FACULTY OF SCIENCE AND TECHNOLOGY

MASTER THESIS

Study programme / specialization:
Biological Chemistry

The spring semester, 2022

Open

Author:
Marthe Tofthagen

Marthe Tofthagen
.....
(Signature author)

Course coordinator:
Cathrine Lillo

Supervisor:
Mark van der Giezen

Thesis title: Third-generation sequencing of IBD patients' gut microbiome

Credits (ECTS): 60 stp.

Keywords:
Inflammatory Bowel Disease
Next-generation sequencing
Third-generation sequencing
Oxford Nanopore Technology
MinION
Gut microbiome
Blastocystis

Pages: ..**75 pages**.....

+ appendix: ..**7 pages**..

Stavanger, June 2022

Acknowledgments

This project was performed from August 2021 until June 2022 at the Department of Chemistry, Bioscience, and Environmental Engineering at the Faculty of Science and Technology, University of Stavanger, as a part of my master's degree in Biological Chemistry. All the work done on this project would not be possible without the collaboration between the University of Stavanger and Stavanger University Hospital (SUS).

First, I would like to thank my supervisor, Professor of Biological Chemistry, Dr. Mark van der Giezen, for all the help this past year. Your research group has all contributed with their knowledge and guidance, making this project much more fun and enjoyable. Your knowledge and encouragement have been inspiring and have led me to an interest in genomic sequencing.

Enough cannot be said to thank postdoc, Dr. Martin Watson, for all your help and encouragement over the past year, for answering all my stupid questions and anxious messages, and for making everyday fun and challenging for me.

Finally, thanks to my friends and family for their support and encouragement in everything I do, even if it involved talking about stool samples over dinner. Frida, Hege, and Ingvild, thank you so much for the past two years. The endless amount of coffee breaks, complaining, and cooperation have made this time so much more memorable. Special thanks to Ingrid and Marthe for taking the time to read my thesis and always giving the best advice.

Stavanger, June 2022

Marthe Tofthagen

Abstract

The current opportunities for thorough gut microbiota profiling using next-generation sequencing (NGS) have opened up for a wide range of metagenomic studies. IBD prevalence is increasing in developed and developing countries that are gradually adapting to a more modern lifestyle. Although the specific pathogenesis is unknown, dysbiosis of the intestinal microbiota is widely believed to cause or promote intestinal inflammation. Intestinal microbial compositions in IBD and healthy individuals have been reported in an increasing number of studies using non-cultured 16S rRNA sequencing technologies. Studying intestinal microbes in relation to their ecological niche, such as relationships with gut microbiota, is an essential step toward fine-tuning our clinical and public health understanding of colonization by intestinal microbes. The main goal of this study was to assess the composition of the gut microbiome in patients diagnosed with IBD using next-generation sequencing. Samples from patients diagnosed with ulcerative colitis (UC) or Crohn's disease (CD), included in a clinical trial at Stavanger university hospital (SUS), were used for this study. DNA was extracted using a modified protocol for fecal DNA extraction in combination with Fast DNA stool kit from Qiagen. Library preparation using protocols provided by Oxford Nanopore Technologies (ONT) was done before sequencing with the MinION sequencer. A post-sequencing pipeline for data analysis provided information about taxonomic classification and diversity of the samples. Samples were also sequenced using Illumina MiSeq to establish the prevalence of *Blastocystis*. The results showed that sequencing with ONTs' MinION provided taxonomic identification down to species level.

The most abundant phyla among the samples were Firmicutes, Bacteroidetes, and Proteobacteria. UC vs. CD was compared at the genus level, showing differences in the abundance of Faecalibacterium, Prevotella, and Roseburia, indicating that dysbiosis may be involved in IBD activity and that there may be differences between patients with CD and UC. A total of 14% of all the samples were *Blastocystis* positive; the positive samples had a more Prevotella-driven enterotype, while the *Blastocystis* negative samples had a more Bacteroides-driven enterotype. Although the changed microbial profiles did not exhibit consistent findings across previous studies, a common trait, namely lower bacterial diversity, surfaced in most of the IBD patients. A comparison of Illumina MiSeq and MinION sequencing concluded that there was little difference in the taxonomic resolution between Illumina MiSeq on higher taxonomic levels.

Abbreviations

BC	Barcode
BSA	Bovine Serum Albumin
bp	Base pairs
CD	Crohns disease
EDTA	Ethylenediamine tetraacetic acid
EtOH	Ethanol
F/B ratio	Firmicutes/Bacteroides ratio
GI	Gastrointestinal
IBD	Intestinal bowel disease
IBD	Irritated bowel syndrome
kb	Kilobases
LRS	Long-read sequencing
Mb	Megabases
NGS	Next-generation sequencing
ONT	Oxford Nanopore technologies
OTU	Operational taxonomic unit
PCoA	Principal Component Analysis
QC	Quality control
RT	Room Temperature
SCFAs	Short Chain Fatty acids
SGS	Second-generation sequencing
STs	Subtypes
SRS	Short-read sequencing
SSU rRNA	Small subunit ribosomal ribonucleic acid
TGS	Third-generation sequencing
UC	Ulcerative colitis

Table of Contents

1 Introduction

1.1 Background

The microbiota is the community of microorganisms that inhabits a specific environment such as the oral cavity, respiratory tract, and gut. The gut microbiome has a mutual symbiotic relationship with the host. It is a nutrient-rich environment where the microorganisms can reside, while the gut microbiota contributes to critical physiological processes and functions to keep the host healthy (Lavelle *et al.*, 2020). Under normal physiological conditions, the gut microbiota acts as a homeostatic unit involved in the fermentation of different complex undigested polysaccharide polymers, as well as the synthesis/production of short-chain fatty acids (SCFAs), certain vitamins, the conversion of energy, and keeping the intestinal mucosa's integrity in a normal homeostatic state (Rodríguez *et al.*, 2015). Some of the symbiotic microorganisms have been reported to have a special effect on the host's immune system, where they can be considered a key factor in immune homeostasis (Khan *et al.*, 2019). The microbiota will affect the organism it inhabits at both a physiological and pathological level. Many studies show the correlation between gut microbiota and human health. Gut dysbiosis is related to Inflammatory bowel disease (IBD), Crohn's disease, and ulcerative colitis (Vich Vila *et al.*, 2018). IBD is an increasing challenge for health care worldwide. It affects 3.5 million people, and the incidence of IBD is growing (Kaplan, 2015). IBD will affect the patient's life, work, and treatment, and research is a substantial cost for health care (Lavelle *et al.*, 2020). Over the past decade, IBD has become a global health challenge, as the number of reported cases is increasing worldwide. The cause of IBD is unclear and undetermined, limiting treatment, and the inflammations are often complicated and unknown. The treatments are often based on maintenance and relieving symptoms and pain rather than recovery and curing it. The lack of proper treatment of IBD is a factor in why the disease has become such a burden (Khan *et al.*, 2019).

The most common method for classifying the composition of a microbial profile is to use amplicon or short-read sequencing of nine highly variable areas within the bacterial 16S ribosomal rRNA gene. Recent studies focus on classification down to species-level by using next-generation sequencing, which provides greater insight into the composition of the gut microbiota. Improved DNA sequencing techniques have transformed the exploration of fundamental biological assessments and questions about evolution and how life itself works. As a result of improvements in sequencing technology and lower costs, more than 200,000

bacterial and archaeal complete or partial genomes have been uploaded to public databases since the first bacterial genome was entirely sequenced in 1995 (Zhang *et al.*, 2020).

