since possible contaminants can complicate downstream analysis to a large extent (Van Dijk, Auger, Jaszczyszyn, & Thermes, 2014).

The reduced sequencing depth combined with inaccurate base-calling and unoptimized protocols makes 3rd generation sequencing machines unreliable to a certain degree when used alone. Nevertheless, the data provided by long-read sequencing can validate the contigs created through short-read sequencing when used in assembly as a reference or framework.
The production of longer reads makes for more continuous reconstructions of genomes, making it easier to detect insertions, deletions and repeating regions, when assembling contigs. This simplifies assembly and helps increase the overall quality of the genomes (Lee et al., 2016).

Long-read sequencing technology is far from optimized but shows a lot of potential. And given the incredible decrease in cost and increase in base-calling quality over the last few years (Mardis, 2008; Wick et al., 2019), new areas, previously deemed too expensive to examine, will open up for researchers. One can speculate this increase in quality and decrease in costs will continue in the future, possibly leading to long-read sequencing being more reliable, and favorable over short-read sequencing. However, as of now it is best utilized in combination with other sequencing technologies.

By implementing a combination of sequencing technologies, the amount of complete/near-complete genomes in online reference databases could increase. In the period of 2007-2011 only 35% of online reference genomes had an accuracy of >99.99% (Koren et al., 2013), illustrating the need for higher quality reference genomes. Through better developed reference databases, deeper analysis of microbiotas is possible. The more completed or near-completed reference genomes a database contains, and the higher quality they are, the easier it will be to annotate microbes' taxonomy and function. This information can be used by researchers to create more advanced simulations of microbiomes, and possibly induce desired properties in them (Mende et al., 2012).

1.6 DNA Extraction

A contested area for metagenomic studies is the manner of which the DNA is extracted from the samples. While the samples are taken from a community and contain a plethora of microorganisms and their DNA, obtaining optimal DNA yield and quality with suitable lengths

for the selected sequence technology can be challenging.
Two commonly used approaches for extraction is Kit-based and a more manual phenol-chloroform approach, each of which has its own benefits and detriments. The aim of your study dictates which approach is best suited.

Through recent development in the field of molecular biology the kit-based approaches for DNA extraction have improved. They now give more DNA of higher quality than before. However, the phenol-chloroform approach still is superior in terms of DNA yield and quality when compared to its kit-based counterpart. Arguably, a DNA extraction primary task is to yield as much decontaminated DNA as possible and a kits' ability to purify sample depends heavily on what type sample is being analyzed, while phenol-chloroform based approaches can more easily be altered to better suit the samples being tested (Janabi, Kerkhof, McGuinness, Biddle, & McKeever, 2016).
Nevertheless, one needs also take into consideration a method ease of use, and how time consuming it is. Moreover, for the study of metagenomics the novelty of the microorganisms being studied dictates which approach if preferable. A kit-based approach could have difficulties providing sufficient amount of DNA for low-abundant species. Despite this, it should provide the same overall number of bacterial species when compared to the phenol-chloroform approach (Peng et al., 2013). Furthermore, the biggest benefit of the kit-based approaches, is their ease of use. The phenol-chloroform methods require foom-hoods, handling and disposal of hazardous substances, also fresh lysozyme solutions will have to be made for each extraction. Most kit-based approaches can easily be done from a bench-top with ordinary lab equipment and precautions, straight out of the box.

1.7 DNA assembly

One of the most important steps in a metagenomic study, especially shotgun and long-read, is the assembly of genomes after sequencing. The process of turning individual reads in to longer continuous fragments (contigs) and then merging these contigs into scaffolds that can ultimately a completed genome is a daunting task. Especially in metagenomic studies where samples contain a plethora of genomes.

Many bioinformatic tools exist for metagenomic data processing and which annotation strategy is best suited depends on the sequencing platform used, and the aim of the study. The different

methods vary in their ability to process different read-lengths. Some are best suited for long read sequencing, while others depend on the massive output and high-coverage produced by high-throughput sequencing. There are four different approaches when assembling genomes, the Naïve approach, the Greedy approach, Overlap-Layout Concecnious and De-brujin graphs.

The Navïe approach is one of the oldest, and simplest approaches for sequence assembly (Staden, 1979). It focuses on finding separate sequences with significant and enough overlap between these sequences and merging them into a longer read. However, errors in sequences, like insertions, deletions, inversion, and repeating regions makes it unreliable and when assembling entire genomes consisting of billions of base-pairs, these errors will scale logarithmic with 4^n depending on how many bases are affected by these errors. This makes it insoluble for the naïve approach to assemble genomes unless the errors and repeating regions are shorter than the reads analyzed (Simpson & Pop, 2015).

Another of the simpler approaches for DNA assembly is the greedy approach. The greedy approach involves continuously combining reads in decreasing order of quality in their overlaps. In summary, it combines the reads with the best overlap first and then adds to it with reads of lower quality until a predefined threshold is reached. If a read overlap contests an already merged read it is ignored. It is a greedy approach as the term implies, as it only involves the most logically optimal assembly for each merging of reads and discards other potential alternatives. While this approach can be inaccurate it often provides a solid approximation for the optimal assembly. However, due to its simple approach and how it assembles reads locally, it suffers when handling repeating regions and have been replaced by more complex graph algorithms that better handles repetitive sequences (Simpson & Pop, 2015).

Through the development of newer Next Generation Sequencing (NGS) platforms both the Naïve and the greedy approach have been replaced. These newer sequencing platforms provide cheaper, faster and higher-throughput sequences than their predecessors, especially platforms like the Oxford MinION and the PacBio SMRT provides exceptionally long reads. These longer reads makes it easier for bioinformatic software to detect repeated regions, insertions, deletions and inversions (Indels) that can take up large portions in a genome. The long reads can span entire open reading frames (ORF).

ORFs can be defined as sequences with a length that is divisible by three and is bound by stop codons (Sieber, Platzer, & Schuster, 2018). ORFs are important when identifying protein coding regions or functional RNA-coding regions in DNA sequences. Although these

sequencing platforms can cover ORFs, they are accompanied with their own inherited flaws. MinION and SMRT both suffer from a low sequencing depth and have a high error-rate for base-calling compared to 2nd generation sequencing platforms. These problems combined with the increase of sequencing output and read lengths and the high species complexity in metagenomic samples raises the computational requirements for assembly, making it more challenging.

Overlap-layout-consensus (OLC) is one of the algorithm approaches that can handle the data output from 3rd generation sequencing platforms. It was developed in the 1980s and was used with sanger sequencing. It functions by turning each read into a node in a graph, these nodes are structured based on their overlaps, meaning one can see how different reads are connected. It then performs a multiple sequence alignment, where the different sequences are structured based on order and overlap, and eventual inconsistencies are removed. OLC can also be modified to construct the map/graph with *k*-mers, which are subsequences of length (*k*).
K-mers is a user specified parameter that can help assembly by covering repetitive and non-unique regions in a metagenome, at the cost of coverage. Using *k*-mers can drastically reduce time spent screening for overlapping reads.
The OLC based approach is best suited to assemble small genomes, or when processing longer reads. This is because the OLC method suffers from bottlenecks especially in the overlap computing step and the vast amount of reads and sequencing output 2nd generation sequencing platform provides makes OLC a very time-consuming and computational demanding approach for high-throughput sequences (Li et al., 2011).

The last of the assembly approaches is De-bruijn graphs. It also was created in the 1980s and is widely used today. It is well suited to study large genomes and metagenomes (Simpson & Pop, 2015). In this assembly method each of the reads are broken into sequences with overlapping *k*-mers. Each of the unique *k*-mers are given a distinct node of the graph and the *k*-mers that comes from adjacent nodes are linked with an edge that indicates direction of the read. After the *k*-mers are mapped an 'Eulerian walk' is performed, which is a "walk" through graph from node to node and crosses each of the edges exactly once. The result of the Eulerian walk should correspond with the original sequence order (Pevzner, Tang, & Waterman, 2001).
However, repeating regions on the sequence can make this challenging. The algorithm will create different Eulerian walks where there is alternative pathing between the nodes. These incorrect reshuffling of the genomes in repeating regions makes it difficult for the algorithm for

select the one corresponding with the original sequence (Limasset, Cazaux, Rivals, & Peterlongo, 2016). Changing the *k*-mer length can help against this problem by covering larger portions of the sequences. However, the created contigs that are unambiguous and non-branching are reliable and provides valuable information when processing high-throughput sequencing data.

The bruijn graphs are more commonly used than OLC because of the significant computational advantage it holds. Unlike OLC, De-bruijn graphs does not require finding the overlapping pair ends of reads and because of this don't require extensive dynamic programming in order to search for said overlaps. Instead the overlaps are inferred by the nodes in the graph, this reduces processing power required. The De-bruijn graph approach can operate very quickly with the right parameters due to this. However, it struggles when finalizing genome assembly.
Because of the short read-lengths from high-throughput sequencing not covering ORFs, the repeating regions hinders its effectiveness. Therefore, it is better suited at creating several 'near-complete' genomes instead of complete genomes. Sequencing errors, like false base-calling, also proves to be difficult for De-Bruijn to process. Since it produces a node for each unique *k*-mer of *k*-length, the number of nodes and edges in the graph will increase with the amount of errors are introduce and can increase the size of the graph considerably. This adds to the already considerable amount of memory De-Brujin requires from the computational hardware and is one of the major problems this approach suffers from.

1.8 Binning & Taxonomy assignation

Binning is the process of turning post assembled contigs into genome bins and assigning taxonomy. It allows the of study individual organisms and their interactions from metagenomic samples (Sedlar, Kupkova, & Provaznik, 2017). In other words, binning is a tool that tries to identify contigs by assigning them, ideally, to a single genome (Kunath, Bremges, Weimann, McHardy, & Pope, 2017). There are currently two approaches for assigning taxonomy in metagenomic studies. You have 16S rRNA amplicon sequencing and whole metagenome shotgun (WMS) sequencing.

Like mentioned previously, 16S rRNA sequencing uses clustering to create OTUs and relies on similarities between reads to do so. And while whole metagenome shotgun approaches utilize all available DNA, 16S rRNA amplicon sequencing uses only phylogenetic marker genes. Based on these genes it screens for species abundance and richness. This approach has a plethora of comprehensive databases that contains extensive amounts of reference marker genes. This makes it easy and reliable to assign contigs based on similarities (Ribeca & Valiente, 2011). However, like previously mentioned, no additional information can be gained besides species richness and abundance from 16S rRNA data.

If one wants to study more than just taxonomy and abundance in a sample, then whole metagenome shotgun sequencing is required. Like mentioned in previous sections the WMS approach sequences all the DNA available instead of just the ribosomal RNA. However, this makes binning even more challenging. There are two ways to bin WMS sequences, the taxonomy dependent and independent approach.

The taxonomy dependent approach performs homology inferences based on online reference databases, meaning it assigns taxonomy based on similarities with already taxonomically assigned contigs. These algorithms assigns taxonomy based sequence composition (McHardy, Martín, Tsirigos, Hugenholtz, & Rigoutsos, 2007), homology (Huson, Auch, Qi, & Schuster, 2007), phylogenetic affiliation (Krause et al., 2008), or a combination of these approaches (MacDonald, Parks, & Beiko, 2012).

Nevertheless, due to the small amount of completed/near-completed reference genomes in the databases, assigning taxonomy can be challenging (Teeling & Glöckner, 2012), especially if you have poorly assembled genomes with a lot of unknown regions. There are currently a multitude of different tools that align sequences and compare then using reference databases for various types of gene material, either it being DNA, RNA viral RNA or proteins. Tools such as MEGAN (Huson et al., 2016), SOrt-ITEMS (Monzoorul Haque, Ghosh, Komanduri, & Mande, 2009) that both operate with read, and Phylopythia (McHardy et al., 2007) that operates with k-mers, are just a few that can assign taxonomy based on aligning sequences to reference databases such as NCBIs' BLAST.

The other approach of WMS binning, often referred to as the 'unsupervised approach' have had a lot development in the recent years. These algorithms use intrinsic information present in samples, like GC-percentage, codon usage and oligonucleotide usage patterns to cluster the reads, meaning grouping data based om similarities, and assign taxonomy based on these clusters (Mande, Mohammed, & Ghosh, 2012). Tools such as Metabin2.0 (Liu, Hou, & Fu,

2015) can assign taxonomy based on k-mer frequencies and TETRA (Teeling, Waldmann, Lombardot, Bauer, & Glöckner, 2004) that clusters based on computed correlations between nucleotide usage pattern between reads are both effective at assigning taxonomy for reference free reads. However, in metagenomic samples there are imbalances in read coverage which can make it a computational challenge compared to the taxonomic dependent approach (Imelfort et al., 2014). Which of these approaches is best suited depends entirely of the novelty of species found in your samples.

1.9 Gene Calling & Functional Annotation

After assigning taxonomy to contigs the process of 'gene calling' can begin. Gene calling revolves around identifying RNA and protein coding regions in the (meta)genomes. It can be performed on both contigs and raw reads from long read sequencing platforms. There are two approaches for gene calling, "Sequence similarity-based" and "Ab Initio" and similarly to binning, the optimal one depends on the novelty of contigs you are analyzing (Kunath et al., 2017). The Sequence similarity-based approach relies on well-developed reference databases and searches for homology between the sample-genes and the database-genes. It provides highly accurate results and can predict functions of processed genes, given it can find matches in the databases.

The "Ab Initio" approach is for the analysis of novel genes with no references in databases. It systematically searches sequences for certain 'signs' that indicate coding regions. These signs are based on either 'signals' or 'content' of the sequences. For prokaryotes, many of the promotor sequences are known to us, making them easy to identify. By analyzing codon frequencies and genome nucleotide composition, ab initio algorithms could differentiate between coding and non-coding regions (Zhu, Lomsadze, & Borodovsky, 2010). Examples of tools that can be used for gene prediction of metagenomes are GeneMark.hmm (Lukashin & Borodovsky, 1998) which operates with a 'hidden Markov framework' and uses ribosomal binding patterns to predict translation initiation codons, and Prodigal (Hyatt et al., 2010) which has a 'trial and error' approach and operates with a self-learning algorithm to differentiate between coding and non-coding regions.

Although powerful, these tools still make mistakes, especially in metagenomic studies that exclusively rely on short-read sequences. The short reads provided by 2nd generation

sequencing platforms often results in fragmented and incomplete genes due to them not being able to cover ORFs. Furthermore, the short contigs makes it hard to identify homologous and will result in a poor identification of novel genes (Kunath et al., 2017). However, these problems can be negated by using a combination of long and short reads when assembling the genomes (M. Kamada et al., 2014)(Price, Hayer, Depledge, Wilson, & Weitzman, 2019). Using short reads to polish the long reads allows for longer ORFs to be examined, which makes it easier to discover the coding regions.

1.10 Pathway Annotation

After contig taxonomy have been assigned and coding regions have been identified, the remaining step is to determine the predicted function of these genes and what pathways they are involved in.

Pathway annotation revolves around comparing the predicted ORFs with already annotated sequences from functional databases. The goal is to produce accurate annotations based on the comparisons and correctly identify orthologous genes, to which we already know the function. There are several approaches for this and a multitude of reference databases to select from. However, what pathways one aims to study dictates what database is best suited.

For our study we used KEGG (Kyoto Encyclopedia for Genes and Genomes) to reconstruct the pathways and annotate gene function. In addition, we utilized a specialized database to identify carbohydrate active enzymes (CAZymes), referred to as the CAZy database (CAZyDB). CAZyDB is a specialized database with detailed information on carbohydrate active enzymes. It analyzes and displays the genomic structural and biochemical information of these enzymes and contains more than 300 families to which to analyze for sequence similarities from (Kunath et al., 2017).

1.11 Aim of Study

In this study we originally planned to explore the metagenomic composition and function of sheep rumen. However, due to time constraints (COVID19), alterations had to be made. Instead, we examine different aspects of a metagenomic study and their strengths & weaknesses. We

perform the different steps involved in a metagenomic study, like DNA-extraction, sequencing, binning and annotation, but on different sample types.

The objectives of our altered study were to obtain an overview of the taxonomy of our samples, understand the function of some of the microbes in these samples and understand how third-generation sequencing can be used to fulfill these roles.