1.1.1 16S rRNA and 18S rRNA

The 16S rRNA gene is widely used for studying microbial taxonomy. The 16S ribosomal RNA is used as a target gene because it is present in all prokaryotic cells and has distinct characteristics used for bacterial identification. The gene is part of the small ribosomal subunit and consists of nine variable regions as well as conserved regions (Figure 1) and has a size of roughly 1500 base pairs, making it an ideal target gene (Campbell, 2015).

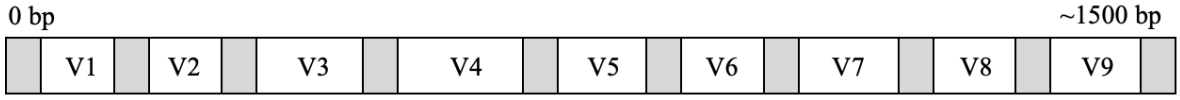


Figure 1. 16S rRNA gene (1500 bp). The gene consists of 9 variable regions (V1-V9) flanked by the conserved regions in grey.

The variable regions can be used to distinguish between different bacterial species due to their different evolutionary developments. The conserved regions make it possible to use universal primers, allowing them to bind to a variety of DNA templates for the identification of different microorganisms (Santos *et al.*, 2020). The v3–v4 region is most commonly used when assessing the gut microbiota for precise taxonomic differentiation (Wei *et al.*, 2020).

Similar sequence variations are typically grouped into operational taxonomic units (OTUs) when sequencing the 16S rRNA gene, with each cluster representing a taxonomic unit of a bacterial species or genus. For comparison with reference databases, OTUs are created to selected threshold lines (e.g., 97% or 99%) (Johnson *et al.*, 2019).

Several pipelines developed for sequencing of 16S rRNA gene is available, the following figure (Figure 2) provides a general overview of 16S rRNA gene next-generation sequencing (NGS) as well as shotgun metagenomics methods. Common for both methods, the process starts with DNA extraction from the microbial sample. The extracted DNA will then be either subjected to PCR amplification or sheared into smaller DNA fragments, for shotgun metagenomics. The 16S rRNA gene amplicons or the smaller sheared DNA fragments will then be sequenced using NGS techniques. The sequencing data is analyzed using a wide range of bioinformatics methods, allowing the investigation of taxonomic compositions as well as the functional capabilities of the tested sample (Boers *et al.*, 2019).

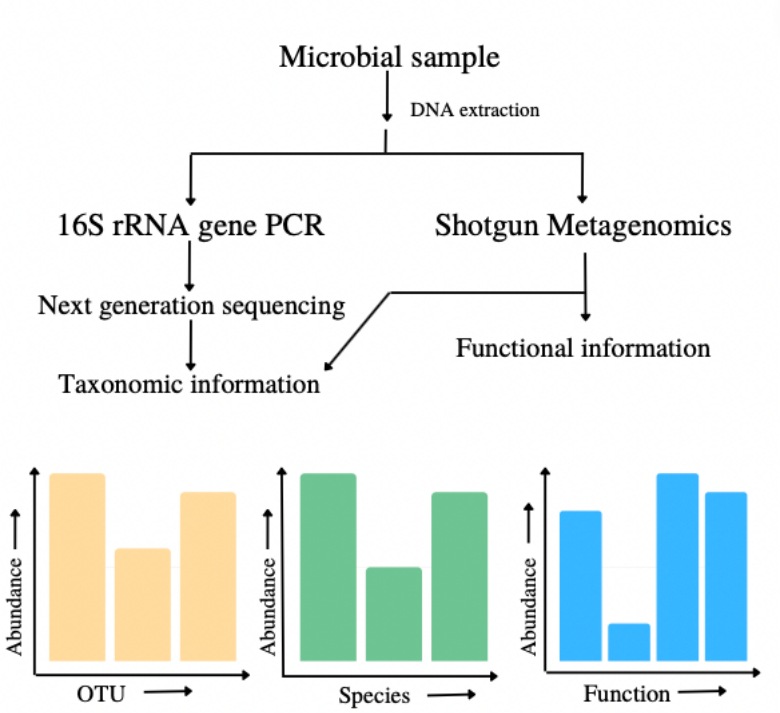


Figure 2. General overview of how microbial DNA gene can be analyzed using next-generation sequencing for taxonomic purposes using 16S rRNA, or for functional analyses using shotgun metagenomics. (Boers *et al.*, 2019)

18S rRNA is a DNA sequence that codes for the small subunit rRNA of eukaryotic ribosomes. The 18S rDNA sequence, like the 16S rDNA sequence, has conservative and variable sections (V1-V9, absence of V6) (Wu *et al.*, 2015).

The 18S rRNA genes are widely used in molecular analysis and phylogenetic studies. The 18S rRNA is great for sequencing due to its accessibility since the gene has highly conserved flanking regions which allow the use of universal primers. V4 is the most widely used and best choice for 18S rRNA gene analysis notes because it contains the most extensive database information and the greatest classification effect among the variable areas. The species distinctions among eukaryotic organisms in each sample are reflected by 18S rDNA sequencing (Black *et al.*, 2014).

One of the significant differences between performing analyses with 18S rRNA gene data and 16S rRNA gene data is the reference database used for OTU selection, taxonomic assignments, and template-based alignment building (Yeh *et al.*, 2021).

1.2 Inflammatory bowel disease

1.2.1 Etiology

IBD is a set of diseases affecting the intestines causing chronic inflammation of the bowels and the digestive tract. Both ulcerative colitis (UC) and Crohn's disease (CD) are grouped under IBD definitions (Khor *et al.*, 2011). IBD has become more frequent over the world in recent years, particularly in Asia's rapidly developing countries but also in Europe and the western parts of the world. Based on research and theories indicated that IBD may be a result of; i) defects in the intestinal epithelial and mucosal barrier, which can contribute to bacterial translocation within the gut; (ii) a microbial imbalance or dysbiosis, in the intestine, and: (iii) a cascade dysfunction in the intestinal inflammatory system, which can ultimately lead to a pathologic proliferation of important inflammatory cytokines. The increasing prevalence of IBD shows that one, if not more of these causative factors have become more prominent in recent years (Kaur *et al.*, 2011).

UC causes inflammation and ulcers on the inner lining of the large intestine due to aberrant immune system responses (Xavier *et al.*, 2007). UC can affect anyone at any age, however people between the ages of 15 and 30 are more prone to develop the condition. Symptoms of ulcerative colitis vary depending on the severity of the inflammation and where it occurs, UC is limited to the colon and rectum. The infected area of the colon is continuous with no patchiness, but ulcers that penetrate the inner lining of the colon are common. Some of the symptoms that occur in the early stage include diarrhea, abdominal pain, and fatigue. Progressed illness can cause cramping, fever, joint pain, and liver disease (Yu *et al.*, 2017).

CD is an inflammatory bowel disease that causes gastrointestinal pain, severe diarrhea, fatigue, weight loss, and malnutrition. The disease affects people of all ages. The signs and symptoms often appear in childhood or early adulthood. CD can affect any part of the GI tract, and is often showing patches of infection along the bowel (Xavier *et al.*, 2007). Both illnesses are long-term and are often aided by treatments to give patients a better life quality (Khor *et al.*, 2011).

Although many different inflammatory diseases affect the gastrointestinal system, most of them may be identified by a distinct underlying etiologic agent or process, as well as the nature and symptoms of the inflammatory activity. The causes of the most common types of inflammatory bowel disease, on the other hand, are unknown (Tamboli *et al.*, 2004).

The most common hypothesis for the cause of IBD is an exaggerated immune response which is triggered by environmental factors affecting the altered gut microbiota, or the invasion of pathogenic microorganisms to a susceptible host (Khor *et al.*, 2011). The alteration of the gut

microbiota raises the question of whether it is the cause of the inflammation or a consequence of it. The central role of the bacteria in this inflammatory disease remains unclear (Khan *et al.*, 2019). In terms of both quantity and immunological reactivity, some bacterial strains may be over-represented in IBD (Schirmer *et al.*, 2019).

The way microbes interact with the guts' mucosal immune compartments appears to play a key role in regulating immunity. Fusobacteria, Actinobacteria, Proteobacteria, Firmicutes, and Bacteroidetes are only a few of the phyla of bacteria. Firmicutes and Bacteroidetes are the most abundant phyla of bacteria, accounting for 90 percent of the gut microbiota (Stojanov *et al.*, 2020). People with IBD develop mucosal lesions because of an excessive or dysregulated immune response to commensal microorganisms in the intestines (Khan *et al.*, 2019). Abnormal microbial colonization of the gastrointestinal tract may be the source of such dysregulation, or dysbiosis, in those who have a genetic predisposition to IBD. Several metagenomic studies has been conducted on the intestinal microbiota. From all the studies conducted, some generalizations can be made regarding the microbiota of people with IBD, increased number of anaerobes, particularly gram-negative anaerobes (Kaur *et al.*, 2011).

Other studies associate patients diagnosed with IBD to have less diversity and lack some bacterial taxa, with Bacteroides, Firmicutes, Clostridia, Lactobacillus, Ruminococcaceae, and Gammaproteobacteria, and Enterobacteriaceae having increased abundances in the gut. (Zhang *et al.*, 2014). A ratio of Firmicutes and Bacteroidetes (F/B ratio) have also been observed and linked to several diseases in the gut due to dysbiosis. Dysbiosis is connected to a decrease in the Firmicutes/Bacteroides ratio, which is seen in IBD patients (Stojanov *et al.*, 2020). Certain intestinal strains can be overrepresented in IBD, for example, increased amount of pro-inflammatory microorganisms and a decreased proportion of anti-inflammatory microorganisms (Pigneur *et al.*, 2016).

Developments in gene-sequencing technologies, as well as increased availability of powerful bioinformatic tools, have enabled novel insight into the microbial composition of the human gut microbiota and the effect of microbial communities on human physiology and disease. Studies that used these technologies indicate that dysbiosis and decreased complexity of the gut microbial ecosystem are common features in patients with Crohn’s disease or ulcerative colitis (Manichanh *et al.*, 2012).

1.2.2 Treatment

The treatment of IBD usually consists of maintaining a state of remission rather than finding a cure. The most common type of treatment is to give patients corticosteroids, aminosalicylates, and immunosuppressive agents (Khan *et al.*, 2019). Anti-inflammatory medications, such as 5-aminosalicylic compounds and systemic corticosteroids, have traditionally been used to treat IBD. According to a recent systematic review and meta-analysis, antibiotics were shown to be effective in inducing remission in active Crohn's disease and ulcerative colitis patients, as well as avoiding recurrence in individuals with quiescent Crohn's disease, according to a recent systematic review and meta-analysis (Khan *et al.*, 2011).

All the treatments mentioned are short-term and will mostly help relieve symptoms and reduce inflammation. Side effects of treatment can be quite serious when treated with the above options, this includes loss of immune tolerance and drug resistance. Alternative treatment options include prebiotics, probiotics and symbiotics can be used as complementary- or alternative medicine to help treat IBD. Fecal microbial transplantation, FMT, can also be used as an alternative treatment for IBD (Pigneur *et al.*, 2016).

FMT is a novel therapeutic approach for rebalancing the gut microbiome (Fang *et al.*, 2018). This has been tried, along with probiotics, prebiotics, and symbiotic and bacterial consortium transplantation to regulate the gut microbiota (Khan *et al.*, 2011). Transplanted healthy donor fecal microbiota can improve the patent's gut microbiota to become non-dysbiotic (Bernstein, 2015).

1.2.3 The role of gut microbiota in IBD

In recent research, the role of the gut microbiota in IBD pathogenesis has been highlighted. The human gut microbiota is a diverse and dynamic environment of commensal bacteria.

The coexistence of the gut microbiota and the host indicates the importance of the microbial flora in host health and maintaining the gut microbiota equilibrium is critical for the host gut and overall systemic physiology, as shown in Figure 3 (Vich Vila *et al.*, 2018). Dysbiosis is defined as a shift in the steady-state structural composition of the gut microbiota that can disrupt microbial homeostasis and is linked to several gut diseases and intestinal inflammation. Many studies report dysbiosis as a direct cause of IBD (Khan *et al.*, 2019). A reduction in commensal bacteria diversity, notably in Firmicutes and Bacteroides, and a relative increase in bacterial species belonging to the Enterobacteriaceae family, is the type of dysbiosis most linked with IBD patients (Ancona *et al.*, 2021). Vich Vila *et al*. (2018) exhibit that strain- and species-level identification of the gut microbiome was necessary to identify gut disease-associated bacteria

(see Figure 3). They assessed the overall composition of the gut microbiome from individuals with both IBS and IBD before taxonomic classification and relative taxonomy abundance.