2 Materials

2.1 Lab equipment

2.1.1 Specific lab equipment

PowerPac™ Basic Power Supply	BioRad	1645050
Gel Doc™ EZ System	BioRad	1708270EDU
UV Sample Tray	BioRad	1708271EDU
P95 DW	Mitsubishi	-
KP95HG	Mitsubishi	-
Telstar AV-100	TELSTAR TECHNOLOGIES, S.L.	-
Heraeus Multifuge X1 Centrifuge	Thermo Scientific™	75004210
913 pH Meter, laboratory version	Metrohm Nordic AS	2.913.0210
Labcycler Gradient, Thermoblock 96, silver	SensoQuest	012-103
Mastercycler® Gradient	Eppendorf®	6311 000.010(?)
Qubit dsDNA BR Assay Kit	Invitrogen	Q32853
Qubit Assay tubes, set of 500	Invitrogen	Q32856
Qubit™ 1 Fluorometer	Invitrogen	Q32857
50x TAR Electrophoresis	Thermo Scientific	B49
PeQGreen DNA/RNA	PeQlab	37-5000
Iproof HF MasterMix	BioRad	1725310
AMPure XP	Beckman Coulter	A63881

Quick-load, Purple 1 kb DNA ladder	New England Biolabs	N0552s
Gel loading dye blue(6x)	New England Biolabs	B7021S
Seakem LE Agarose	Lonza	50004
Surebeads Magnetic Rack	BioRad	-
TruSeq Index Plate Fixture	illumina	15028344
Adhesive Sealing Sheets	Thermo Scientific	AB-0558
Centrifuge 5418R	Eppendorf®	-
Galaxy 14D	VWR	-
Refrigerator	BOSCH	-
Freezer	BOSCH	-
Ultra-Low Temperature Freezer C585 Innova	New Brunswick	-
FastPrep-24 TM	MP Biomedicals	-
Thermomixer C + (1,5 mL block)	Eppendorf®	-
NanoDrop Spectrophotometer ND-1000	Saveen Werner	-
Sartotius Quintex 124-1S	VWR	-
GS	Kern	-
AV-100	Tellstar	13472
MS2 minisloaker	IKA	-
Galaxy14D	VWR	-
RCT Classic	IKA	-
Quintix 124-1s	Sartorious	-
Mini Sub cell GT	BioRad	-
Tisch-autoclav	Certoclav	-

2.1.2 General lab equipment

Pasture pipette 5 mL non-sterile graduated up to 1 mL	VWR	612-1684
Biospehere filter tips 0.1-20µ	VWR	70.1114.210
Biospehere filter tips 2.0-20 µl	VWR	70.760.213
Biospehere filter tips 20-300 µl	VWR	70.765.210
Biospehere filter tips 200 µl	VWR	70.1189.215
Biospehere filter tips 1250 µl	VWR	70.1186.210
ART™ Barrier Hinged Rack Pipette Tips	Thermo Scientific™	2139-HR
Finntip™ Pipette Specific Pipette Tips, 10mL	Thermo Scientific™	9400303
Ultra fine pipette tip 0.1-10 µl	VWR	613-0364
Ultra fine pipette tip 1.0-250 µl	VWR	613-0362
Ultra fine, FlexTop, extended pipette tip 100-1250 µl	VWR	613-0272
Axygen® 1.5 mL MaxyClear Snaplock Microcentrifuge Tube	Axygen	MCT-150-C
Axygen® 0.2 mL Thin Wall PCR Tubes with Flat Cap	Axygen	PCR-02-A
Axygen® 2.0 mL MaxyClear Snaplock Microcentrifuge Tube	Axygen	MCT-200-C
Finnpipette F1, 8 channels, 0.5-10 µl	Thermo Scientific™	OH68580
Finnpipette F1, 8 channels, 5-50 µl	Thermo Scientific™	OH69611
Finnpipette F1, 8 channels, 30-300 µl	Thermo Scientific™	PH78657
Finnpipette F1, single channel, 0.5-10 µl	Thermo Scientific™	NH70705

Finnpipette F1, single channel, 5-50 µl	Thermo Scientific™	CH50877
Finnpipette F1, single channel, 30-200 µl	Thermo Scientific™	CH20500
Finnpipette F1, single channel, 100-1000 µl	Thermo Scientific™	LH37761
Finnpipette F1, single channel, 1.0-5.0 mL	Thermo Scientific™	LH47208
Finnpipette F1, single channel, 2-10 mL	Thermo Scientific™	T23916

2.2 Chemicals, manufactured reagents and kits

DNeasy PowerLyzer PowerSoil Kit	QIAGEN	12855-100
Iproof HF MasterMix	BioRad	1725310
AMPure XP	Beckman Coulter	A63881
Nextera XT Index Kit	illumina	15055294
PhiX control v3	illumina	15017666
AMPure XP	Beckman Coulter	A63881
Emsure Methanol	Merck	-
Emsure chlroroform	Merck	-
phenol:chloro	Sigma-Aldrich	P2069
NEBnext FFPE DNA Repair Mix	New England Biolabs	M6630S
Blunt/TA Ligase Master Mix	New England Biolabs	M0367S
NEBnext Ultra II End-Repair/dA-tailing module	New England Biolabs	E7546S
Flow cell wash kit	Oxford nanopore	EXP-WSH003
Flow cell priming kit	Oxford nanopore	EXP-FLP002
Ligation sequencing kit	Oxford nanopore	SQK-LSK109
Primere	Eurofins genomics	-
PeQGreen DNA/RNA	PeQlab	37-5000

50x TAR Electrophoresis	Thermo Scientific	B49
Gel loading dye blue(6x)	New England Biolabs	B7021S
Seakem LE Agarose	Lonza	50004

2.3 Buffers

We only needed to mix two buffers ourselves, since the rest came complete from the suppliers. The buffers we created was 10 mM Tris pH 8.5 and 0.2 M NaOH

Our 10mM Tris pH 8.5 was comprised of the following:

Tris Buffer, 0.1M Solution, Ph 7.4 500Ml	aMRESCO	A611-E553-10
Nuclease-free water	Thermofisher scientific	AM9920
Hydrochloric Acid, Concentrated	VWR Chemicals	470301-260

Our 0.2 M NaOH was comprised of the following:

Sodium chloride 5 M in aqueous solution, autoclaved	VWR Chemicals	7647-14-5
miliQwater (created in lab)	-	-

2.4 Software tools

Name	Function	Supplier	Reference
Rcommander, DADA2 pipeline	Assembly, binning and illustration of 16S rRNA taxonomy	Benjamin J Callahan et.al	(Benjamin J. Callahan et al., 2016)
GhostKoala	Annotation of MAGs	Minoru Kanehisa et.al	(M. Kanehisa, Sato, & Morishima, 2016)
EPI2ME	Sequence alignment and binning	Oxford Nanopore technologies	https://nanoporetech.com/nanopore-sequencing-data-analysis

MetaGeneMark	Gene annotation	John Besemer, Alexandre Lomsadze and Mark Borodovsky	(Besemer, Lomsadze, & Borodovsky, 2001)

3 Methods

Due to severe time limitations as a result of problems with DNA extraction, PCR amplification and the Covid-19 outbreak, alterations to our study were made. While we originally set out to perform 16S rRNA, metagenomic shotgun and minION sequencing on rumen samples collected from sheep fed with seaweed, we instead performed 16S rRNA like planned, but switched our shotgun metagenomic analysis to samples from a human gut derived enrichment culture. This analysis included generating long-read shotgun data using Oxford Nanopore technology and analysis of previously constructed metagenome-assembled genomes via short-read Illumina technology (Ostrowski et al., 2020).

3.1 Sampling

3.1.1 16S rRNA samples

The 16S rRNA samples were collected from lamb that was fed with different levels of seaweed (*Saccharina latissima*/Sugar kelp) over a period of 30 days. Temporal rumen samples were collected by esophageal tubing throughout this period; however, we only analyzed the last samples (taken at the slaughterhouse). The different feeding groups have 8 biological replicates and are divided into A (0% seaweed,) B (5% seaweed) and C (2.5% seaweed). Each of the feed groups contain four samples that contain the fluid phase and four that consists of particle phase. The samples were then immediately frozen and stored at -80°C.

3.1.2 Metagenomic DNA and shotgun data

As mentioned above, within in this thesis we analyzed MAGs from a previously generated metagenome (Ostrowski et al., 2020), and provide a summary of the methods used hereafter. Fecal samples from 80 healthy 18-20-year-old adults were collected and immediately placed in an anaerobic jar (2.5 L AnaeroJar; Oxoid) equipped with a gas-generating kit (AnaeroGen; Oxoid). Samples were diluted at 10% (wt/wt) in phosphate-buffered saline (PBS) (0.1 M, pH

7.4) and a 100 µl aliquot was grown in Defined Medium (DM) supplemented with 10 mg/mL xanthan gum (XG, Sigma). Samples that showed growth on xanthan gum, as evidenced by loss of viscosity and increased culture density (20 samples), were sub-cultured 10 times by diluting an active culture 1:100 into fresh DM-XG medium. Multiple samples harvested at different time points were stored for gDNA extraction. Samples (44 in total) from 2 mL cultures were harvested by centrifugation and stored a -20 C until further use.

A phenol:chloroform:isoamylalcohol and chloroform extraction method was used to extract high molecular weight DNA as previously described (Pope et al., 2011). The DNA was quantified using a Qubit™ fluorimeter and the Quant-iT™ dsDNA BR Assay Kit (Invitrogen, USA), and the quality was assessed with a NanoDrop One (Thermo Fisher Scientific, USA).

A total of 44 samples were subjected to metagenomic shotgun sequencing using a combination of Illumina HiSeq 3000 and Illumina HiSeq X platforms (Illumina, Inc.) at the Norwegian Sequencing Center (NSC, Oslo, Norway). Samples were prepared with the TrueSeq DNA PCR-free preparation and sequenced with paired ends (2 × 150 bp) on two lanes. Quality trimming of the raw reads was performed using Cutadapt (Martin, 2011), removing all bases on the 3′-end with a Phred score lower than 20 and excluding all reads shorter than 100 nucleotides, followed by a quality filtering using the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/).

Reads with a minimum Phred score of 30 over 90% of the read length were retained. Remaining reads were co-assembled using metaSPAdes v3.10.1 with default parameters and k-mer sizes of 21, 33, 55, 77 and 99 (Nurk, Meleshko, Korobeynikov, & Pevzner, 2017). The resulting contigs were binned with MetaBAT v0.26.3 in “very sensitive mode” (Kang, Froula, Egan, & Wang, 2015). The quality (completeness, contamination, and strain heterogeneity) of the MAGs was assessed by CheckM v1.0.7 with default parameters (Parks, Imelfort, Skennerton, Hugenholtz, & Tyson, 2015). Open reading frames were annotated using PROKKA v1.14.0 (Seemann, 2014).

3.2 Cell lysis and DNA extraction

3.2.1 Bead beating cell lysis and DNA extraction

The sheep rumen contains a variety of microorganisms, Gram-positive, gram-negative bacterial cells, fungal, archaeal and protozoal cells all inhabit the rumen. Due to this variety of organisms to examine we chose bead beating to extract the DNA from both the liquid and the solid samples. Bead beating is a harsh and mechanical way to disrupt the cell membrane in order to acquire the DNA the beads used were 0.1mm glass. Before the cells were lysed the samples were thawed on ice (around +4°C) and vortexed to homogenize.

For the DNA extraction we utilized the "DNease, Powerlyzer, Powersoil Kit from QIAGEN and proceeded in accordance with the protocol accompanying the kit. The kit was chosen over the traditional Phenol/chloroform approach due to its ability to effectively extract DNA from multiple samples at the same time, while the Phenol/Chloroform technique require more labor per sample, so to save time while still obtaining adequate amounts of DNA this approach was chosen.

We transferred roughly 0.25g of sample material into the provided Powerbead Tube added 750 µl Powerbead Solution, 60 µl C1 solution and used a Powerlyzer 24 homogenizer to lysate the cells. We decided to use the Powerlyzer instead of vortexing by hand to save time, and to ensure equal amount of stress was put on each of our samples. The machine ran on 4,000 RPM for 45 seconds after which the cells were lysed through the shaking process, and the intramitochondrial DNA was supposedly released into the solution, the sample was centrifuged at 9,900 RCF to form a pellet, the fresh supernatant was transferred in to a clean 2 mL collection tube (provided in kit).

While the protocol expects 400-500 µl supernatant we often got more (around 600), and always extracted as much as possible. After adding 250 µl of C2 solution (provided in kit) and briefly vortexing the samples were incubated for 5 minutes in a fridge with +5°C. After which the samples were centrifuged for 1 minute at 9,900 RCF and roughly 750 µl supernatant were extracted to a new clean 2 mL tube (provided in kit). A 1200 µl aliquot of C4 solution (provided in kit) was then added to the samples, and then vortexed for 5 sec.
A total of 675 µl supernatant was then transferred to the MB Spin Column (provided in kit) and then centrifuged at 9,900 RCF for 1 minute, with the flow through discarded. This process was repeated until all the supernatant was used. A 500 µl aliquot of C5 solution (provided in kit)

was added and centrifuged for 30 seconds at 9,900 RCF. The flow through was discarded, and the tube was then centrifuged again at 9,900 RCF for 1 minute to dry out the MB Spin Column for flow through. The filter contained in the MB Spin Column was transferred to a clean 2 mL collection tube and 100 µl of C6 solution (provided in kit) was added on top of the filter membrane. The tube was then centrifuged at 9,900 RCF for 30 seconds before the filter was removed and we were left with pure DNA in the bottom of the 2 mL collection tube.

All the centrifuging was done in room temp (around 20°C) in accordance with the protocol provided by the kit. The kit used were customized towards lysing cells found in processed soil, fecal, water, food, insects, swabs with PCR inhibitors.

3.2.2 Measuring DNA concentration

DNA-concentrations were measured after DNA extractions as well as before and after the PCR cleanup process start, using Qubit machines for quantification and by validating on agarose gel.

The Qubit measures the nucleic acids ability to absorb ultraviolet radiation with the wavelength of 260 nm, the more DNA you have the more of the radiation will be absorbed. Nucleic acids cannot absorb any UV radiation consisting of wavelengths longer than 260, therefore the Qubit machine also measures absorbance on 280 nm wavelength to detect any foreign particles in the sample. By comparing these 2 values it can provide indications on the purity of your samples. If the 260/280 value precedes 1.7, the sample can be considered “pure” from contamination.

However, quantifying DNA through spectrophotometry can be unreliable. The machine is not able to distinguish between DNA, RNA and proteins, free nucleotides and other particles will also affect the purity score.
By using fluorescent dyes, the downsides of spectrophotometer quantification can be reduced (Haque et al., 2003). Qubit machines require the use of specific coloring molecules that binds to the particles being examined in order to distinguish between them, whether it is DNA, RNA or proteins.
We mixed 1 µl DNAdye 199 µl buffer, from the dsDNA HS assay kit from Thermo Fischer Scientific, for each sample. The 198 µl buffer/dye dilution was mixed with 2 µl DNA from our samples and vortexed to homogenize. We then used a Qubit machine to measure absorption.

We used a electrophorese gel to validate the length of our DNA. For our samples we tested out different gels and run times depending on which stage of the library preparation we were on. After DNA extraction we used a 1% agarose gel and ran it on 70 volts for 40 min, while post PCR DNA ran at 1.5% gel at 90 volts for 40 min. We created the gels using electrophoresis grade agarose from VWR and concentrated TAE-buffer from Thermo Scientific that we diluted in distilled water.

3.3 16S rRNA gene amplicon analysis

The 16S ribosomal RNA is one of the most used sequences for determining phylogeny. The main benefits of sequencing this region is its presence in all bacteria and archaea, also the slow mutation rate providing more accurate measures of time when used as a molecular clock. In addition, different regions within the 16S rRNA gene contain hypervariable sections resulting in ease of differentiating and the large size of the region makes it usable for informatic purposes.

We chose 16S rRNA gene sequencing to examine taxonomy of our samples, the complex composition of microorganisms found in rumens makes it nearly impossible to utilize other techniques like DNA-DNA hybridization. While DNA-DNA hybridization is reliable when examining small samples containing fewer bacteria, it falls short when utilized on samples containing massive amounts of different microorganisms, it would be too costly and time-consuming to utilize(J.-C. Cho & Tiedje, 2001). While 16S rRNA gene data often is too conserved to determine taxonomy on species and subspecies level, it is better suited than DNA-DNA hybridization when determining microbiological compositions in a more general manner, which was our focus for this paper.

When sequencing the 16S rRNA gene we looked at the v3-v4 region. The combination of V3 and V4 regions strikes a nice balance with their hypervariable regions and how conservative they are. The V4 region contains few hypervariable regions and can reliably identify to the phylum level as well as the whole 16S rRNA gene (Yang, Wang, & Qian, 2016). V3 contains more hypervariable region which allows for better identification down to genus level. The combination of these regions allows us to analyze for new and unique bacteria while retaining the ability to identify well conserved species. Although other regions, like V2-V3 can give more accurate depictions of species richness (Yang et al., 2016) we opted to determine taxonomy based on the V3-V4 regions due to availability .

All the amplicon sequencing was done in accordance with the protocol provided by Illumina. https://support.illumina.com/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf

3.3.1 PCR Amplification

The function of the PCR Amplification step is to amplify the desired DNA template, using primers designed for specific regions and overhang adapters to induce ligation. We used 341F and 805R primers for our amplicon PCR, designed for the V3-V4 region creating strands of 428 basepairs.

The PCR Amplification was done under sterile conditions in a sterile cabinet using sterilized equipment and containers. We created a mastermix(MM) containing 1 µl DSMO, 1 µl forward primer, 1 µl reverse primer, 12.5 µl polymerase and 7 µl H2O per sample in a 1.5mL Eppendorf tube. In individually marked PCR-tubes we mixed 2.5 µl DNA (diluted to 10 ng/µl) from each sample with 22.5 µl MM to a sum total of 25 µl. The PCR tubes were placed in a Sensoquest Labcycler and ran on the following program (98°C for 3 min, 25 cycles of (95°C for 30 sec, 53°C for 30 sec and 72°C for 30 sec) 72°C for 5 min, hold at 4°C. The PCR product ran on a 1.5% agarose gel at 70v for 40 minutes to check for contamination and general quality.

3.3.2 PCR Clean-up 1

The PCR Clean-up step uses AMPure XP beads to purify the PCR product from primer and primer dimer species, leaving purer strands of the V3-V4 regions desired.

AMPure XP beads were vortexed for 30 seconds to evenly disperse the beads and 20 µl was transferred to each well in the Amplicon PCR plate using a multipippette. The beads and PCR product were mixed up and down 10 times using the multipipette. The plate was sealed and shaken at 1800 RPM for 2 minutes using a MIDI plate. After shaking the plate was placed at room temperature and incubated for 5 min. The plate was placed on a magnetic stand for 2 min, so the beads could separate from the supernatant.

The supernatant was discarded after separation from the beads, by using a mulitpiptette. A 200 µl aliquot of fresh 80% ethanol was added to each well, using a multipipette to wash the beads. The plate was incubated for 30 seconds while still on the rack before the supernatant was carefully discarded. This step was repeated once. After supernatant was removed the plate was

placed to airdry for 10 minutes, after which the plate was removed from the magnetic stand and 52.5 µl 10mM Tris, with a pH of 8.5 was added to each well. The beads and the Tris were gently mixed by pipetting, before being incubated at room temperature for 2 minutes. The plate was placed back on the magnetic stand for the beads to gather and supernatant to clear. After which 50 µl of the supernatant was transferred to a clean PCR-plate for further use.

3.3.3 Index PCR

Index PCR consists of different pairs of index primers added in unique pairs to individual samples. This allows us to mix samples together and sequence them as such. After sequencing the Illumina MiSeq, different steps are taken to recognize the reads and which sample they originate from due to their unique indexes and can sort them accordingly. This makes sequencing a lot more time and cost efficient, since pooling of the different samples negates the problem of having to run individual runs for each sample.

We used a multipipette to transfer 5 µl of the previous PCR product to a new clean 96-well plate. The various primers were arranged on a TruSeq Index Plate Fixture Index primer 2(white caps) arranged from A-H on the plate, while the index primer 1(orange caps) was arranged from 1-6. A 2.5 µl aliquot of each primer type, 12.5 µl HiFi HotStart ReadyMix and 5 µl PCR grade water was mixed and added to each well and mixed gently by pipetting up and down. The plate was centrifuged at 1000 x g at room temp for 1 minute. The PCR plate was then placed in a Sensoquest Labcycler where we used the same program as the Amplicon PCR, consisting of 95°C for 3 min, 8 x (95°C, 30 sec. 53°C, 30 sec. 72°C, 30 sec.) 72°C for 5 minutes and hold at 4°C after which it was frozen at -20°C.

3.3.4 PCR Clean-up 2

PCR clean-up 2 is the final cleaning step before the DNA be quantified, normalized and sequenced.

After defrosting the PCR product, the AMPure XP beads were vortexed for 30 seconds to homogenize and 28 µl was added to each well in our index PCR plate containing the PCR product and mixed gently with pipetting. The PCR plate was then incubated at room temperature for 5 minutes before it was placed on a magnetic rack till the supernatant cleared.

The clear supernatant was discarded and 200 µl fresh 80% ethanol was added to the beads and was incubated for 30 seconds at room temperature, before the clear supernatant again was discarded. The ethanol wash was repeated once. After the ethanol was discarded following the second wash, the beads were air-dried for 10 min. After the beads were dried the well was placed off the magnetic rack and 27.5 µl 10mM Tris with a pH of 8.5 was added to each well and gently pipetted up and down to mix before incubating for 2 min. After which the plate was again placed on a magnetic rack till the supernatant cleared and all the supernatant was transferred to a new 96-well PCR plate by pipetting.

3.3.5 Troubleshooting

When initially running the PCR amplification we had major difficulties getting it to work. We ran it according to protocol, but it barely amplified our DNA. While it can be expected that the amount of DNA will increase logarithmically with the amount of cycles you run, we barely saw a doubling of the DNA, even after 25 cycles.

We tried adding different solutions to enhance the amplification and increase yield, and we tried different dilutions of the DNA, in case contaminants were blocking for the primers or deactivating the polymerase. We ran two sets of amplification PCR to test different PCR enhancers, the first with 1 µl DNA and the second with 2.5 µl DNA. In each of the sets we added different volumes of PCR amplifying solutions, the first with 1 µl and the second with 2 µl, totaling up to 4 different tests for each solution.

DNA/solution volume	BSA	MgCl2	DSMO
1 µl DNA	1 µl	1 µl	1 µl
	2 µl	2 µl	2 µl
2.5 µl DNA	1 µl	1 µl	1 µl
	2 µl	2 µl	2 µl

Tabel 1.1 Showing the different solvents and their concentration for improving the PCR.

Figure 2.1 Showing from left to right in increasing order what results we from the different amplifying solvents. The first four wells contain BSA, the next four contain DMSO and the last four contain MgCl2

Out of the different solutions we added we had best success with the DMSO. DMSO makes the GC rich regions more heat-labile and reduces the melting temperature for the reaction. It directly binds to the Cytosine residue in the GC rich regions and changes its conformation and reduces the strength of the triple-hydrogen bond, which is what makes the DNA less accessible.

This increased our amplification and gave us notably more DNA, but still less than expected. We believe the reason as to why we struggled with amplification is because we chose the kit-based approach for DNA extraction. Although the kit was designed to handle soil samples, which in theory should be harder to handle than rumen samples, some contaminants probably got through.

We noticed a time-gradient between amount of amplification and time spent prepping samples, since once we reduced the number of samples handled at once and focused on working rapidly with them, we saw a much larger yield through amplification than previously (see figure 2.2). We speculate that the contaminants might have influenced our polymerase, decreasing its efficiency.

While we still managed to extract enough DNA for 16S rRNA gene analysis, some downstream biases might have arisen due to these complications.

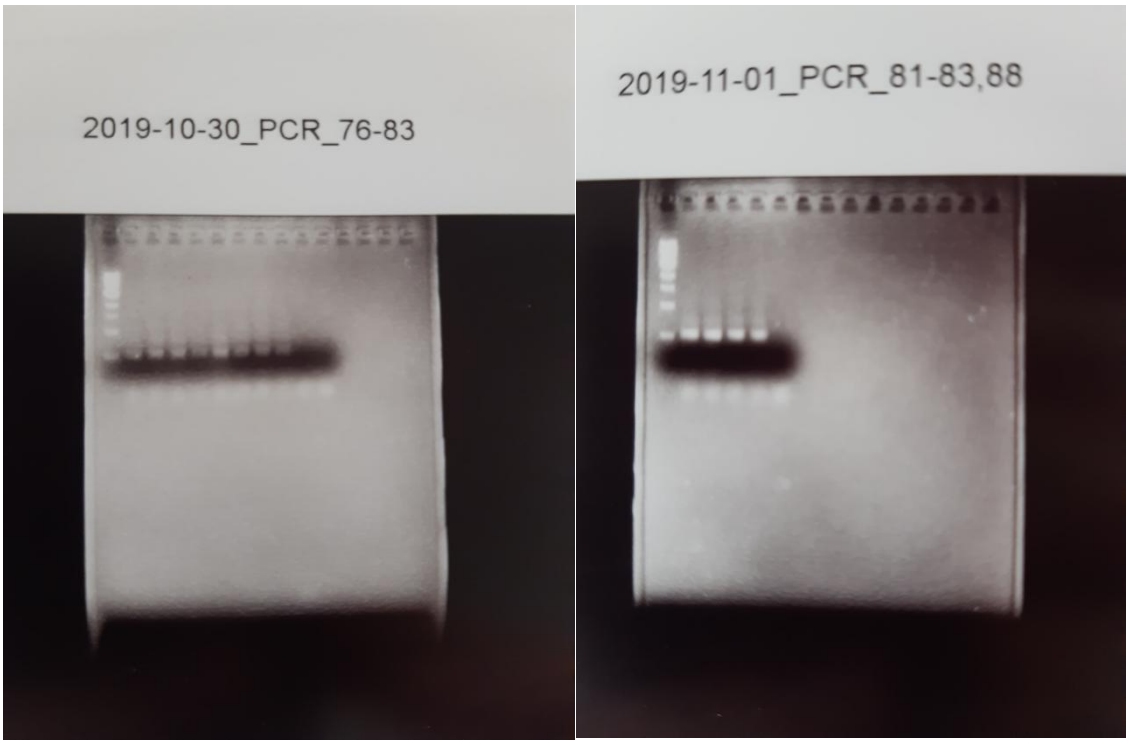


Figure 2.2 PCR results after reducing the number of samples handled at once. The four outmost wells on the left picture are the same samples in the second picture and shows a considerable increase in band-strength.

3.4 Library preparation & Sequencing

3.4.1 16S rRNA gene sequencing

Before 16S rRNA gene sequencing took place, the DNA samples were quantified. The formula for quantifying and the quantified measurements can be found in Appendix 1.

After 16S rRNA gene amplification, PCR products were sequenced on an Illumina MiSeq. A heatblock, suited for 1.7 ml microcentrifuge was heated up to 96°C. The MiSeq reagent cartridge was removed from the freezer and thawed at room temperature. For denaturing our DNA, 4nM pooled library (5 µl) and 0.2 N NaOH (5 µl) was mixed in a microcentrifuge tube. The samples were then vortexed to mix and centrifuged at 280 x g at 20°C for 1 minute. The samples were then incubated at room temperature, to allow the DNA to denature into single strands. A 990 µl aliquot of pre-chilled hybridization buffed (HT1 from Illumina kit) was added to the tube containing the denatured DNA (10 µl). The samples were then kept on ice between dilutions. We desired a final concentration of 6pM, so 180 µl of our 20pM denatured library

was mixed with 420 µl pre-chilled HT1 solution. The tubes were then inverted several times and pulse centrifuged, before being placed back on ice.

The next step was denaturing and diluting the PhiX Control to the same loading concentration of our Amplicon library. A 2 µl aliquot of the 10nM PhiX library and 10 mM Tris pH 8.5 was mixed in a sterile 1.5mL Eppendorf tube. After diluting the PhiX library with 10 mM Tris pH 8.5 until a 6 nM concentration was achieved, 5 µl of 6 nM PhiX library and 5 µl of 0.2 N NaOH was mixed and vortexed briefly. After a 5-minute incubation at room temperature to allow for denaturation, 990 µl of pre-chilled HT1 solution and 10 µl denatured PhiX library was combined in a sterile 1.5mL Eppendorf tube.
The PhiX library was then diluted with 180 µl denatured library in 420 µl pre-chilled HT1 solution inside a sterile 1.5mL Eppendorf tube. The tube was then inverted, and pulse centrifuged to mix. A 20 µl aliquot of the PhiX library was then mixed with 480 µl amplicon library. The combined sample was heated in a heat block at 96°C for 2 minutes and loaded onto the MiSeq v3 reagent cartridge which was inserted into the MiSeq for sequencing.

3.4.2 MinION sequencing

MinION sequencing preparations were made in accordance with Oxford Nanopores "Genomic DNA by Ligation (SQK-LSK109)" protocol.

While there are possibilities to fragment the DNA in your samples, which increases the durability of the flowcells we chose to omit this, because we were interested in obtaining as long reads as possible.

3.4.2.1 DNA repair and end-prep

The DNA and the reagents were thawed slowly on ice. After running a Qubit we normalized the two samples, we transferred 6.4 µl(1µg DNA) of sample XDC.orginial and 17.6 µl(1µg DNA) of XDC.03 to separate Eppendorf DNA LoBind tubes and added nuclease free water till the total volume reaches 49 µl and flicked the tubes to homogenize. A 47 µl aliquot of this DNA solution was transferred to a new Eppendorf DNA LoBind tube and the following reagents was added, 3.5 µl NEEBNext FFPR DNA Repair Buffer, 2 µl NEEBNext FFPE DNA Repair Mix, 3.5 µl Ultra 2 End-prep reaction buffer and 3 µl Ultra 2 End.prep enzyme mix. The mixture was

flicked gently, spun down and placed on a thermal cycler and incubated at 20°C for 5 min, followed by 65°C for 5 min.

The AMPure XP beads was vortexed to homogenize and added to the mixture and flicked before it was placed on a holamixer(rotator mixer) for 5 minutes at 11 RPM in room temperature. Afterwards the Eppendorf DNA LoBind tube was spun down and placed on a magnetic rack to pellet the beads. After the solution was clear the supernatant was removed with a pipette. The beads were washed with 200 µl freshly made 70% ethanol, without disrupting the pellet and the supernatant was removed again, before another 200 µl fresh ethanol was added and again removed.

The Eppendorf DNA LoBind tube was spun down and placed again on the magnetic rack where any ethanol residues was removed, and the pellet was dried for 30seconds. The Eppendorf DNA LoBind tube was removed from the magnetic rack and 61 µl of nuclease free water was added before the sample was incubated at room temperature for 2 minutes. The Eppendorf DNA LoBind tube was again placed on the magnetic rack and once the eluate was clear it was transferred to a clean 1.5 ml Eppendorf DNA Eppendorf DNA LoBind tube.

We ran another Qubit on the samples to validate DNA concentration. The XDC03 samples had almost twice the amount of DNA than XDC original, because of this we diluted it till both had roughly 1000 ng DNA in each sample.

3.4.2.2 Adapter ligation and clean-up

Adapter ligation is the process of attaching oligonucleotides to the DNA fragments. These Oligonucleotides can perform various functions and are a critical step in library preparation. For nanopore sequencing the oligonucleotides consists of motor proteins that attach to the 3’ and 5’ ends of the DNA fragments and allows them to attach with the nanopores. Furthermore, these adapters increases the probability for the complementary strand to immediately follow the template strand during sequencing (Gilpatrick et al., 2020; Schalamun et al., 2019)

The reagents were thawed at room temperature and placed on ice. In a DNA Eppendorf DNA LoBind tube 60 µl DNA sample from the previous step, 25 µl Ligation buffer (LNB), 10 µl NEBNext Quick T4 DNA Ligase and 5 µl Adapter Mix (AMX) was mixed and flicked gently to combine, before it was spun down and incubated for 10 minutes at room temperature. The AMPure XP beads were vortexed to homogenize before 40 µl was added to the reaction and

mixed by flicking and the reaction was placed on a Hula mixer (rotator mixer) for 5 minutes with 11RPM at room temperature.

The LoBind tube was spun down and placed on a magnetic rack, after the solution was clear, the supernatant was removed. The beads were then washed with 250 µl Long Fragment Buffer (LFB) and the tube was flicked to resuspend the pellet and then placed back on the magnetic rack. The supernatant was removed and discarded after the pellet reformed. The previous step was repeated once. The LoBind tube was spun and placed back on a magnetic rack where any residual supernatant was removed, and the pellet airdried for 30 seconds. After which the tube was removed from the magnetic rack and 15 µl Elution buffer (EB) was used to resuspend the pellet.
After a 10-minute incubation at 37°C the tube was placed back on a magnetic rack for a pellet to form. The clear eluate was the transferred into a clean 1.5ml Eppendorf DNA LoBind tube.

We used a Qubit to quantify the samples and it seemed the previous dilution had worked since both were close to the desired DNA concentration range of 700ng.

3.4.2.3 Priming and loading the Flow Cell

All the reagents were thawed at room temperature and placed on ice. After the priming port on the flowcell was opened air was removed using a pipette. A 30 µl aliquot of Flush Tether (FLT) was added directly to the Flush Buffer (FB) tube and mixed by vortexing. An 800 µl aliquot of this priming mix was then added to the priming port followed by 5 minutes of waiting.

Meanwhile, in a separate tube 37.5 µl Sequencing Buffer (SQB), 25.5 µl Loading Beads (LB) that was mixed before use, and 12 µl of our DNA library was added and gently mixed by flicking. A 200 µl aliquot of the priming mix was then loaded to the flow cell via the priming port, the SpotON sample port cover was removed, and 75 µl of our sample was added via the SpotON sample port in a dropwise fashion. The SpotOn sample port cover was then replaced and the priming port was sealed. The sequencing was then performed on a minION device, connected to a laptop using the MinKNOW software for basecalling and analysis.

3.5 Bioinformatic processing

3.5.1 16S rRNA gene amplicon analysis

The 16S rRNA Amplicon data was analyzed using the DADA2 pipeline(Benjamin J Callahan et al., 2015). We trimmed our reads to 275 on forward length and 235 on reverse length and trimmed off the edges at 17bp on the forward reads and 21 on the reverse to make sure the index primers were excluded in the downstream analysis. We tried altering the maximum number of "expected errors" in order to optimize our pipeline but found that the default setting of MaxEE=c (2,2) provided the best results. After accounting for abundancies of our chimeric variants we found that 15.55% of our reads were chimeras.

OTUs were created with the standard settings and taxonomy was assigned using a "Naive Bayesian Classifier" based method (Wang, Garrity, Tiedje, & Cole, 2007). Alpha diversity, Bray-Curtis distance and phylogenetic distribution graphs were all created using the 'Phyloseq' data package.

When running the dada2 pipeline we had problems merging our reads. We tried trimming our reads at different lengths in order to negate this problem, but with little success. Out of the 120 000 reads, roughly 70 000 were merged. As a result of this we decided to instead assign taxonomy based on the forward reads alone. We speculate this problem can have occurred due to problems with our primers and potential bias as a result of complications with PCR amplification.

3.5.2 Metagenmoic Shotgun Analysis

After receiving completed assembled MAGs we analyzed them using GhostKoala version 2.2 (M. Kanehisa et al., 2016). GhostKoala is an annotation server for genomes and metagenome sequences and performs Kegg orthology assignments which identifies individual gene functions. It also recreates KEGG pathways BRITE hierarchy and KEGG modules to infer functions to the annotated organisms.

3.5.3 MinION analysis

After sequencing was completed the long-reads were analyzed using EPI2ME version. 2020.2 10-3247478. EPI2ME is Oxford Nanopores own analysis platform and was used to assign taxonomy and align sequences.

We used 16S rRNA gene data, from the same samples our long-reads were from, to select

reference genomes from public databases (i.e. NCBI) to align our shotgun sequences against. We also used previously constructed MAGs that was that was provided for us and was well represented in our sample as reference genomes to align with. The next step in our study would have been assembling and performed genome binning of these reads. However, due to time limitations as a result of the COVID-19 outbreak this could not be done.

4 Results

Table 4.1 Sample Summary Table. The different samples we used for the different sequencing platforms, what we did with them, and what we would have done, if time allowed.