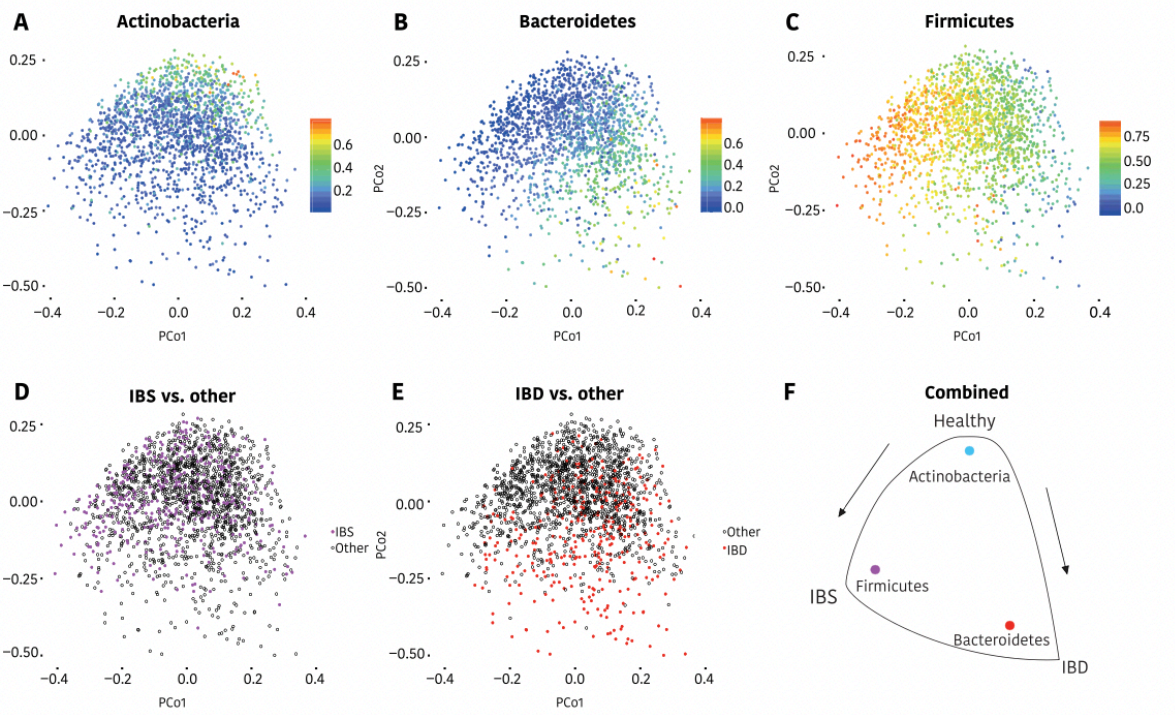


Figure 3. Principal coordinate analysis of Bray-Curtis dissimilarities showing the gut microbiome spectrum of 1792 human fecal metagenomes. Taxonomic endpoints were used to determine Bray-Curtis dissimilarities. The lowest nonredundant taxonomic level was used to designate endpoints. **(A), (B)** and **(C)** show the relative abundance of the three most abundant phyla – Actinobacteria, Bacteroidetes, and Firmicutes. The metagenomes of patients with IBS **(D)** or IBD **(E)** differed from the control patients. **(F)** displays controls or samples combined, and patients with IBD or IBS had more Actinobacteria in their feces than controls. IBS patients had more Firmicutes and fewer Bacteroidetes than healthy individuals. Patients with IBD, on the other hand, had fewer Firmicutes and more Bacteroidetes than controls (Vich Vila *et al.*, 2018).

Some studies show that there might be a link between gut microbiota and the development of IBD while others state that the correlation is between environment, genetics, and diets. General changes in the gut microbiome influence one's susceptibility to diseases like CD and UC (Ancona *et al.*, 2021). A lack of exposure to different microorganisms during childhood and early life along with living conditions and usage of antibiotics will affect the immune system's development and maturation of the gut microbiome (Khan *et al.*, 2011). This lack of exposure is discussed to be a leading factor in the loss of some negative regulatory pathways, which will ultimately lead to an over-active immune response. Discoveries made indicate that IBD is a polymicrobial disease where a combination of abnormal immune responses weakened intestinal mucosal barrier, and various gut microbial factors lead to dysfunctional host-microbial interactions (Schirmer *et al.*, 2019).

It is however unclear whether a specific individual bacterium or microorganism, or a group of a certain strain might be the cause of IBD (Nell *et al.*, 2010). Based on several studies, there is an indication that specific bacterial species/strains are common in IBD patients, however, none of these species or strains have been proven to be directly causative of IBD (Lavelle *et al.*, 2020). Around >90% of the healthy human gut microbiota belongs to four different phyla: Bacteroidetes, Actinobacteria, Firmicutes, and Proteobacteria. The human gut encounters many more microorganisms than any other part of the body, making the gut the most complex component of the immune system (Koboziev *et al.*, 2014).

1.2.4 Dietary influence on inflammatory bowel disease

The development of IBD is affected by complex interactions between changes in the intestinal flora, environmental changes, and some genetic properties. The role of diet has been an underestimated factor in the course of the disease (Owczarek *et al.*, 2016).

As many as 70% of patients with IBD have reported employing an elimination diet during remission and this kind of major change in habits and eating can affect not only the patient itself, but also family and social life (Zallot *et al.*, 2013).

It is well known that a diet consisting of high levels of fat and carbs combined with low amounts of fruits and vegetables will increase the risk of metabolic disorders/conditions. The diet plays a major role in the influence of microbiota in the gut (Tomasello *et al.*, 2014).

1.3 *Blastocystis*

Blastocystis is a genus consisting of single-celled, atypical, non-flagellated, anaerobic stramenopiles (Stensvold *et al.*, 2016). *Blastocystis* is the most common eukaryotic microbe infecting human intestines, with an estimated 1 billion people infected worldwide. Even though *Blastocystis* has been related to gastrointestinal disorders, its pathogenicity is still debated due to most carriers being asymptomatic (Gentekaki *et al.*, 2017). *Blastocystis* is a diversified species in terms of genetics. *Blastocystis* has been divided into many morphologically similar but genetically diverse lineages, based mostly on the sequences of their small subunit (SSU) ribosomal RNA genes (Maloney *et al.*, 2020). To date, up to 28 subtypes (STs) of the SSU rRNA gene have been isolated from the genus *Blastocystis*, and the approved subtypes have to meet the criteria of having SSU rDNA sequences that differ by 4% or more (Maloney *et al.*, 2020; Stensvold *et al.*, 2016). Only four of the nine known subtypes seen in

humans are common — ST1, ST2, ST3, and ST4. In studies that include subtyping, they make up over 90% of all human *Blastocystis* (Alfellani *et al.*, 2013).

Blastocystis is thought to play a functional role in bowel diseases such as IBS and potentially IBD, in addition to symptomatic similarities to several gut-related diseases (Scanlan, 2012). It can be challenging to research the impact *Blastocystis* might have on these diseases because of the absence of specific markers for both IBD and IBS. Both IBS and IBD share similar symptoms with parasitic infections, like diarrhea, fatigue, inflammation, and abdominal pain (Clark *et al.*, 2013; Scanlan, 2012).

Both direct and indirect diagnostic approaches have been established for most parasites. Approaches based on morphology (microscopy) and detection of DNA (generally PCR) or antigens are direct, whereas indirect methods are primarily focused on the detection of antibodies (Stensvold *et al.*, 2016). There has been a development in molecular methods, which can detect *Blastocystis* in genomic DNA, extracted directly from human stool samples (Stensvold *et al.*, 2016). For *Blastocystis*, the SSU rRNA gene is the best phylogenetic marker that is available to date, and the most used *Blastocystis* gene for sequencing. The gene has available reference sequences that include all reported subtypes (Maloney *et al.*, 2019).

1.4 DNA sequencing

1.4.1 The importance and development of DNA sequencing

The world of genomics has been revolutionized over the past 20 years after the development of first- and second-generation sequencing. The development of SGS (second-generation sequencing) technologies has led to the completion of major genetic projects like the human genome project (Park *et al.*, 2016). The demand for quicker, cheaper, and more efficient technologies has resulted in newer sequencing methods – NGS, or third-generation sequencing (TGS). SGS is a synthesis-based sequencing that relies on a Polymerase chain reaction (PCR) to amplify the targeted DNA template. PCR can often lead to bias and errors in technique which will ultimately affect the sequencing results. NGS on the other hand, uses single DNA molecules for sequencing, enabling real-time sequencing so the reads can be directly analyzed as soon as they pass through the sequencer (Lu *et al.*, 2016).

In recent years, sequencing technologies involving single-molecule sequencing without the need for amplification have been stepping out into the spotlight of genomic research.

NGS is also known as long-read sequencing and will produce long reads with an average of 10-20 kb, meaning it will detect longer fragments (Li *et al.*, 2021). Unlike SGS, the NGS technologies can produce average read lengths of up to several thousand bases and a maximum

read length of more than 100 kb. NGS enables the use of universal primers to amplify DNA from many different organisms all from within one sample (Peters *et al.*, 2018). This amplicon-based technique enables parallel processing of several samples during a single sequencing run and can read up to several hundred samples in a single run (Guardiola *et al.*, 2015). The amount of DNA reads from a sequencing run can reach the order of up to 20 billion sequencing reads per flow cell used. To maximize the number of targeted reads during a sequencing process, the specific DNA sequence derived from the gene of interest must be amplified using PCR. During the PCR process, the complementary strands will separate, allowing for the designed primers to amplify and bind to the targeted DNA segments and proceed with the production of nucleotide sequences, increasing the number of copied DNA molecules (Garibyan *et al.*, 2013). Although these technologies seems promising and have a lot of potential for quicker assembly and expanded application areas, one disadvantage is a rather high error rate of the sequencing. It used to be up to 40% error rate when NGS was newly developed, but as the technologies and data processing tools has improved much, the error rate has decreased. Pacific Biosystems (PacBio) and Oxford Nanopore Technologies (ONT) are currently the major players when it comes to the development of third-generation sequencers, using Single Molecule Real-Time-, and nanopore sequencing respectively (Bleidorn, 2016). Several studies have shown that the data produced by the MinION was accurate enough to generate a consensus sequence from a single species sample for species identification with >99% accuracy (Vasiljevic *et al.*, 2021).

1.4.2 The use of third-generation sequencing vs next-generation sequencing

Identification of bacteria at the species level can be crucial in several situations like disease outbreaks in hospitals or contamination of food and water (Boers *et al.*, 2019). In these cases, rapid and accurate identification of species is key.

Park et al. demonstrate how the NGS compares to the previous sequencing technologies in both efficiency and cost, and how the parallel DNA sequencing methods have opened a whole era of genomics and molecular biology (Park *et al.*, 2016).

The setback of first- or second-generation sequencing, or short-read sequencing (SRS), is the obvious limitation to the DNA's size. All SGS cannot sequence longer stretches of DNA than about 400 bp. To be able to use SGS to sequence whole genomic DNA one must fragment the targeted DNA and amplify it in clones of between 75 to 400 base pairs. The shorter sequences will later be spliced by computer programs to form a contiguous sequence for further analysis. One of the necessary steps in SRS is PCR to amplify the DNA, SRS may often fail to create sufficient overlapping sequences from the DNA fragments (Adewale, 2020).

Long read sequencing (LRS), or TGS, will allow amplification of the full-length gene, providing a more realistic representation of the taxa in a sample. Due to the poor sensitivity of the nanopore and the inability to control the DNA's translocation speed through the pore, the sequencing outcome will result in high error rates. Despite these high error rates of the nanopore sequencing, the overall increased read length and full-length 16S rRNA gene sequencing allow for species-level classification which ultimately leads to improved taxa resolution compared to previous technologies (Ciuffreda *et al.*, 2021).

There are many uses for NGS and TGS, and even though the technology itself is well developed, there are continuously new methods and protocols being produced and published for metagenomic studies. Some of the experimental usages of NGS include *de novo* genome assembly, measuring genetic variations by using an already existing reference genome, analyzing transcriptome results, and determination of methylation of CpG dinucleotides, to mention a few (Park *et al.*, 2016).

1.5 Oxford Nanopore Technology and the MinION sequencer

Although the concept of using nanopores for sequencing was introduced back during the 1990s, it had its breakthrough when a functional sequencing platform using nanopores was developed after several important discoveries within the technology (Feng *et al.*, 2015). ONT developed the nanopore sequencer, which is used for TGS or NGS in 2012, but it was first released for commercial use in 2014 (Jain *et al.*, 2018). In contrast with other sequencing technologies, this does not involve any form of fluorescence or synthesis, but directly detects the sequence of ssDNA molecules in real-time and it is defined as an NGS sequencing device based on its single-molecule sequencing ability (Slatko *et al.*, 2018). The MinION device is palm-sized and runs individual DNA or RNA molecules through a nanopore, meaning that only a single strand of nucleic acid can pass through it. The four bases, A, T, C, and G, all have different electrical properties, so the electrical signals for each specific base can be detected using the MinION. Whereas older sequencing technologies relied on sequencing by synthesis, the MinION uses an ionic current, which is passed through the flow cell. The different nucleotide bases are recognized by the variations in currents as they travel through the nanopores (Petersen *et al.*, 2019). The initial stage in NGS is library preparation. It enables DNA or RNA to adhere to the sequencing flow cell and to identify the sample. Preparation of the sequencing library includes ligation of adapter sequences along with a motor protein and a hairpin adapter which allows for double-stranded DNA to be sequenced.

The sample is loaded into a flow cell, where the nanopores and sensors are located, before loading the flow cell into the device and starting a new sequencing run. Barcodes and adaptors used for each sample are added during the library preparation step before loading the flow cell. The flow cell itself is an array with approximately 512 sensors controlled by an application-specific circuit. Each of these sensors is connected to four nanopores, even though only one nanopore per sensor is active at any given time. Flow cells can only be reused a certain number of times before the integrity of the pores deteriorates (Deamer *et al.*, 2016).

The adaptors will interact with the proteins attached to the nanopores allowing the DNA strand to enter the pores. Each of the pores has an ionic current that is disrupted when a single nucleotide passes it, which creates a "squiggle" plot as shown in Figure 4. Each squiggle is created by disrupting the ionic current. By decoding each alteration of the current, one can identify the molecule, in this case, the bases, that disrupted the ionic current (Plesivkova *et al.*, 2018).

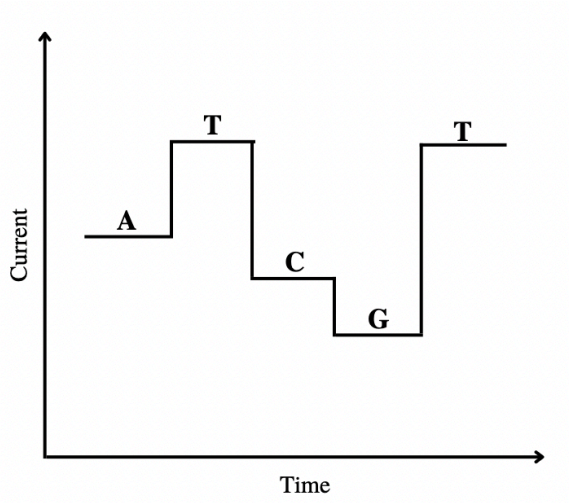


Figure 4. Squiggle plot generated by the ionic currents from each nucleotide as they are sequenced. A Squiggle plot is created as the different nucleotides pass through the ionic current of the nanopore (Plesivkova *et al.*, 2018).

The MinION can sequence up to thousands of base pairs in one single run, meaning that the whole 16S rRNA genome can be sequenced in real-time. One of the major issues with using the MinION is the high error rates. The error rates were up to 15% per base, but due to development of newer protocols for library preparation and post sequencing analysis over the past years, the error rate has improved (Quainoo *et al.*, 2017). MinION generates reads up to 883 kb. This feature contributes to creating better insight into the complexity of genomes. ONTs MinION can create long reads which also has applications outside of genome studies as it can be used for sequencing full-length genes used in epidemiological studies and taxonomic studies

(Ciuffreda *et al.*, 2021; Matsuo *et al.*, 2021). The MinION also generates data in real-time, meaning that the read results are available as soon as each base pass through the nanopore. This has a significant clinical impact since it allows one to identify problems or abnormalities and, if necessary, treat them sequentially as the sequencing occurs. The combination of real-time data generation with MinION's compact size, low cost, quick and easy library preparation due to the lack of amplification PCR, and portability gives it a significant advantage over other NGS devices. In a study conducted by Leggett *et al*. (2020) they used the MinION sequencer to perform shotgun metagenomics to profile mock communities and fecal samples from healthy and ill preterm infants. They were able to classify a 20-species mock community and capture the diversity of the gut microbiota of the infants. The response to treatments such as probiotic supplementation, antibiotic treatment or episode of suspected sepsis was all reported using the MinION (Leggett *et al.*, 2020). Their results showed that the MinION along with real-time analysis software (NanoOK RT) could process metagenomic samples to a rich dataset in under 5 hours. This sort of study creates a base for future studies, emphasizing the importance of developing these tools and approaches for rapid feedback that provides information about tailored patient treatment options. The absence of an amplification PCR step during library preparation is a crucial component in reducing sequencing biases (Madoui *et al.*, 2015; Matsuo *et al.*, 2021).