	16S rRNA	Metagenome (MAGs)	Metagenome (Nanopore)	Metagenome (hybrid MAGs)
Rumen	543, 548, 550, 552, 559, 563, 568, 549, 556, 570, 575, 580, 581, 582, 584, 588, 591, 595, 600, 602,	TBD	TBD	TBD
Human Gut	XDC.Original, XDC.03	INDI.25.1, INDI25.2, INDI.25.6, INDI25.8, INDI25.9, INDI25.10 *INDI.01 *INDI.03 *INDI.10	XDC.Original, XDC.03	TBD

* Used for Nanopore vs Illumina MAG allignments

4.1 16S rRNA Amplicon results

The bacterial diversity in our samples was analyzed using 16S rRNA gene amplicon data that was generated from the sheep rumen samples collected that were subjected to three different feed groups. Our data was analyzed for differences in microbial composition as a result of different concentrations of sugar kelp in their diet. We analyzed the MiSeq results using the DADA2 pipeline and created graphs illustrating alpha diversity (see figure 4.1, 4.2), Bray-Curtis dissimilarities (See figure 4.3) and prokaryotic family distribution (see figure 4.4).

Our 16S rRNA sequencing produced 4,660,772 reads from the different feed types. After filtering and removing chimeras 3,424,480 reads remained. However, due to only utilizing forward reads, sequencing depth was reduced and could have hindered the creation of OTUs. Further sequencing and binning using merged reads could be beneficial for further discovery of microbial composition. The Shannon-Wiener index (Figure 4.1 & 4.2) represents both unique species and their evenness (Shannon, 1949). A high Shannon-Wiener index represent high microbial diversity and an even distribution of the bacteria, while a low index indicates low diversity and uneven distribution. Most our samples showed a high Shannon-Wiener index despite the index between our samples varied greatly, ranging from roughly 5.5 up to almost 7.5. The Simpsons index (Figure 4.1 & 4.2) adds to this information. A high Simpson index indicates high diversity within samples, most our samples had a Simpson index ranging between 0.996 - 0.999. These are high values and confirm the Shannon-Wiener index, indicating our samples had high microbial diversity evenly distributed. These is no apparent pattern between feed-groups nor sample type when measuring alpha-diversity.

The Bray-Curtis dissimilarities (Figure 4.3) measures the difference in species populations between samples, with the lower the values, the more similar the samples are. In our samples there are no apparent pattern between the different feed types, however there as distinct separation between fluid samples and particle samples. The fluid samples appear to be more similar in terms of species, while the particle appears more diverse. The prokaryotic family distribution appeared even across all the samples (Figure 4.4). *Prvotellaceae* was the most abundant closely followed by *Lachnospiraceae* and *Rikenellaceae*. Although the abundance varies greatly between the samples, the distribution still remains similar.

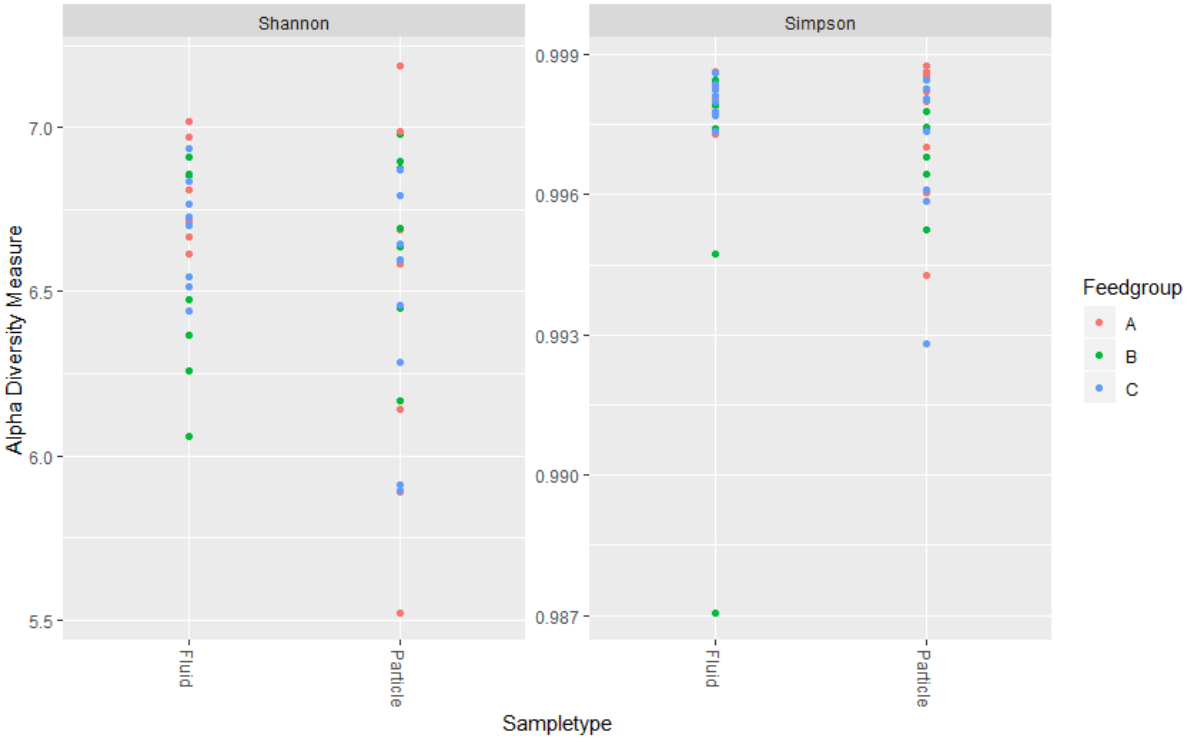


Figure 4.1 **Illustrating the alpha diversity of samples of the different feedgroups A, B and C**. Shannon-Wiener index in the left graph and Simpson diversity in the right. Y-axis indicate the amount of diversity in the Shannon index and how even the distribution is in the Simpson index. X-axis show particle and fluid samples, while the color indicates the feed group.

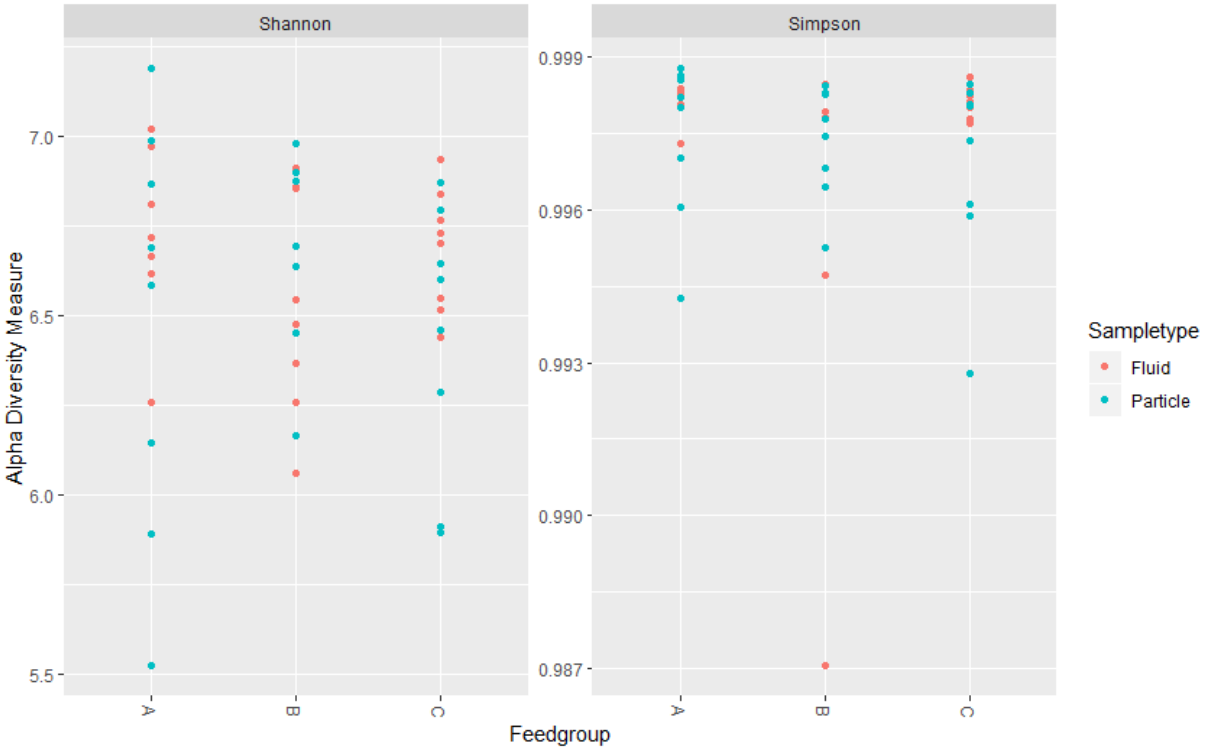


Figure 4.2 **Alpha diversity by sample type**. Left is Shannon Index, right is Simpson index. Each node represents a sample and A, B, C indicate feed group. Y-axis indicate the amount of diversity in the Shannon index and how even the distribution is in the Simpson index. The blue nodes are particle samples and orange are Fluid samples.

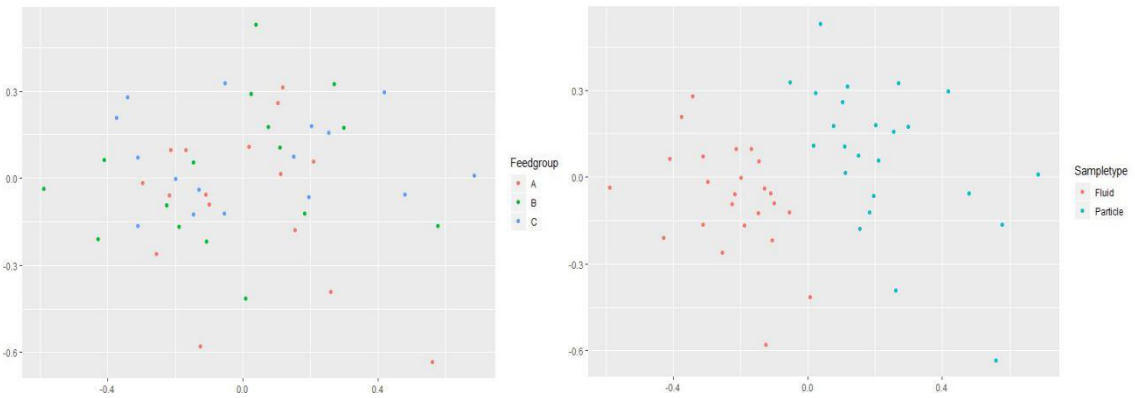


Figure 4.3 **Illustrating the Bray-Curtis dissimilarities between the samples**. Each node represents a sample and color represents feed group on the left graph and the sample type on the right side. The distance between the node represents how different they are in terms of specie composition.

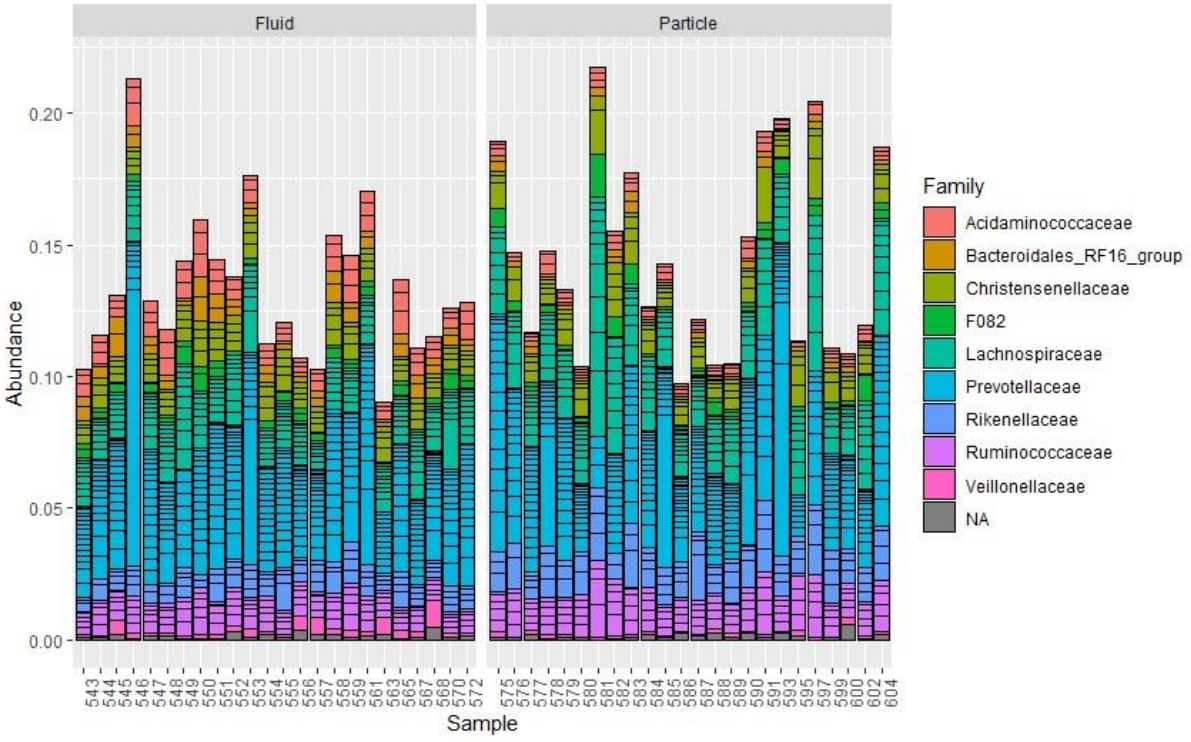


Figure 4.4 **Phylogenetic distribution of the 50 most represented Prokaryotic families in each sample.** Each node in a column represents a genus or species. The most represented families being *Rikenellaceae*, *Prevotellacea* and *Lachnospiraceae* across all samples.

Based on these results we selected 2-3 samples from each feeding group (A, B and C) with at least one from each sample type (particle/fluid). We selected our samples for further sequencing and analysis based on their microbial composition. Especially diverse samples, or samples containing bacterial families of interest for lignocellolytic degradation, like Lachnospiraceae, was prioritized.

4.2 Shotgun Metagenomic Results

Given the above-mentioned delays that affected this project, we used previously generated Illumina data that was generated from a human gut microbiome enrichment and assembled into MAGs (Ostrowski et al., 2020).

In total we selected six MAGs for genomes annotation and comparison against our long-read shotgun data. For MAG annotation, we used open reading frame (ORF) FASTA amino acid files (.faa) as input for GhostKoala (table 4.2). GhostKoala searches for orthology and assigns KEGG Orthology (KO) numbers based on distinct orthologs. The KO assignment of genes utilizes the SSDB database containing 'SSEARCH' computation, which compares all possible

genomes (Minoru Kanehisa, Sato, Kawashima, Furumichi, & Tanabe, 2016).

Based on these ortholog annotations, we used Ghostkoalas' KEGG mapping for network recreation which recreates the pathways and what predicted enzymes could be found in our MAGs. The KEGG pathway diagram (Figure 4.5 & 4.6) represents the relationship of genes and gene products. The enzyme commission numbers (EC numbers) are a numerical classification system for enzymes. Every EC number specify a specific enzyme catalyzed reaction and each node in the pathway illustrates a chemical compound these enzymes interact with and produces.

We decided to study two pathways involved in the degradation of lignocellulose across six different MAGs. These six MAGs were selected out of a set of 30 based on their genomic content. The Starch & Sucrose metabolism pathway (Figure 4.5) and Glycolysis pathway (Figure 4.6) was selected due to their importance in plant cell wall degradation and anaerobic fermentation in the gut ecosystem. In the Starch and Sucrose pathway we can find three active pathways that produces D-Glucose, which is the lignocellulolytic component that cellulose is comprised of. The D-Glucose in the MAG 'INDI25.2' are derived from trehalose, maltose and cellulose.
D-Glucose was predicted to be one of the most produced compounds in our samples and is further broken down in the glycolysis pathway. In our KEGG pathway we found most of the enzymatic reactions needed for breaking down D-Glucose to Pyruvate and Acetyl-CoA across all six MAGs (table 4.3).

These KEGG maps combined shows how there were enzymes present for breaking down Cellulose, Maltose and Glucose into the VFAs Pyruvate and Acetyl CoA. We also used GhostKoala to screen for CAZymes present it our samples (Table 4.4). Among the different CAZyme families, glycosyltransferase (GT) and glycoside hydrolases are the only one present in the 'starch and sucrose metabolism' pathways. GHs make out 68% of all CAZymes while GTs make out the remaining 32%. The most representated CAZyme is the GH13, which was present in all samples and account for 37.5% of all the CAZymes.

Table 4.2 **Annotation results from GhostKoala summarized. '**Number of nucleotides' and 'Contigs' shows the content of our faa files used as input, while 'Entries' and '% Annotation' shows the amount of MAGs and the percentage of our samples GhostKoala managed to annotate.