1.6 Bioinformatics and data analysis

After running a sequencing run with the MinION, the raw data can be error corrected before being assembled into contigs. Because prior software methods did not perform well with the long, error-prone NGS reads, many different software packages have been developed specifically for processing Nanopore or NGS data (Lu *et al.*, 2016). It is critical to have access to the necessary analytic tools as sequencing technology advances, delivering both larger amounts of data, and more complex data due to the increase in read length. When analyzing data from the sequencing of bacteria, operational taxonomic units, or OTUs, are used. OTUs are based on similarities between the different strands of bacteria and the DNA strands. If two sequenced bacteria have a 16s rRNA gene that is 97% or more alike, they are said to be the same OTU. If the similarity is less than 97%, they are two different OTUs and can then represent two different bacterial species (Johnson *et al.*, 2019).

1.6.1 Diversity

The mean diversity of different species in different habitats or sites on a local scale is what alpha diversity (α-diversity) displays. Because numerous perturbations might impact alpha diversity. Summarizing and comparing alpha diversity is a common strategy in community assessment (Willis, 2019). Analyzing alpha diversity of any amplicon sequencing data is a common preliminary stage in determining differences among microbial environments in terms of microbial ecology. The simplest measure of alpha diversity is species richness, which is the count of the number of species or OTUs present in each area (Kim *et al.*, 2017; Thukral, 2017). There are several various indices that consider the abundance or frequencies of OTUs in a sample. Some of the most common alpha diversity indexes are Shannon, observed, Simpson, Chao, and ACE Index. The simplest measure is richness, the number of species (or OTUs) observed in the sample (Thukral, 2017; Willis, 2019).

The ratio of alpha diversity to regional diversity is known as beta diversity (β-diversity). It's the difference in species richness between two environments or the measure of how similar or unlike two regions are. Beta diversity can be described as the differences between communities, or their dissimilarities. There are more than 17 indices for measuring beta diversity and some of the most used are Bray-Curtis, Jaccard, and Sorensen (Schroeder *et al.*, 2018).

1.6.2 Bioinformatics

Bioinformatics is the branch of biology that deals with the storage, analysis, and interpretation of experimental data. Some of the bioinformatics tools are computers, databases, and the statistical tools and algorithms used for data analysis (Twyman *et al.*, 2006). The purpose of bioinformatics is to extract information and identify relationships between datasets. The datasets are often nucleotide or protein sequences, protein structures, gene-expression profiles, molecular weights, or biochemical/metabolic pathways (Gauthier *et al.*, 2018; Twyman *et al.*, 2006). The basis of sequence comparison is the ability to align two sequences and determine the number of shared residues. The result is an alignment score, which represents the quality of the alignment. For nucleotide sequences, comparisons are always made based on sequence identity, which is the percentage of identical residues in the alignment. For protein sequences, identity can be suitable for the comparison of very closely related sequences, but a more useful measure is sequence similarity, which considers conservative substitutions between chemically or physically similar amino acids (Twyman *et al.*, 2006).

The 16S ribosomal RNA gene has been sequenced for identification and analyzing bacterial communities since the 1970s. This method focuses on the 16S rRNA gene, which has highly conserved and highly varied (hypervariable) sections across bacterial species. The highly conserved regions enable the design of universal PCR primers to target and amplify the 16S rRNA sequence. Because of these characteristics, 16S rRNA sequencing techniques can capture almost all the bacteria in a microbial community, which can then be compared to huge 16S rRNA databases to identify their identities (Lu *et al.*, 2020). The species, genera, and even phyla present in many metagenomic samples are mostly unknown at the time of sequencing, and the purpose of sequencing is to properly establish this microbial composition (Wood *et al.*, 2014).

1.6.3 Basecalling

When sequencing, DNA fragments pass through a detector, in this case, a nanopore, and generate a signal. The identity of the nucleotides can be assigned by the information passed through the detector by base-specific dye or voltage which the nucleotide emits. This information contains the production of four traces of signal, which corresponds to each of the four bases over the length of a sequencing run. By using tools and algorithms, the four unique traces can be converted into the actual sequence of nucleotides. This process is known as basecalling (Twyman *et al.*, 2006).

Basecalling can directly impact the quality of the resulting sequencing, ideally, the traces would be free of noise and bias, and all the peaks would be evenly distributed, of equal height, and have a Gaussian shape. But the peaks have variable spacing, height, and shape. Basecalling is known for being error-prone for these reasons, and for accurate sequence assembly, it is important to give an estimate of quality for each of the assigned bases (Rang *et al.*, 2018). The rapid succession and development of basecallers demonstrate that their performance is a key factor in the quality of the base pair sequence retrieved from the raw signal (Bože *et al.*, 2017).

1.6.4 Taxonomy

An important phase in metagenomics research is determining the taxonomic entities present in a sample using a metagenomic sequencing dataset. Due to the growing utilization of high technologies, more accurate and efficient methods for metagenomic investigations are required (Boers *et al.*, 2019; Ciuffreda *et al.*, 2021). Assigning taxonomic labels to sequencing reads and inferring the composition of a microbial community are becoming more attractive study fields (Jain *et al.*, 2018). Because of the larger information content contained in the sequence, long

reads often allow better taxonomic and functional analysis than short reads. However, most widely used metagenomic classification tools or pipelines rely on algorithms based on short reads, which do not scale well with long-read datasets (ranging from 13 kb to 2 Mb) and do not account for the higher error profile of nanopore reads by default (Ciuffreda *et al.*, 2021). The increase of long-read datasets in sequencing studies along with the constant updates of bioinformatics software has offered a lot of information about the performance of metagenomic tools using nanopore readings. Error corrections and methods to solve read error-related challenges are constantly being developed (Adewale, 2020).

Many computational genomics workflows for metagenomics research include assigning taxonomic labels to sequencing data. Several approaches to completing this work in a timely manner have emerged in recent years. Kraken 2 is the most recent version of Kraken, a taxonomic classification system that provides high accuracy and rapid classification speeds by employing precise k-mer matches (Lu *et al.*, 2020). This classifier compares each k-mer, a k-mer is a substring of length *k* in each given string, in a query sequence against the lowest common ancestor (LCA) of all genomes that include the k-mer in question. The classification method is informed by the k-mer assignments (Wood *et al.*, 2019).