Samples	**Number of nucleotides**	**Contigs**	**Entries**	**% Annotation**	**Taxonomy (Phylum)**
INDI25.1	2796691	44	1198	48.1%	Monoglobus pectinilyticus
INDI25.2	6599619	124	2104	40.3%	Parabacteroides distasonis
INDI25.6	10827299	166	4582	49.9%	Blautia producta
INDI25.8	3520264	61	1670	50.4%	Faecalitalea cylindroides T2
INDI25.9	4409231	346	2926	70.7%	Escherichia coli ED1a
INDI25.10	4489977	190	1976	50.7%	butyrate producing bacterium SM4 1

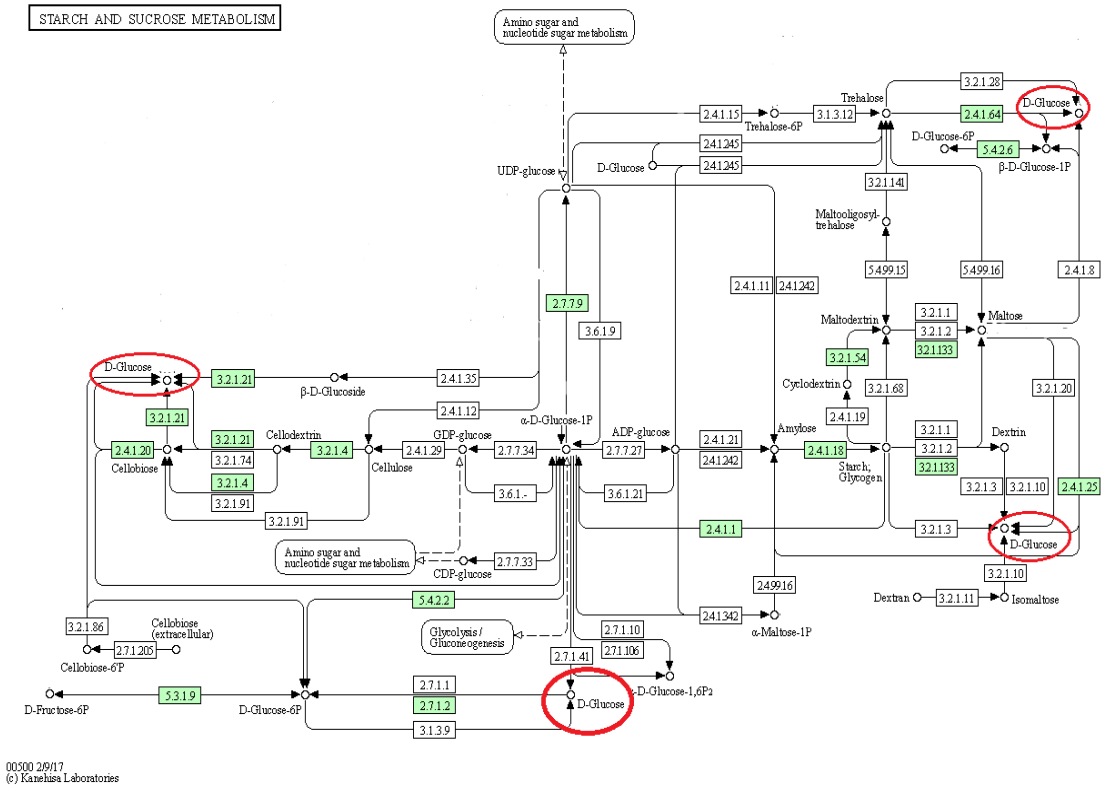


Figure 4.5 **Chemical pathway for Starch and sucrose metabolism**. E.C numbers marked in green are present E.C numbers in MAG INDI25.2. Marked in red are pathways connected to the production of D-Glucose. Arrows between E.C numbers and chemical compounds indicate how they interact.

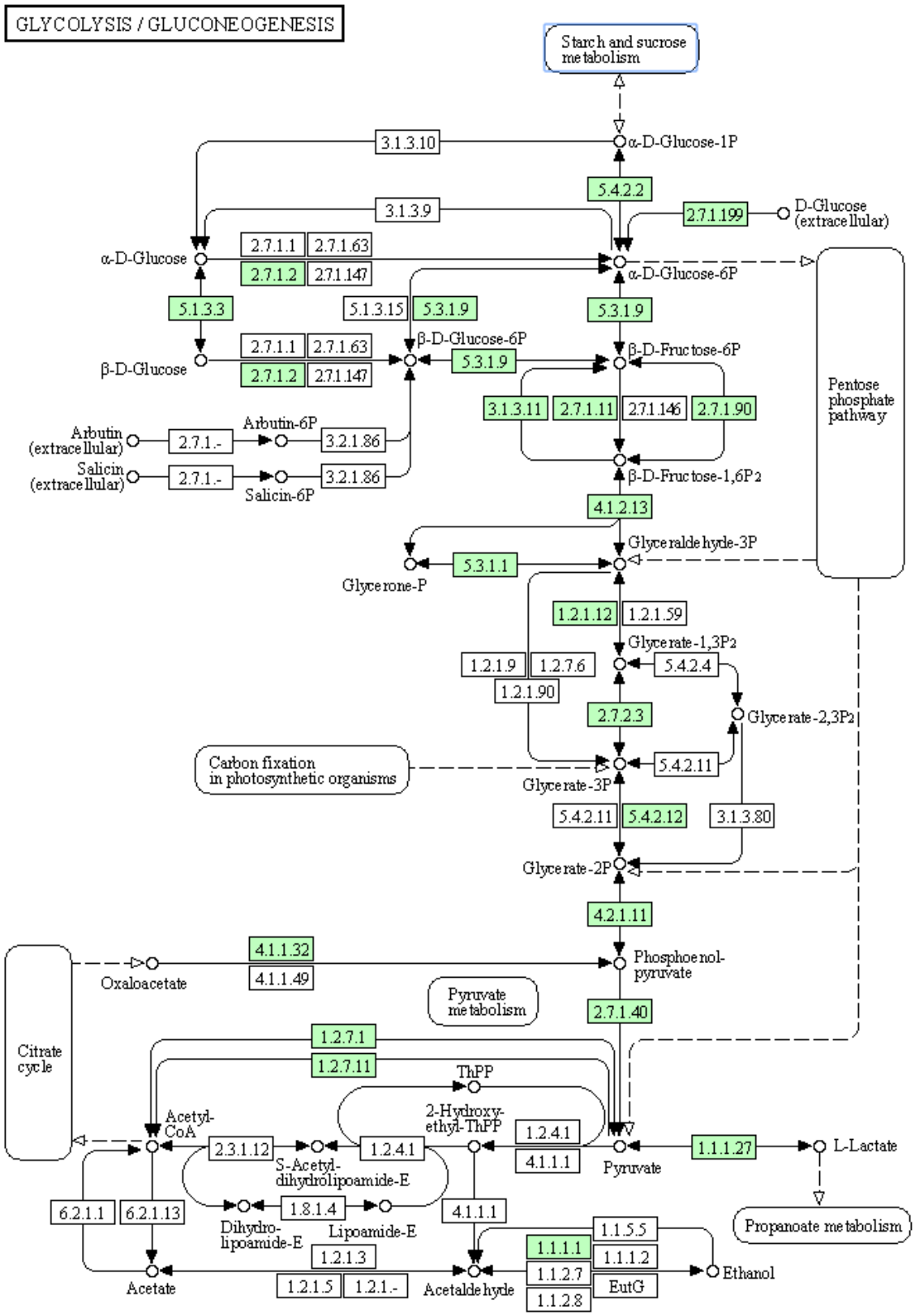


Figure 4.6 **Chemical pathways for Glycolysis.** E.C numbers marked in green are enzymes present in MAG INDI 25.2. Arrows between E.C numbers and chemical compounds indicate how they interact.

Table 4.3 **The presence of enzymes needed for breaking down D-Glucose to Pyruvate and/or Acetyl-CoA.** 'X' indicates presence and '-' indicates absence across six samples. E.C numbers are grouped based on similar functions.

Essential E.C	INDI25.1	INDI25.2	INDI25.6	INDI25.8	INDI25.9	INDI25.10
1.2.1.12/1.2.1.59	X,-	X,-	X,-	X,-	X,-	X,-
1.2.1.9/1.2.7.6/1.2.1.90	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-
1.2.7.1	X	X	X	3(X)	X	X
1.2.7.11	2(X)	2(X)	X	X	-	-
2.7.1.199	X	X	X	2(X)	3(X)	X
2.7.1.40	X	X	X	X	X	X
2.7.2.3	X	X	X	X	X	X
2.7.9.1/2.7.9.2	-,-	-,-	-,-	-,-	-,-	-,-
3.1.3.11/2.7.1.11/2.7.1.146/2.7.1.90	X,X,-,X	X,X,-,X	X,X,-,X	-,X,-,X	2(X),X,-,-	X,X,-,X
4.1.1.32	X	X	X	X	X	X
4.1.1.49	-	-	X	-	X	X
4.1.2.13	X	2(X)	3(X)	3(X)	3(X)	3(X)
4.2.1.11	X	X	X	X	X	X
5.3.1.1	X	X	-	-	-	-
5.3.1.9	X	X	2(X)	2(X)	2(X)	2(X)
5.4.2.11/5.4.2.12	-,2(X)	-,X	X,2(X)	X,X	X,2(X)	-,2(X)

Table 4.4 **Heatmap of present CAZymes associated with starch and lignocellulose hydrolysis across six MAGs.** Each row represents a CAZyme family and each column represent a sample. 'X' indicate presence of said E.C number, while '-' indicate absence. Numbers in front of 'X' indicate how many times this E.C number was discovered in a MAG.

MAGs	INDI25.1	INDI25.2	INDI25.6	INDI25.8	INDI25.9	INDI25.10	total
GH4				1			1
GH5	2	2	2	1	1		8
GH13	3	3	8	4	5	4	27
GH31		1	1		1	1	4
GH32		1	1	1		1	4
GH37					1		1
GH57		1					1
GH65	1						1
GH94	1						1
GH97		1					1
GT1			1	1	1	1	4
GT2			1	1	1		3
GT4		1		1	1		3
GT5			1	1	1	1	4
GT20		1			1		2
GT35	1	1	1	1	1	1	6
GT36			1				1
total	8	12	17	12	14	9	72

4.3 MinION Sequencing Results

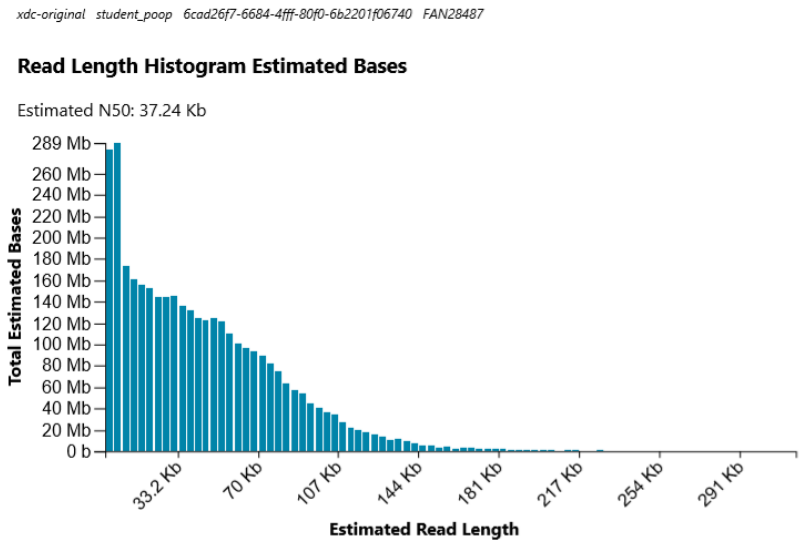
4.3.1 Sequencing

To complement our Illumina-generated MAGs, two DNA samples from the same human gut microbiome enrichment were running on the MINion for long-read sequencing (XDC.original & XDC.03).

Two sequencing runs were performed on two separate flowcells and MINion machines. The machine running the 'XDC-original sample' created 346.05 K reads and 3.47 Gb of bases. The estimated N50 value of 37.24 Kb indicate that half of our genome's sequences can be found in contigs longer than 37.24 Kb. Exceptional long reads were created in this run with some reads well above 200 Kb (Figure 4.7). The Q-score for the XDC-original run (Figure 4.8) was consistently high and scored 11. It never dipped below the median which indicates the generated reads was of high quality.

In the XDC.03 sample 325.86 K reads, and 3.08 Gb of bases was created. The estimated N50 value was 25.62 Kb, which indicate more fragmented reads compared to our XDC-original run. However, long reads above 146 K can be found, but not as extreme as in the other run (Figure 4.7). The average Q-score for our reads ranged between 10-11 (Figure 4.8), although lower than the XDC-Original sample, it still fluctuates within the expected median, despite dropping towards the end of the run.

(A)



(B)

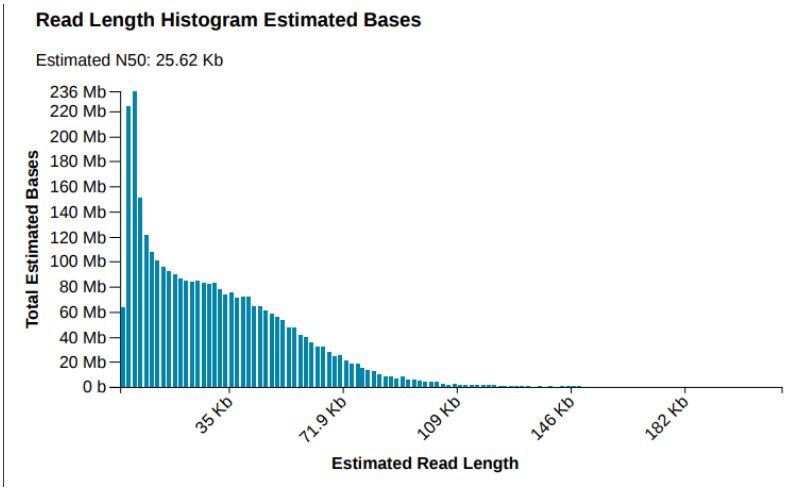


Figure 4.7 **Estimated read length from the MINion sequencing of XDC.original sample(A) and XDC.03 sample (B).** Kb = Kilo bases, Mb = Mega bases. X-axis indicate read length while Y-axis shows estimated amount of bases of X length

(A) (B)

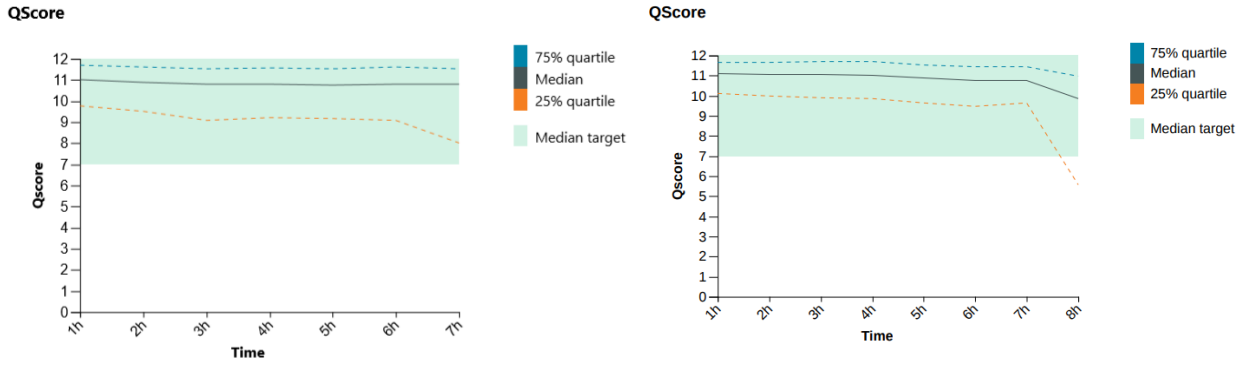


Figure 4.8 Average Q-score for the XDC-original sample (A) and XDC.03 sample (B). Q-score indicate the quality of identification of nucleotides generated in the sequencing run

4.3.2 Binning

The taxonomic binning of both samples was performed using Oxford Nanopores' WIMP. WIMP is a quantitative analysis tool for real-time species identification. It can be used to identify bacteria, fungi, archaea and viruses. We used the raw reads from our sequencing for this. Due to the high-quality reads that were generated and a shortage of time, DNA assembly was skipped.

In the XDC-original sample 290,568 reads were analyzed and 202,115 reads was classified, leaving 88,453 unclassified. The classified reads consisted of 99% bacteria <1% eukaryote, virus and archaea. The most abundant bacteria genus in the sample was *Bacteroides* with 101,358 reads. The most represented species among the *Bacteroides* was *Bacteroides cellulosilyticus,* which accounted for 84.33% of these reads. Other notable genera identified using WIMP was *Escherichia* with 33,867 reads and *Lachnoclostridium* with 31,693 reads (Figure 4.9).

In the XDC.03 sample 568,303 reads were analyzed, but only 335,260 of these reads was classified, leaving 233,043 unclassified. The classified reads consisted of 98% bacteria, 1 % eukaryote and <1% archaea and viruses.

In this sample the most predominant genus present was *Parabacteroides,* these made out 120,954 of the classified reads. Other notable genuses that was present is *Escherichia,* with 58,695 reads, *Bacteroides* with 34,217 reads and *Lachnocolstridium* with 25,032 reads (Figure 4.10). The most represented specie in the *Parabacteroides* genus was *Parabacteroides sp. CT06,* which is an uncharacterized species in the NCBI database with 55,680 reads and *Parabacteroides distasonis* with 53,967 reads.

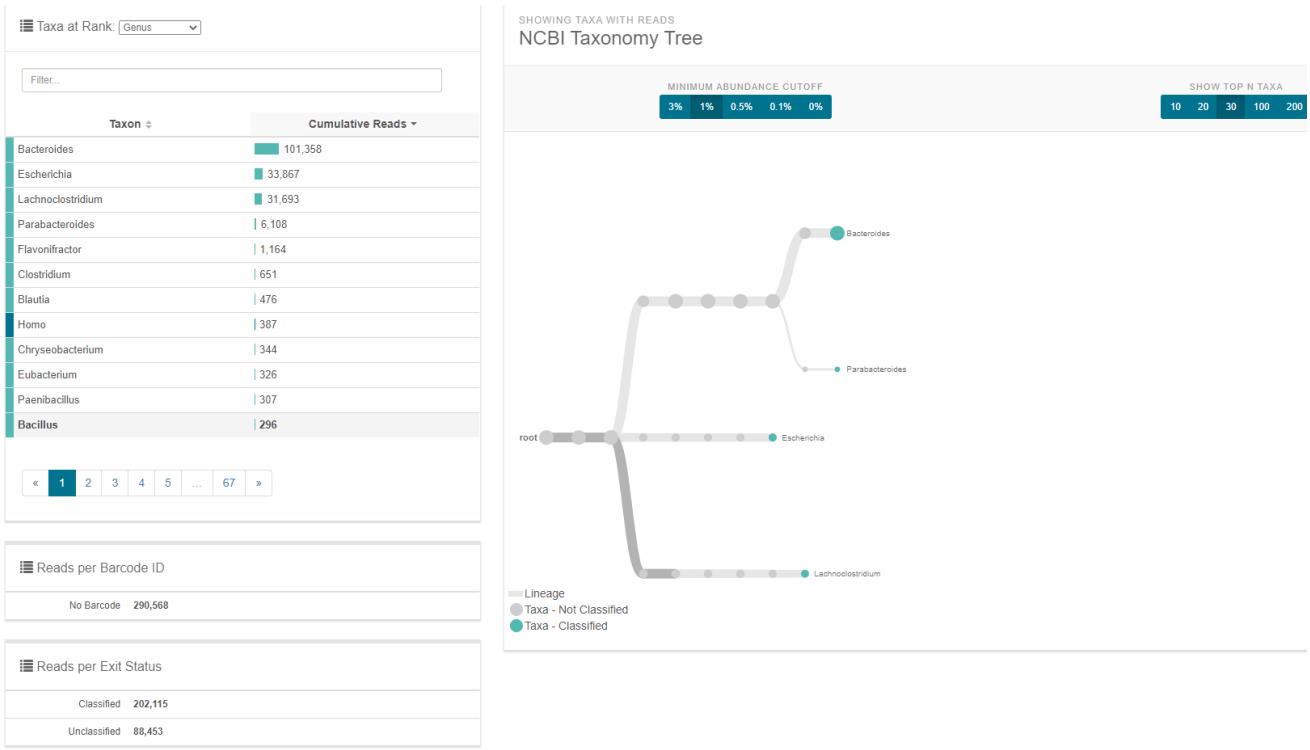


Figure 4.9 WIMP results from XDC-original. List in the left shows taxon and how many reads are allocated to which genus, while the taxonomy shows the most abundant taxa in form of relatedness.

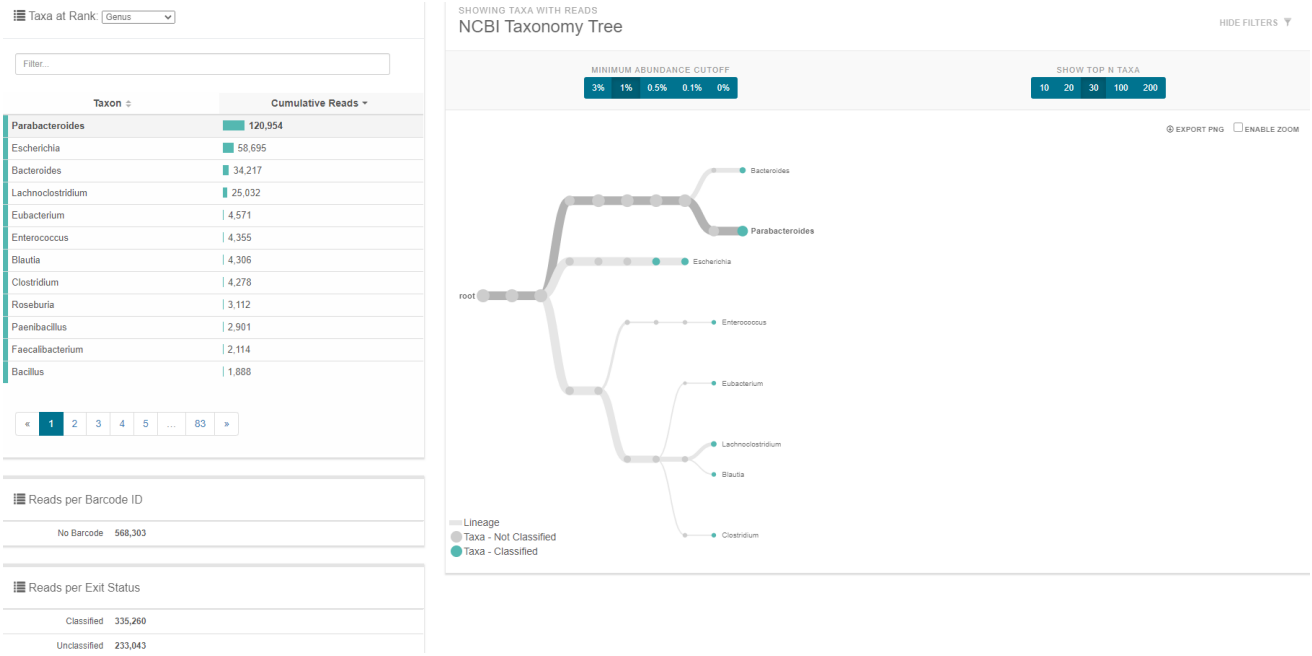


Figure 4.10 WIMP results from XDC.03. List on the left shows taxon and how man reads are allocated to which genus, while the taxonomy shows the most abundant taxa in the form of relatedness.

4.3.3 Sequence Alignment

To compare our long-read shotgun data against the previously generated Illumina dataset, we aligned our nanopore reads against the abovementioned MAGs. As our original research aims were to use a genome-centric approach, we used MAGs as opposed to the unassembled Illumina

metagenome. When aligning sequences, we previously analyzed 16S rRNA data containing taxonomic data for the same samples (Ostrowski et al., 2020) to determine what MAGs to align against our long-reads.

This 16S rRNA gene data also enabled us to predict the expected coverage of taxonomically assigned MAGs from both samples, which were further characterized based on level of completeness and contamination.

Finally, to estimate the coverage of our long-reads and see if the values compared to the relative abundances determined via 16S rRNA gene analysis, we used the MAGs reference to assess the coverage of our long-read binning, where highly represented genus like *Bacteroides* and *Escherichia* were detected.

For the XDC-Original sample a MAG taxonomically assigned as *Parabacteroides distasonis* with 99.22% completeness and 0% contamination and 98.73% Average Nucleotide and Amino Acid Identity (AAI) was selected for alignment.

For the XDC.03 sample a MAG taxonomically assigned as *Escherichia flexneri*, with 99.65% completeness and 0.08% contamination and 99.57% AAI was selected for alignment.

In our XDC-original MAG alignment 203,752 nanopore reads were analyzed and 22,134 were successfully aligned against the *P. distasonis* MAG with an average alignment accuracy of 79.6%. The coverage in this alignment was quite low where the majority sequences ranged between 20-40x coverage.

However, in these samples there were sections of the references that had massive amounts of coverage, in the 200x-300x range, while the remaining sequence positions had a coverage of roughly 20-25x, which draws down the average (Figure 4.12).

For the XDC.03 alignment 326,198 reads were analyzed and 69,406 was successfully aligned against the *E. flexneri* MAG with an average alignment accuracy of 92%. Surprisingly for this alignment, some sequences consisting of 28,513 alignments, stood out from the rest (Figure 4.11). Where average coverage resided between 50x-70x for the other sequences, this one stood out with 992x coverage. For these sequences, roughly the first 100 positions of the sequence had a coverage of 28,304 and 24,722, which defies expectations.

We initially expected this to be a 16S rRNA gene or ribosomal RNA, however by running the sequence through NCBI's BLASTN we deduced it to be transposons. This goes to show our long-reads have gaps and most likely didn't manage to cover entire ORFs, resulting in an

abnormal coverage of these regions, since these transposons can come from a multitude of species.

Both previous alignments are far from optimal. They both show a relative low coverage that is inconsistent over the positions of the contigs. However, one of the MAGs from XDC.03 showed great promise.

It was classified as *P. distasonis* with 98.27% completeness, 1.54% contamination and a 96.75% AII. Our Nanopore alignment against this MAG consisted of 335,260 reads, where 128,523 was aligned successfully with a 91.2% average alignment accuracy. The coverage for this MAG ranged between 240-320x for the majority of contigs. The coverage was more consistent across the sequence positions when compared to the other alignments (Figure 4.13). Although a great alignment, the result is not too surprising since the Parabacteroides was the most represented genus from our XDC.03 WIMP result based on 16S rRNA gene analysis (Figure 4.10).

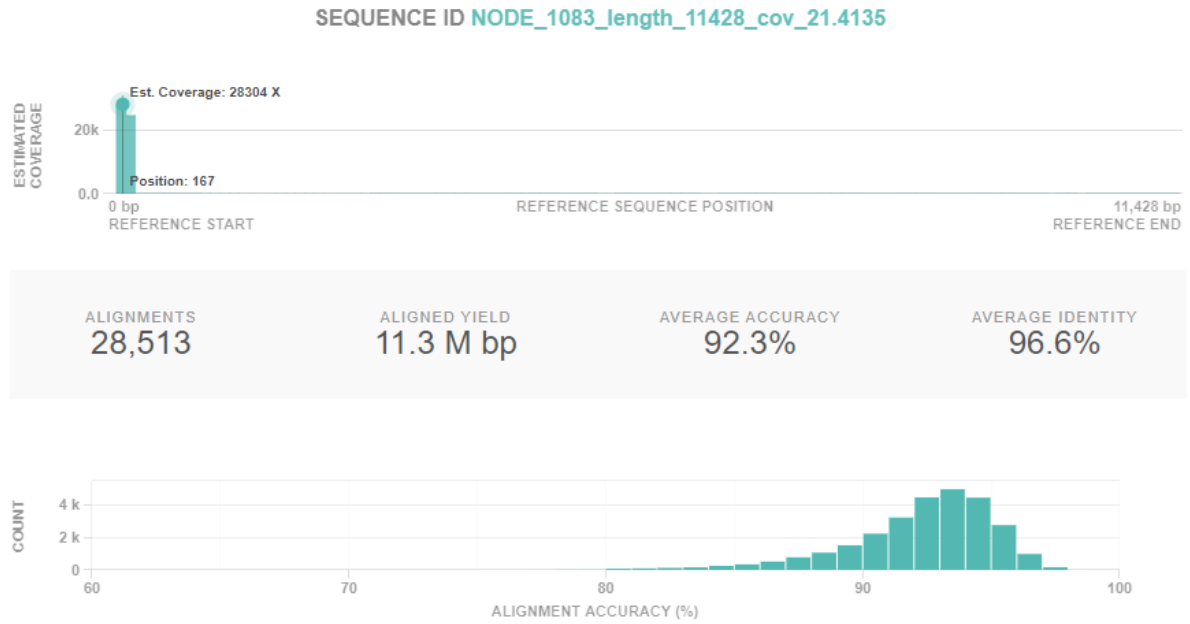


Figure 4.11 sequence alignment abnormality for XDC.03. This alignment was between our XDC.03 long reads and XDC.03. INDI.03 short reads. The first positions on the coverage bar has a coverage of >24,000, while the remaining, which are hard to see due to the difference, resides in the 50x-70x coverage range.

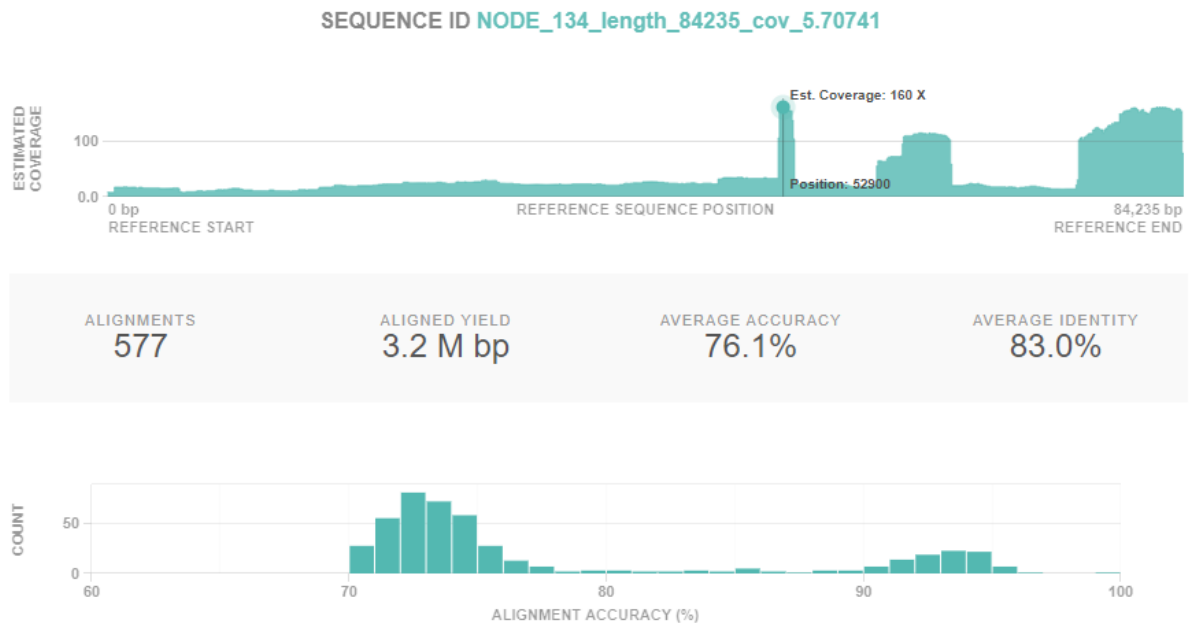


Figure 4.12 Sequence alignment for XDC-original. This alignment was between our XDC-original long reads and XDC-original. INDI.10 short reads. The average coverage for this sequence was 38x, but if we exclude the inconsistent 'spikes' of coverage, we can see that most positions were in the 15-25x coverage range.

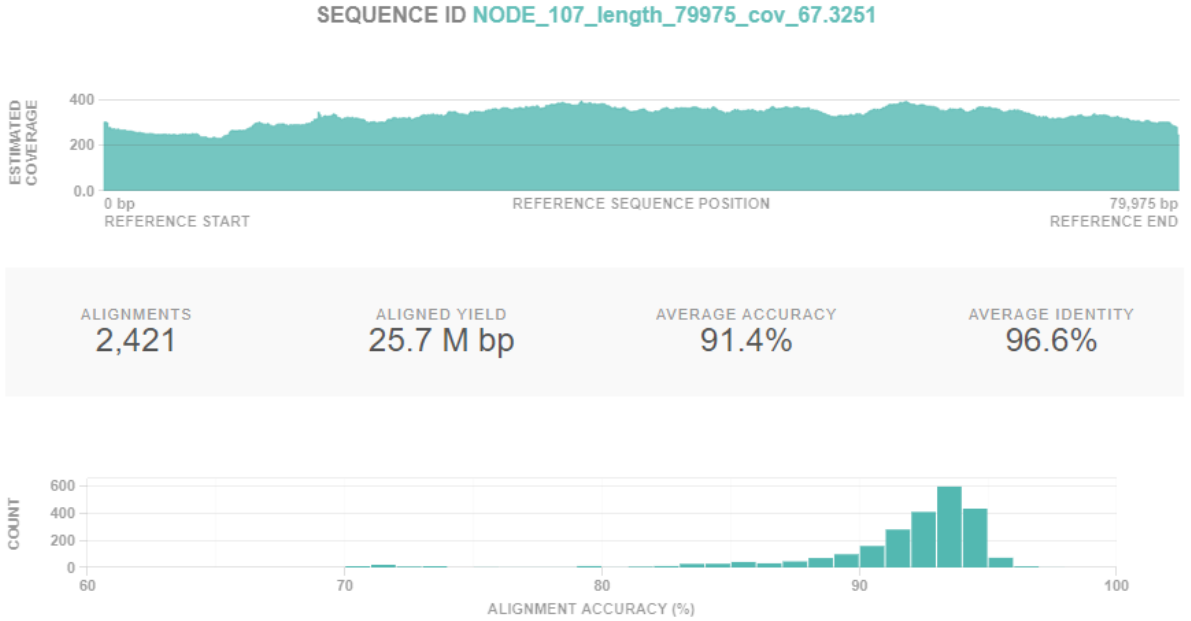


Figure 4.13 The best sequence alignment from XDC.03. This was an alignment between XDC.03 long read data and XDC.03. INDI01 short reads, this alignment had higher and more consistent coverage across the sequence positions.

4.3.4 Annotation

Based on the sequencing alignments we decided to annotate our INDI.03 sample due to the abnormal amounts of coverage this MAG contained. This was done in order to confirm or deny our suspicions of the coverage originating from transposons. The annotation was performed using Ghost Koala.

The alignment of MAG INDI 0.3 to our XDC.03 long-reads, showed it contained variable levels of coverage consisted of 69 contigs and 4518942 bases. From these, 3115 genes were annotated, which account for 74.6% of the input data which is roughly 25% more than our other annotations (table 4.2).

Fascinating, the most represented genes were involved in coding proteins involved in genetic information processing. Followed by proteins involved in signaling and cellular processing and carbohydrate metabolism (Figure 4.14). This sample also contained a considerable increase in genes involved environmental information processing (Figure 4.14).

We annotated the other MAGs we used in sequence alignment (INDI.01 & INDI.10)While Ghost Koala managed to assign 29 ABC-transporters in the INDI.10 sample, and 27 ABC-transporters in the INDI.01 sample, the INDI.03 alignment had 176 ABC-transporters assigned to it.

Furthermore, this MAG contained a considerable increase of genes involved in signal transduction, especially genes involved in two-component system. Our other MAGs (INDI.01 & INDI.10) contained 27 and 30 orthologs assigned to them involved in two-component signal transduction, while the INDI.03 sample with abnormal coverage had 148 orthologs assigned to it. This solidifies our suspicions of this MAG containing large amounts of transposons.