Kraken 2 improves on Kraken by lowering memory use by 85%, allowing for the use of larger amounts of reference genomic data while retaining great accuracy and speed. In addition, Kraken 2 has a translated search mode, which improves sensitivity in viral metagenomic studies (Kibegwa *et al.*, 2020). Wood *et al*. found that Kraken 2 showed superior accuracy to other nucleotide classifiers and an increase in processing runtime and memory requirement. As the amount of assembled genome databases grows, so will the number of reference sequence databases used in metagenomics studies. The readings from a 16S rRNA sequencing experiment are usually compared with a reference database to identify the bacterial population. Greengenes, NCBI, SILVA, and RDP are the some of the standard 16S rRNA databases, each with somewhat different content (Cole *et al.*, 2014; DeSantis *et al.*, 2006; Quast *et al.*, 2013).

Kraken and Kraken 2 uses a custom-built database to attempt to assign a taxonomic label to every read in a metagenomics sample. Although several accurate methods for aligning a sequence read to a database of microbial genomes have been developed, this step alone is insufficient to assess how much of a species is present in a sample (Dilthey *et al.*, 2019; Lu *et al.*, 2017). When closely related species are present in the same sample, a situation that frequently occur because many reads align as well to many species. To fix this problem, a separate abundance estimation algorithm is needed. Bracken (Bayesian Reestimation of

Abundance After Classification with KrakEN) goes beyond just simply classifying the individual reads and assign the abundance of species, phylum, or other taxonomic categories from the sequences. Bracken can estimate species abundance in metagenomics by re-distribute reads in the taxonomic tree/rank. Reads that are assigned to nodes above species level will be distributed down to species level, and the reads assigned to strain level are re-distributed up to their parent species (Lu *et al.*, 2017). Bracken estimates the number of reads originating from each species present in a sample using the taxonomy labels provided by Kraken, a highly accurate metagenomics classification method (Sun *et al.*, 2021).

1.7 Aim of the study

The cause of inflammatory bowel disease is most likely a combination of microbial and environmental factors. A dysbiosis in the gut can contribute to a variety of gut diseases and is often linked with IBD (Buttó *et al.*, 2016). The main objective of this study is to assess the presence of microbial eukaryotes in the human intestinal microbiome and to evaluate the composition and abundance of microorganisms in patients with IBD. Total DNA will be extracted and isolated from stool samples before preparing it for third-generation sequencing. Targeted microbial eukaryotes PCR will be conducted on isolated DNA as well as primer construction for 16S rRNA targeted sequencing and primer design for Oxford Nanopore technologies. Subgoals for this study is to optimize and evaluate DNA extraction protocols to obtain high quality genomic input for sequencing.

This study was designed to assess the process of third-generation sequencing, and to prove the efficacy of TGS in assessing the microbial composition of IBD patients' stool samples.

2 Materials and Methods

2.1 Biological material

A total of 61 stool samples from 55 patients were received from Stavanger University Hospital (SUS, Helse Stavanger HF, Rogaland) from August 2021 to March 2022 for further analysis and sequencing. The samples are from patients included in a clinical trial by SUS (SUSI), the trial aims to study the outcomes of protocol-based handling of newly diagnosed IBD patients. The primary outcome of the trial is to look at the clinical efficacy of IBD drug therapy. The second outcome measures are to correlate fatigue in patients with coeliac disease, fatigue in IBD and the intestinal microbiome of patients with IBD. Inclusion criteria for this trial is that patients must be newly diagnosed with IBD, and the study is eligible for patients of all sexes in the ages 16-80 years. Exclusion criteria for patients is previous IBD with specific treatment within 10 years, inability to consent and inability to adhere to the treatment protocol. All patients included in the trial have been diagnosed with either Ulcerative colitis or Crohns disease, meaning the patients had clinical signs of disease. Samples were taken at different time stamps, V0 indicates the time of diagnosis, and V3, V11, and V60 are 3, 11, and 60 months after diagnosis, respectively.

2.2 Other materials

2.2.1 Prepared solutions

10 mM Tris-HCl pH 8.0 with 50 mM NaCl.

A volume of 750 µL of 1 M NaCl, and 150 µL of 1 M Tris-HCl was combined in a 15 mL Falcon tube. The volume was adjusted to 15 mL using 14.1 mL of nuclease-free water. The final product is a 15 mL solution with 50 mM NaCl along with 10 mM Tris-HCl.

10x Tris-acetate EDTA buffer (TAE buffer).

In a conical flask, 48.5 g of Tris base was dissolved in 800 mL of deionized water along with 11.4 mL of glacial acetic acid and 20 mL of 0.5 M EDTA (pH 8.0). Deionized water was used to adjust the volume to 1 L as the final volume. The final product was 1L of 10x Tris-acetate EDTA buffer.

Ammonium acetate.

38.5 g of ammonium acetate was dissolved in 50 mL of deionized water before the solution was filtered and stored in a sterilized bottle at RT. The final product was 50 mL of 10 M ammonium acetate.

2.2.2 Kits and other reagents

All kits used for this project are listed in Table 1, and the different reagents are listed in Table 2. The reagents and kits listed in Table 2 were used for DNA extraction, DNA purification, PCR, and gel electrophoresis.

Table 1. Kits used for the laboratory work.

Kit	Manufacturer	Production site	Cat. no	Use
Fast DNA stool kit	Qiagen	Hilden, Germany	51604	DNA extraction
Genomic DNA Clean & concentrator	Zymo research	California, USA	D4064-D4065	DNA purification
1x dsDNA High sensitivity assay	Thermo Fisher	Waltham, Massachusetts	Q33231	DNA quantitation
16S barcoding kit, 1-12	Oxford Nanopore technologies	Oxford, UK	RAB204	Library preparation
16S barcoding kit, 1-24	Oxford Nanopore technologies	Oxford, UK	SQK-16S024	Library preparation
Flow cell loading kit	Oxford Nanopore technologies	Oxford, UK	RAB204/ SQK-16S024	Loading sample on flow cell
Flow cell Wash kit	Oxford Nanopore technologies	Oxford, UK	EXP-WSH004	Washing of the flow cell

Table 2. Materials and reagents used besides the kits included in the previous table.

Material/reagent	Manufacturer	Production site	Cat. no	Use
Sterile zirconia beads, Ø 0,1mm	BioSpec	Bartlesville, USA	11079101z	Sample preparation
Lysis buffer (ASL buffer)	Qiagen	Hilden, Germany	51604	Cell lysis
Ammonium acetate	VWR	Geldenaaksebaan, Leuven, Belgium	153164R	DNA extraction
Ethanol (abs.)	Supelco	Darmstadt, Germany	64-17-5	DNA extraction, purification, sequencing
AMPure XP magnetic beads	Beckman Coulter	California, USA	A63881	DNA purification, library prep.

DNase-free RNase	Thermo Fisher scientific	Massachusetts, USA	89836	DNA extraction
Agarose, low EEO	Sigma	Darmstadt, Germany	A0576-100G	Gel electrophoresis
GelRed Nucleic Acid Strain, 10 000X in water	Biotium	Oslo, Norway	41003	Gel electrophoresis
HyperLadder (500 lanes) and 5X DNA Loading Buffer, blue.	Meridian Bioscience, Inc	London, UK	BIO-33026	Gel electrophoresis
DreamTaq PCR Master Mix (2X)	Thermo Fisher Scientific	Massachusetts, USA	K1071/K1072	PCR
KAPA HiFi HotStart ready mix	Roche sequencing	California, USA	N/A	PCR
LongAmp Taq 2X master mix	New England BioLabs	Massachusetts, USA	M0287S	PCR
Bovine Serum Albumin	Thermo Fisher scientific	Massachusetts, USA	B14	PCR

2.3 Sample preparation

150 to 200 mg of frozen stool samples were added, using a biopsy punch, to a 2 mL tube, containing sterile zirconia beads (Ø 0,1mm). The frozen samples were either used for further extraction or kept frozen at -80 °C until further use.