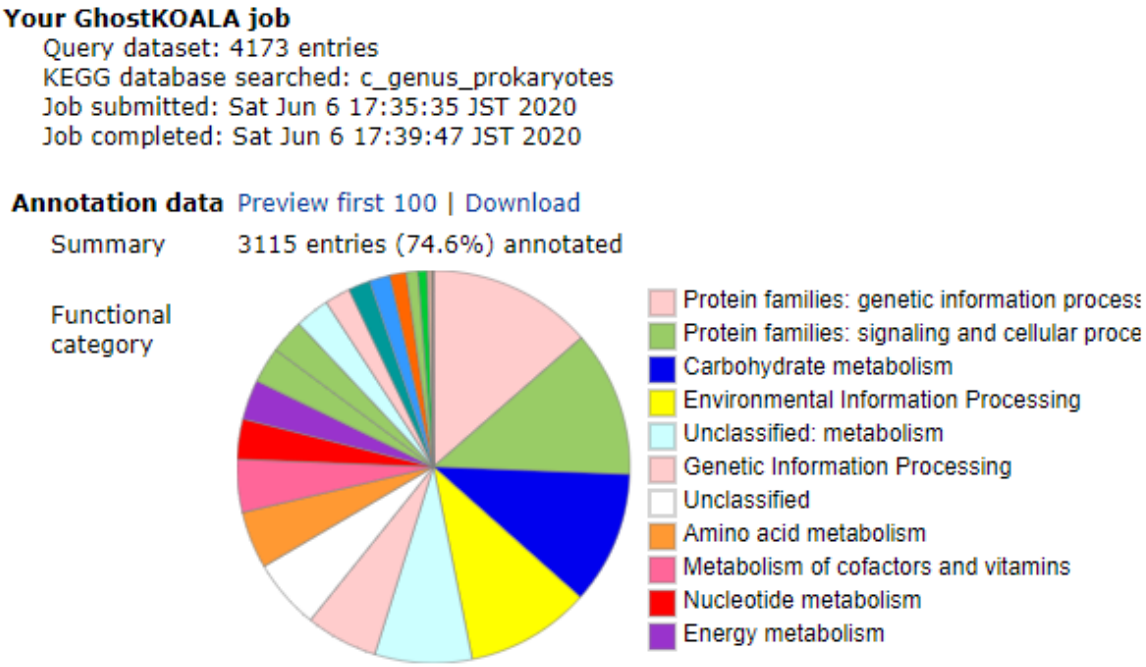


Figure 4.14 Annotation performed on MAG with inconsistent and abnormal amounts of coverage long-read coverage.

5 Discussion

The aim of this study was to explore metagenomic taxonomy and biochemical potential of gut microbiome samples from both the rumen of sheep and feces of human (i.e. distal gut), in addition to survey the potential strengths and detriments of both short and long read sequencing.

5.1 Sample & Library preparation

Sample preparation in this study was conducted using DNeasy Powerlyzer Powersoil Kit's Quick-Start Protocol for DNA-extraction, Illumina's 16S rRNA gene amplicon Metagenomic Sequencing Library Preparation for 16S rRNA gene sequencing and Oxford Nanopores "Genomic DNA by Ligation (SQK-LSK109)" protocol for Oxford's MINion long-read sequencing.

The success of our different sample preparation varied extensively with the different sequencing platforms, the library preparation we struggled the most with was when working the 16S rRNA gene. For our 16S rRNA gene sequencing we decided to use a kit for DNA extraction as

opposed to a traditional Phenol-chloroform approach, which proved to be troublesome. Our samples did not transition through the 16S rRNA gene library preparation protocol in an optimal manner. The most challenging step was the amplicon PCR, where our DNA was barely amplified, even though we tried to optimize it.

Arguments can be made for the benefits of a kit-based approach, such as time-efficient extraction, low costs, less chance of cross-contamination between samples and sufficient amount of DNA for 16S rRNA gene sequencing (Kramvis, Bukofzer, & Kew, 1996). However, in our experience more time was spent troubleshooting for our ineffective library preparation than what we saved using this approach, rendering one of the kit-approach greatest assets useless. Surprisingly, the kit we used was designed to handle soil samples, which should be harder to process than rumen samples, despite this, contaminants seemingly still got through. However, we cannot say for certain if our problems were due to using a kit approach, or if they originated somewhere else, like in sampling. Our problems could also be a result of poor handling on our part; however, it is unlikely given we had trouble amplifying all our original 48 samples.

5.2 16S rRNA gene analysis

The 16S rRNA data generated in this study, should have been used to map out the microbial composition of our samples, before shotgun sequencing could commence. However, due to time limitations, and trouble with the library preparation we had to swap sample type from sheep rumen samples to human gut samples between 16S rRNA gene sequencing and shotgun sequencing. If we had continued with the original plan, comparing the 16S rRNA gene taxonomy with our binned shotgun sequences would have provided important information in how well our assembly and binning were. In addition, 16S rRNA gene analysis and WGS analysis can provide different data from the same sample (Ranjan, Rani, Metwally, McGee, & Perkins, 2016).

While WGS taxonomy can more precise and assign on a species level, due to high coverage, the ability to utilize different databases for referencing by using both 16S rRNA gene and WGS analysis will provide a better understanding of the microbial composition of the sample. Furthermore, using 16S rRNA gene analysis is more cost-effective than WGS and can be used to validate WGS data, as well as a tool for selection of samples for further sequencing.

Despite not being able to use our 16S rRNA gene data for validation of the WGS data, our data can still provide useful information on the differences between feed-groups and sample-type.
There was no apparent pattern between the different feed-type and microbial composition when looking at alpha diversity and Bra-Curtis dissimilarities. This indicates different concentrations of sugar kelp, as a substitute to concentrates, will have little impact on the microbial composition of the rumen. Since seaweed is little used in today's society, it could be a valuable source of nutrients for our ruminants.
Utilizing this untapped potential could help reduce farm areas that is currently used for production of concentrates, which in turn will allow for more efficient use of suitable farm areas. This could help increase food production on a global scale.

Despite there being little difference between our feed-types, there were seemingly a difference in sample type. When looking at our Bray-Curtis dissimilarities for sample types, our fluid samples showed lesser values than our particle samples. This indicates that the microbial community associated with our fluid samples was structurally distinct to that associated with our particle samples. Based on this, using both fluid and particle samples for DNA-extraction should be valuable when looking at diversity, since only extracting the fluid seemed to filter out some of the microbes.
While our 16S rRNA gene data should show the taxonomic composition accurately down to family level, we have to take into consideration the troubles we experienced with amplification PCR and merging of forward and reverse read in our DADA2 pipeline. The troubles we experienced with amplification could have impacted our results. Although we used the recommended amount of amplification cycles, the small amount of actual amplification that took place could have resulted in inaccurate representations of microbial composition, this combined with the reduced coverage from only using forward reads, could have had negative repercussion on our results. Ways to improve our 16S rRNA gene data would be to re-extract DNA using a phenol-chloroform approach and use primers designated to the V2-V3 regions instead of V3-V4, which are better for taxonomic assignment (Bukin et al., 2019).

5.3 Shotgun sequence annotation

To circumvent the technical issues with our sheep rumen samples, we worked with fully assembled and binned MAGs from human gut samples, which enabled us to proceed with genome annotation and functional predictions.

The selected MAGs we used had varying degree of ORFs with functional assignments, but the majority ranged between 40-50% annotation. Despite the low amount of annotation, we still managed to recreate two important biochemical pathways, with E.C numbers involved to the full degradation of the highly prevalent fiber component D-glucose. This indicate that even though fiber fermentation to VFA contributes a relatively small amount of human energy consumption, it still is a valuable source of nutrients that our digestive tract is well adapted to process.

Furthermore, in our annotated MAGs we discovered several CAZymes involved in these pathways, however, the GH13 family was represented more than the others. The GH13 family consists of several enzymes, such as different amylases, glucosidases and hydrolases. Given the well represented glycolysis pathway we can assume the GH13 CAZyme family plays an important role in breaking down D-Glucose and turning it in to Pyruvate and Acetyl CoA.

Although we managed to annotate our MAGs to an extent, it is important to consider the amount of information we lost by just utilizing short reads for their assembly. The fact that roughly half our ORFs were annotated could come from the fact that assembly was hindered by the inherent flaws of just using the short reads.

The ability to detect and resolve assembly problems that originate with repeating elements such as, inverted transposable elements, transposons, gene duplications and prophage sequences is of major importance in order to assemble high quality MAGs (Moss, Maghini, & Bhatt, 2020). Similar studies, that utilized short-reads for their MAG assembly, have found that roughly half of their MAGs met the MIMAG values(Stewart et al., 2018) of 90% completeness and <5% contamination, set out by (Robert M. Bowers et al., 2017).This shows that even though MAGs assembled with short reads are a valuable source of information and can be used to discovery of novel species, their inability to assembly high quality MAGs leaves them sub-optimal for annotation.

If our study had gone as planned, we would have used the short-read MAGs as the bulk force of information regarding annotation, however, we would have polished said MAGs with our long-read data.

5.4 Long-read sequencing

For our long-read sequencing run we produced longer than average reads. When comparing our N50 values to the average presented by PacBio's SMRT sequencing technology of roughly 20 Kbp, our average N50 values of 25.62 Kbp and 37.24 Kbp is quite good. Our values even exceeded that of studies utilizing MAGs created from just short-reads (Datema et al., 2016).
These results shows the power and ease of use of the third-generation sequencing platform, since this was the first time we tested out the equipment and still managed to get good results. However, due to time limitations our use of these long reads was limited. Because of this we could only use the long-reads for assessing coverage and assigning taxonomy.

Based on the taxonomy assignment performed by WIMP we could see the most represented bacteria phylum was Bacteroides, Firmicutes and Proteobacteria which corresponds with the typical microbial composition of the human gut (Rinninella et al., 2019). We can also see the similarities on represented families between out 16S rRNA gene taxonomy and our long-read taxonomy, despite being samples from different animals (table 4.1). Lachnospiraceae and Bacteroidaceae were among the most represented families in both samples analyzed in this study. Based these families known function in fermentation we can assume they are important across species in fiber degradation. However, we cannot properly explore their actual biochemical potential due to lack of high-quality MAGs for functional annotation.

One of the challenges with utilizing third-generation long-read sequencing as a standalone platform, has been its lack of coverage(Lui, Nielsen, & Arkin, 2020). While 2nd generation platforms generate a massive array of reads, the long consistent reads generated with long-read sequencing hinders copies of the same DNA fragments to be sequenced simultaneously. Low amounts of coverage can make the discovery of structural variations unreliable, especially if base-calling errors are not accounted for.
However if moderate amounts of coverage 15-17x is reached, the long-reads should be able to effectively detect these genomic variants better than short-reads are able to (Cretu Stancu et al., 2017; De Coster et al., 2019).
For our long-reads, most of our custom alignment reached this 15x-17x threshold and could, in theory, be used for detecting genetic variations between microbes. However, since we did not assemble nor bin our reads it is hard to say if the coverage was consistent enough across the contigs to be reliably used. Many of the MAG contigs that were aligned against our nanopore data had an average coverage much higher than the modest 15x-17x. The most consistent

alignment we found had a coverage ranging between 200x-300x across several MAG contigs. With that much coverage, evenly dispersed across the contigs, binning and annotation could be easily performed using just long-reads, instead of co-assembling short and long reads.

To generate HQ-MAGs from just third-generation sequencing platforms would be more cost-effective and less time-consuming than having to rely on both short and long-read assembled MAGs. If one could consistently generate this amount of coverage evenly dispersed, one could argue third-generation sequencing platforms would replace second-generation sequencing platforms entirely. However, as of today, long-read sequencing is still costly and based on our results, this type of coverage is not common, but given the rapid development this technology this may no longer be the case, in the near future.

The inconsistencies we found in our long-read data provides a strong argument for how long-reads are better utilized when assembled with short-reads, acting as a correction tool in genome assembly instead of as a standalone analysis method, despite massive recent improvements for error-correction and polishing (Amarasinghe et al., 2020). If both short and long-reads are utilized for genome assembly, recreating "connected high quality (HQ)-MAGs" encoding entire 16S rRNA genes in addition to several other rRNA genes would be easier, especially for de-novo assembly (Singleton et al., 2020).

If time would have allowed, we could have linked our co-assembled HQ-MAGs to our 16S rRNA gene amplicon data. This would have allowed for better assembly contiguity, and the recovery of multicopy and conserved single copy genes which are normally missing in short-read assemblies (Lui et al., 2020). Furthermore, being able to link complete genomic data to well-developed 16S rRNA gene databases would allow us to explore microbial function and link it to structural trends in microbial communities.
This would improve our understanding of microbial communities and microbe's biochemical potential since we would have better taxonomic data that has been collected over years of study and our annotation would be more complete using the HQ MAGs.

6 Conclusion

The original aim of this study was to explore the microbial composition of sheep using second and third generation sequencing, and to better understand if sea kelp could be a viable supplement in feed. However, due to time limitations we instead explored the taxonomy of sheep rumen and human gut. We also tried annotating based on short reads alone and used MinION to better understand its limitation in terms of coverage.

The different aspects of our metagenomic study each provide valuable information regarding both taxonomy and function of microbes. Based on our results we cannot conclude much, since sample types was swapped during the study, making it hard to draw parallels across the different sequencing platforms.
Nevertheless, we can validate the different methods we used and explore their strengths and weaknesses. Our 16S rRNA gene analysis provided detailed information of the microbial composition of the sheep samples. In many cases microbes were identified down to a species level. It was a powerful and easy tool for us to use to validate our samples and better understand which samples to further explore. Furthermore, we could have used that data to validate our future sequencing, binning and annotation, improving their quality.

The short-read MAGs we obtained were useful for understanding the function of microbial communities found in the gut. Based on these MAGs we managed to recreate vital biochemical pathways for our metabolism, illustrating how powerful and useful annotation can be, despite not being optimized, in terms of the MAGs reaching MIMAG's threshold of 95% completeness and <5% contamination. Our long-read sequencing gave us detailed information of the sample's taxonomy. However, since we did not assemble our reads, we could not use them for annotation, which would have been interesting. Based on our alignments we could see the coverage these reads provided was inconsistent.
Despite this, most of our contigs had a coverage higher than the minimal amount recommended for differentiating between genetic variants, which means our long reads could have been just as a stand-alone platform.

Based on our findings we can conclude that each of these techniques utilized to explore microbial communities have their own niches and that the best way to use them, based on today's technology, is in unison. By assembling MAGs based on both short and long-reads we can obtain more complete genomes and using well developed 16S rRNA databases we can

explore the compositional trends of the microbes of interest, more valuable information can be obtained this way.

7 Appendix

7.1 Appendix 1

Measurements for our quantified DNA samples.

Formula:

(concentration in ng/µl) (660 g/mol × average library size) × 106 = concentration in nM

				DILUTION	
				tot vol:	
				50	
Sample	Well	(ng/ul)	nM	Dil. DNA	Water
543	A1	13.5	37.19008264	5.4	45
544	B1	10.1	27.82369146	7.2	43
545	C1	14.4	39.66942149	5.0	45
546	D1	12.9	35.53719008	5.6	44
547	E1	13.8	38.01652893	5.3	45
548	F1	6.83	18.815427	10.6	39
549	G1	5.83	16.06060606	12.5	38
550	H1	5.82	16.03305785	12.5	38
551	A2	6.61	18.20936639	11.0	39
552	B2	8.65	23.8292011	8.4	42
553	C2	12.4	34.15977961	5.9	44
554	D2	14.4	39.66942149	5.0	45
555	E2	19.7	54.26997245	3.7	46
556	F2	21.8	60.05509642	3.3	47
557	G2	16.7	46.00550964	4.3	46
558	H2	15.5	42.69972452	4.7	45
559	A3	8.89	24.49035813	8.2	42
561	B3	25.4	69.97245179	2.9	47
563	C3	21.4	58.95316804	3.4	47
565	D3	17	46.83195592	4.3	46

567	E3	20.4	56.19834711	3.6	46
568	F3	21.6	59.50413223	3.4	47
570	G3	27.7	76.30853994	2.6	47
572	H3	14	38.56749311	5.2	45
575	A4	10.5	28.92561983	6.9	43
576	B4	42.9	118.1818182	1.7	48
577	C4	23.8	65.56473829	3.1	47
578	D4	16.1	44.35261708	4.5	45
579	E4	21.8	60.05509642	3.3	47
580	F4	23.8	65.56473829	3.1	47
581	G4	20.1	55.37190083	3.6	46
582	H4	4.59	12.6446281	15.8	34
583	A5	14.2	39.1184573	5.1	45
584	B5	32.6	89.80716253	2.2	48
585	C5	25.5	70.24793388	2.8	47
586	D5	11.1	30.5785124	6.5	43
587	E5	22.7	62.53443526	3.2	47
588	F5	23.7	65.2892562	3.1	47
589	G5	20.5	56.4738292	3.5	46
590	H5	8.71	23.99449036	8.3	42
591	A6	22.1	60.8815427	3.3	47
593	B6	18.9	52.0661157	3.8	46
595	C6	26.9	74.1046832	2.7	47
597	D6	38.9	107.1625344	1.9	48
599	E6	30.8	84.84848485	2.4	48
600	F6	19.7	54.26997245	3.7	46
602	G6	19.6	53.99449036	3.7	46
604	H6	11.7	32.23140496	6.2	44

8 References

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), 30. doi:10.1186/s13059-020-1935-5

Barzon, L., Lavezzo, E., Militello, V., Toppo, S., & Palù, G. (2011). Applications of next-generation sequencing technologies to diagnostic virology. *International journal of molecular sciences, 12*(11), 7861-7884.

Bergman, E. (1990). Energy contributions of volatile fatty acids from the gastrointestinal tract in various species. *Physiological reviews, 70*(2), 567-590.

Besemer, J., Lomsadze, A., & Borodovsky, M. (2001). GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Research, 29*(12), 2607-2618. doi:10.1093/nar/29.12.2607

Boeke, J. D., Church, G., Hessel, A., Kelley, N. J., Arkin, A., Cai, Y., . . . Holt, L. (2016). The genome project-write. *Science, 353*(6295), 126-127.

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T., . . . Eloe-Fadrosh, E. A. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature biotechnology, 35*(8), 725-731.

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., . . . The Genome Standards, C. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature biotechnology, 35*(8), 725-731. doi:10.1038/nbt.3893

Browne, H. P., Neville, B. A., Forster, S. C., & Lawley, T. D. (2017). Transmission of the gut microbiota: spreading of health. *Nature Reviews Microbiology, 15*(9), 531.

Bryce, J., Boschi-Pinto, C., Shibuya, K., Black, R. E., & Group, W. C. H. E. R. (2005). WHO estimates of the causes of death in children. *The Lancet, 365*(9465), 1147-1152.

Bukin, Y. S., Galachyants, Y. P., Morozov, I. V., Bukin, S. V., Zakharenko, A. S., & Zemskaya, T. I. (2019). The effect of 16S rRNA region choice on bacterial community metabarcoding results. *Scientific Data, 6*(1), 190007. doi:10.1038/sdata.2019.7

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J., & Holmes, S. P. (2015). DADA2: High resolution sample inference from amplicon data. *bioRxiv*, 024034. doi:10.1101/024034

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), 581-583. doi:10.1038/nmeth.3869

Cardona, S., Eck, A., Cassellas, M., Gallart, M., Alastrue, C., Dore, J., . . . Manichanh, C. (2012). Storage conditions of intestinal microbiota matter in metagenomic analysis. *BMC Microbiology, 12*(1), 158. doi:10.1186/1471-2180-12-158

Castillo-González, A. B.-B., ME; Domínguez-Viveros, J; Chávez-Martínez, A. (2014). *Rumen microorganisms and fermentation* (Vol. 46).

Cho, I., Yamanishi, S., Cox, L., Methé, B. A., Zavadil, J., Li, K., . . . Teitler, I. (2012). Antibiotics in early life alter the murine colonic microbiome and adiposity. *Nature, 488*(7413), 621-626.

Cho, J.-C., & Tiedje, J. M. (2001). Bacterial species determination from DNA-DNA hybridization by using genome fragments and DNA microarrays. *Appl. Environ. Microbiol., 67*(8), 3677-3682.

Cretu Stancu, M., van Roosmalen, M. J., Renkens, I., Nieboer, M. M., Middelkamp, S., de Ligt, J., . . . Kloosterman, W. P. (2017). Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nature Communications, 8*(1), 1326. doi:10.1038/s41467-017-01343-4

Czerkawski, J. W. (1986). An Introduction to Rumen Studies. *Pergamon Press*, 7-10.
Datema, E., Hulzink, R. J. M., Blommers, L., Espejo Valle-Inclan, J., van Orsouw, N., Wittenberg, A. H. J., & de Vos, M. (2016). The megabase-sized fungal genome of Rhizoctonia solani assembled from nanopore reads only. *bioRxiv*, 084772. doi:10.1101/084772
De Coster, W., De Rijk, P., De Roeck, A., De Pooter, T., D'Hert, S., Strazisar, M., . . . Van Broeckhoven, C. (2019). Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. *Genome research, 29*(7), 1178-1187.
Dijkstra, J. (1994). Production and absorption of volatile fatty acids in the rumen. *Livestock Production Science, 39*(1), 61-69. doi:https://doi.org/10.1016/0301-6226(94)90154-6.
Fulde, M., & Hornef, M. W. (2014). Maturation of the enteric mucosal innate immune system during the postnatal period. *Immunological reviews, 260*(1), 21-34.
Garbeva, P., Veen, J. A. v., & Elsas, J. D. v. (2004). MICROBIAL DIVERSITY IN SOIL: Selection of Microbial Populations by Plant and Soil Type and Implications for Disease Suppressiveness. *Annual Review of Phytopathology, 42*(1), 243-270. doi:10.1146/annurev.phyto.42.012604.135455
Gilpatrick, T., Lee, I., Graham, J. E., Raimondeau, E., Bowen, R., Heron, A., . . . Timp, W. (2020). Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nature biotechnology, 38*(4), 433-438. doi:10.1038/s41587-020-0407-5
Haque, K. A., Pfeiffer, R. M., Beerman, M. B., Struewing, J. P., Chanock, S. J., & Bergen, A. W. (2003). Performance of high-throughput DNA quantification methods. *BMC Biotechnology, 3*(1), 20. doi:10.1186/1472-6750-3-20
Heyer, R., Schallert, K., Zoun, R., Becher, B., Saake, G., & Benndorf, D. (2017). Challenges and perspectives of metaproteomic data analysis. *Journal of biotechnology, 261*, 24-36.
Hugenholtz, P. (2002). Exploring prokaryotic diversity in the genomic era. *Genome Biology, 3*(2), reviews0003. 0001.
Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome research, 17*(3), 377-386.
Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., . . . Tappu, R. (2016). MEGAN community edition-interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS computational biology, 12*(6).
Huttenhower, C., Kostic, A. D., & Xavier, R. J. (2014). Inflammatory bowel disease as a model for translating the microbiome. *Immunity, 40*(6), 843-854.
Hyatt, D., Chen, G.-L., Locascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics, 11*, 119-119. doi:10.1186/1471-2105-11-119
Ijssennagger, N., Belzer, C., Hooiveld, G. J., Dekker, J., van Mil, S. W., Müller, M., . . . van der Meer, R. (2015). Gut microbiota facilitates dietary heme-induced epithelial hyperproliferation by opening the mucus barrier in colon. *Proceedings of the National Academy of Sciences, 112*(32), 10038-10043.
Imelfort, M., Parks, D., Woodcroft, B. J., Dennis, P., Hugenholtz, P., & Tyson, G. W. (2014). GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ, 2*, e603.
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. doi:10.1186/s13059-016-1103-0
Janabi, A. H., Kerkhof, L. J., McGuinness, L. R., Biddle, A. S., & McKeever, K. H. (2016). Comparison of a modified phenol/chloroform and commercial-kit methods for extracting DNA from horse fecal material. *Journal of microbiological methods, 129*, 14-19.
Jose C. Clemente, Luke K. U., Laura Wegener Parfrey, Rob Knight,. (2012,). The Impact of the Gut Microbiota on Human Health: An Integrative View. *Cell, 148*(6), 1258-1270. doi:https://doi.org/10.1016/j.cell.2012.01.035

Kallmeyer, J., Pockalny, R., Adhikari, R. R., Smith, D. C., & D'Hondt, S. (2012). Global distribution of microbial abundance and biomass in subseafloor sediment. *Proceedings of the National Academy of Sciences, 109*(40), 16213-16216.

Kamada, M., Hase, S., Sato, K., Toyoda, A., Fujiyama, A., & Sakakibara, Y. (2014). Whole genome complete resequencing of Bacillus subtilis natto by combining long reads with high-quality short reads. *PloS one, 9*(10), e109999-e109999. doi:10.1371/journal.pone.0109999

Kamada, N., Chen, G. Y., Inohara, N., & Núñez, G. (2013). Control of pathogens and pathobionts by the gut microbiota. *Nature immunology, 14*(7), 685.

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. *J Mol Biol, 428*(4), 726-731. doi:10.1016/j.jmb.2015.11.006

Kang, D. D., Froula, J., Egan, R., & Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ, 3*, e1165. doi:10.7717/peerj.1165

Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L., & Gordon, J. I. (2011). Human nutrition, the gut microbiome and the immune system. *Nature, 474*(7351), 327-336. doi:10.1038/nature10213

Koren, S., Harhay, G. P., Smith, T. P. L., Bono, J. L., Harhay, D. M., McVey, S. D., . . . Phillippy, A. M. (2013). Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biology, 14*(9), R101. doi:10.1186/gb-2013-14-9-r101

Krajmalnik-Brown, R., Ilhan, Z.-E., Kang, D.-W., & DiBaise, J. K. (2012). Effects of Gut Microbes on Nutrient Absorption and Energy Regulation. *Nutrition in Clinical Practice, 27*(2), 201-214. doi:10.1177/0884533611436116

Kramvis, A., Bukofzer, S., & Kew, M. C. (1996). Comparison of hepatitis B virus DNA extractions from serum by the QIAamp blood kit, GeneReleaser, and the phenol-chloroform method. *Journal of Clinical Microbiology, 34*(11), 2731-2733. Retrieved from https://jcm.asm.org/content/jcm/34/11/2731.full.pdf

Krause, L., Diaz, N. N., Goesmann, A., Kelley, S., Nattkemper, T. W., Rohwer, F., . . . Stoye, J. (2008). Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Research, 36*(7), 2230-2239. doi:10.1093/nar/gkn038

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.

Lee, H., Gurtowski, J., Yoo, S., Nattestad, M., Marcus, S., Goodwin, S., . . . Schatz, M. (2016). Third-generation sequencing and the future of genomics. *bioRxiv*, 048603.

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., . . . Fan, W. (2011). Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. *Briefings in Functional Genomics, 11*(1), 25-37. doi:10.1093/bfgp/elr035

Limasset, A., Cazaux, B., Rivals, E., & Peterlongo, P. (2016). Read mapping on de Bruijn graphs. *BMC Bioinformatics, 17*(1), 237. doi:10.1186/s12859-016-1103-9

Liu, Y., Hou, T., & Fu, L. (2015). *A new unsupervised binning method for metagenomic dataset with automated estimation of number of species* (2167-9843). Retrieved from

Lloyd-Price, J., Abu-Ali, G., & Huttenhower, C. (2016). The healthy human microbiome. *Genome medicine, 8*(1), 1-11.

Lui, L. M., Nielsen, T. N., & Arkin, A. P. (2020). A method for achieving complete microbial genomes and better quality bins from metagenomics data. *bioRxiv*, 2020.2003.2005.979740. doi:10.1101/2020.03.05.979740

Lukashin, A. V., & Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Research, 26*(4), 1107-1115. doi:10.1093/nar/26.4.1107

MacDonald, N. J., Parks, D. H., & Beiko, R. G. (2012). Rapid identification of high-confidence taxonomic assignments for metagenomic data. *Nucleic Acids Research, 40*(14), e111-e111. doi:10.1093/nar/gks335

Mande, S. S., Mohammed, M. H., & Ghosh, T. S. (2012). Classification of metagenomic sequences: methods and challenges. *Briefings in bioinformatics, 13*(6), 669-681. doi:10.1093/bib/bbs054
Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in genetics, 24*(3), 133-141.
Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal, 17*(1), 10-12.
McHardy, A. C., Martín, H. G., Tsirigos, A., Hugenholtz, P., & Rigoutsos, I. (2007). Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods, 4*(1), 63-72. doi:10.1038/nmeth976
Mende, D. R., Waller, A. S., Sunagawa, S., Järvelin, A. I., Chan, M. M., Arumugam, M., . . . Bork, P. (2012). Assessment of metagenomic assembly using simulated next generation sequencing data. *PloS one, 7*(2).
Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., . . . Edwards, R. A. (2008). The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics, 9*(1), 386. doi:10.1186/1471-2105-9-386
Monzoorul Haque, M., Ghosh, T. S., Komanduri, D., & Mande, S. S. (2009). SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics, 25*(14), 1722-1730. doi:10.1093/bioinformatics/btp317
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).
Moran, J. (2005). Feeding Management for Small Holder Dairy Farmers in the Humid Tropics. *Landlink Press*, 42-49.
Morgan, X. C., Tickle, T. L., Sokol, H., Gevers, D., Devaney, K. L., Ward, D. V., . . . Huttenhower, C. (2012). Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biology, 13*(9), R79. doi:10.1186/gb-2012-13-9-r79
Moss, E. L., Maghini, D. G., & Bhatt, A. S. (2020). Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nature biotechnology, 38*(6), 701-707. doi:10.1038/s41587-020-0422-6
Neuman, H., Debelius, J. W., Knight, R., & Koren, O. (2015). Microbial endocrinology: the interplay between the microbiota and the endocrine system. *FEMS microbiology reviews, 39*(4), 509-521.
Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Res, 27*(5), 824-834. doi:10.1101/gr.213959.116
Ostrowski, M., Rosa, S. L. L., Kunath, B., Robertson, A., Pereira, G. V., Yao, T., . . . Martens, E. (2020). Tu1926 MICROBIAL DIGESTION OF XANTHAN GUM IN THE HUMAN GUT. *Gastroenterology, 158*(6), S-1221. doi:10.1016/S0016-5085(20)33713-6
Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res, 25*(7), 1043-1055. doi:10.1101/gr.186072.114
Peng, X., Yu, K.-Q., Deng, G.-H., Jiang, Y.-X., Wang, Y., Zhang, G.-X., & Zhou, H.-W. (2013). Comparison of direct boiling method with commercial kits for extracting fecal microbiome DNA by Illumina sequencing of 16S rRNA tags. *Journal of microbiological methods, 95*(3), 455-462.
Pevzner, P. A., Tang, H., & Waterman, M. S. (2001). An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences, 98*(17), 9748-9753.
Pope, P. B., Smith, W., Denman, S. E., Tringe, S. G., Barry, K., Hugenholtz, P., . . . Morrison, M. (2011). Isolation of Succinivibrionaceae implicated in low methane emissions from Tammar wallabies. *Science, 333*(6042), 646-648. doi:10.1126/science.1205760
Price, A. M., Hayer, K. E., Depledge, D. P., Wilson, A. C., & Weitzman, M. D. (2019). Novel splicing and open reading frames revealed by long-read direct RNA sequencing of adenovirus transcripts. *bioRxiv*, 2019.2012.2013.876037. doi:10.1101/2019.12.13.876037

Ranjan, R., Rani, A., Metwally, A., McGee, H. S., & Perkins, D. L. (2016). Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochemical and biophysical research communications, 469*(4), 967-977. doi:10.1016/j.bbrc.2015.12.083

Reinhardt, C., Bergentall, M., Greiner, T. U., Schaffner, F., Östergren-Lundén, G., Petersen, L. C., . . . Bäckhed, F. (2012). Tissue factor and PAR1 promote microbiota-induced intestinal vascular remodelling. *Nature, 483*(7391), 627-631.

Ribeca, P., & Valiente, G. (2011). Computational challenges of sequence classification in microbiomic data. *Briefings in bioinformatics, 12*(6), 614-625.

Rinninella, E., Raoul, P., Cintoni, M., Franceschi, F., Miggiano, G. A. D., Gasbarrini, A., & Mele, M. C. (2019). What is the Healthy Gut Microbiota Composition? A Changing Ecosystem across Age, Environment, Diet, and Diseases. *Microorganisms, 7*(1), 14. doi:10.3390/microorganisms7010014

Robertson, R. C., Manges, A. R., Finlay, B. B., & Prendergast, A. J. (2019). The human microbiome and child growth–first 1000 days and beyond. *Trends in microbiology, 27*(2), 131-147.

Rodríguez-Valera, F. (2004). Environmental genomics, the big picture? *FEMS Microbiology Letters, 231*(2), 153-158.

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

Schalamun, M., Nagar, R., Kainer, D., Beavan, E., Eccles, D., Rathjen, J. P., . . . Schwessinger, B. (2019). Harnessing the MinION: An example of how to establish long-read sequencing in a laboratory using challenging plant tissue from Eucalyptus pauciflora. *Molecular ecology resources, 19*(1), 77-89. doi:10.1111/1755-0998.12938

Sedlar, K., Kupkova, K., & Provaznik, I. (2017). Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. *Computational and Structural Biotechnology Journal, 15*, 48-55.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics, 30*(14), 2068-2069. doi:10.1093/bioinformatics/btu153

Shannon, C. E., & Weaver, W. (1949). The mathematical theory of information. *University of Illinois Press, 97*.

Sieber, P., Platzer, M., & Schuster, S. (2018). The definition of open reading frame revisited. *Trends in genetics, 34*(3), 167-170.

Simpson, J. T., & Pop, M. (2015). The theory and practice of genome sequence assembly. *Annual review of genomics and human genetics, 16*.

Singleton, C. M., Petriglieri, F., Kristensen, J. M., Kirkegaard, R. H., Michaelsen, T. Y., Andersen, M. H., . . . Nielsen, P. H. (2020). Connecting structure to function with the recovery of over 1000 high-quality activated sludge metagenome-assembled genomes encoding full-length rRNA genes using long-read sequencing. *bioRxiv*.

Staden, R. (1979). A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research, 6*(7), 2601-2610. doi:10.1093/nar/6.7.2601

Staley, J. T., & Konopka, A. (1985). Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annual review of microbiology, 39*(1), 321-346.

Stewart, R. D., Auffret, M. D., Warr, A., Walker, A. W., Roehe, R., & Watson, M. (2018). The genomic and proteomic landscape of the rumen microbiome revealed by comprehensive genome-resolved metagenomics. *bioRxiv*, 489443. doi:10.1101/489443

Teeling, H., & Glöckner, F. O. (2012). Current opportunities and challenges in microbial metagenome analysis—a bioinformatic perspective. *Briefings in bioinformatics, 13*(6), 728-742.

Teeling, H., Waldmann, J., Lombardot, T., Bauer, M., & Glöckner, F. O. (2004). TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics, 5*(1), 163. doi:10.1186/1471-2105-5-163

Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics, 13*(1), 36-46.

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.

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol., 73*(16), 5261-5267.

Wick, R. R., Judd, L. M., & Holt, K. E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biology, 20*(1), 129. doi:10.1186/s13059-019-1727-y

Yang, B., Wang, Y., & Qian, P.-Y. (2016). Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC Bioinformatics, 17*(1), 135.

Yano, J. M., Yu, K., Donaldson, G. P., Shastri, G. G., Ann, P., Ma, L., . . . Hsiao, E. Y. (2015). Indigenous bacteria from the gut microbiota regulate host serotonin biosynthesis. *Cell, 161*(2), 264-276.

Ye, C., Hill, C. M., Wu, S., Ruan, J., & Ma, Z. (2016). DBG2OLC: Efficient Assembly of Large Genomes Using Long Erroneous Reads of the Third Generation Sequencing Technologies. *Scientific Reports, 6*(1), 31900. doi:10.1038/srep31900

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

Postboks 5003
NO-1432 Ås
Norway