2.4 DNA extraction and quantitation

2.4.1 DNA extraction using Protocol Q

A modified extraction protocol suggested by Costea *et al.* (2017) and widely used for metagenomic studies, along with the Fast DNA stool kit from Qiagen was followed and Figure 5 shows the workflow used for sample preparation, DNA extraction, and library preparation for sequencing (Costea *et al.*, 2017).

1 mL of ASL lysis buffer was added to homogenize the stool samples containing sterile zirconia beads before incubation at 95 °C for 15 minutes. The cells were mechanically lysed using TissueLyser LT (Qiagen) before centrifugation at 16000 x *g* for 5 minutes at 4 °C. Supernatants were transferred to new tubes. The residual pellet was homogenized in 300 mL Lysis buffer and once again incubated at 95 °C for 15 minutes followed by centrifugation at 16000 x *g* for 5 minutes at 4°C. The supernatants were pooled in the new 2 mL tube. A volume of 260 µL of 10M ammonium acetate was added to each lysate tube, before mixing well and incubating on ice for 5 minutes. The lysate tubes were centrifuged at 16000 x *g* for 10 minutes at 4 °C. The supernatant from each sample was transferred into two 1.5 mL Eppendorf tubes before one

volume of isopropanol was added to all samples and incubated on ice for 30 minutes. The chilled samples were centrifuged at 16000 x *g* for 15 minutes at 4 °C before discarding the supernatant and washing the pellet with 0.5 mL of 70% EtOH letting the pellet dry for 5-10 minutes. The pellet was then dissolved in 100 µL of TE (Tris-EDTA buffer) and the two aliquots were pooled in a new tube. To ensure no RNA contamination in the sample, 2 µL of DNase-free RNase (10 mg/mL) was added to the samples and incubated at 37 °C for 15 minutes. A total of 200 µL of AL buffer and 15 µL of proteinase K (Qiagen) were added to the samples, mixed by vortex, and incubated at 70 °C for 10 minutes. 200 µL of ethanol (96-100%) was added to the lysate before samples were transferred to a QIAamp spin column (Qiagen) and centrifuged at 16000 x *g* for 1 minute at RT. The flow-through was discarded before adding 500 µL of wash buffer AW1 (Qiagen) and repeating the centrifugation for 1 minute. This was repeated using wash buffer AW2 (Qiagen). The columns were dried by adding the columns to new, dry 2 mL tubes and centrifugation for 1 minute at 16000 x *g* at RT. To retain the DNA after washing, 75 µL of AE buffer (Qiagen) was added to the column before spinning the sample down for 30 seconds, 16000 x *g* at RT. This step was repeated, so the final eluate with the retained DNA was 150 µL of extracted DNA. All samples were quantified using NanoDrop (NanoDrop One, Thermo Scientific, Waltham, Massachusetts) and Qubit fluorometer, using 1X dsDNA high sensitivity assay (Invitrogen, Waltham, Massachusetts) to ensure good quality genomic material. Some of the samples were run on a 1% agarose gel to determine fragment lengths and quality. The genomic DNA was stored at -20 °C until further use.

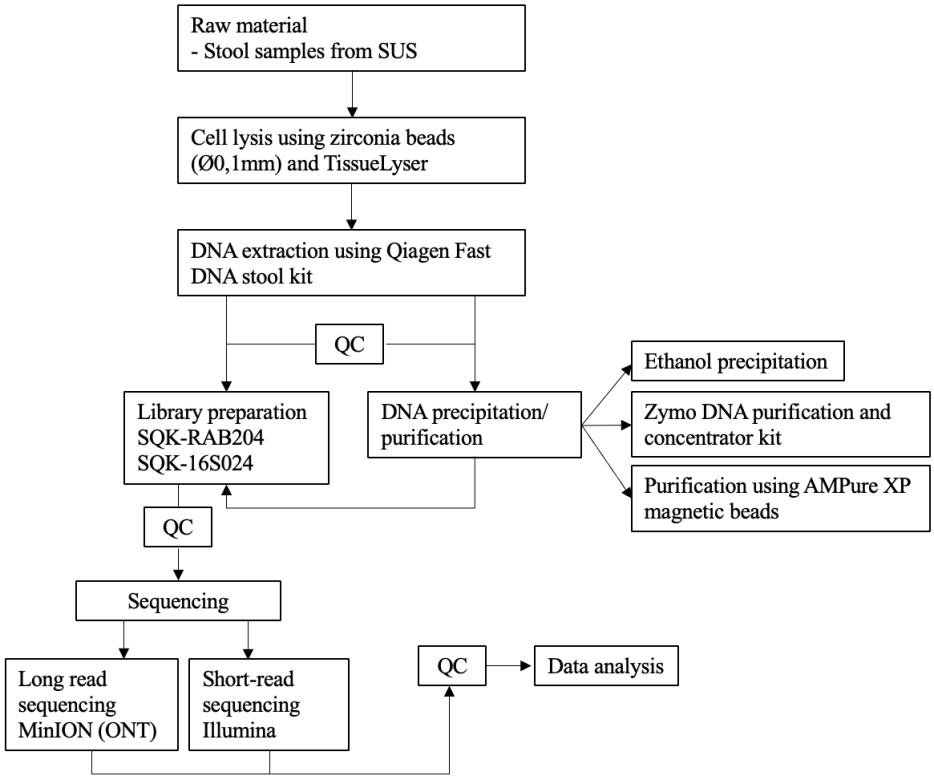


Figure 5. Flowchart for the general workflow from sample collection to sequencing and data analysis. Extraction was performed using cell lysis by bead beating and a DNA extraction kit from Qiagen. All samples were quantified using NanoDrop and then purified if necessary. Library preparation was performed using barcoding kits from Oxford Nanopore technologies before starting sequencing runs and further data analysis.

2.4.2 Nanodrop protocol

The lower and upper pedestal of the NanoDrop One was cleaned properly before use by applying a small amount of deionized water and wiping it off carefully using laboratory wipes (lint-free, VWR). The AE buffer (Qiagen) was used as a blank before loading 1-3 µL of sample onto the lower pedestal and measuring.

2.4.3 Qubit protocol

The Qubit 1X dsDNA high sensitivity assay kit is an easy and accurate quantitation method. To prepare the standards before measuring, 10 µL of each of the Qubit standards was added to a clean 0.5 mL Qubit assay tube, before adjusting the volume to 200 µL using the 1X dsDNA working solution. For measuring, 1-2 µL of each sample was added to a new Qubit assay tube before adjusting the final volume to 200 µL using the working solution. The samples were vortexed quickly before incubating at RT for 2 minutes.

2.5 DNA purification

After the samples were isolated, DNA purification was performed to remove unwanted salts, proteins, or reagents as high-quality genomic material is crucial for further analysis and sequencing. All samples were measured using either NanoDrop or Qubit fluorometer before and after DNA purification.

2.5.1 Zymo purification and DNA concentrator kit

A volume of 70 µL of the extracted DNA was mixed with 140 µL of the ChiP DNA binding buffer and added to a spin column provided in the kit (Zymo Research). The tubes were spun down at RT for 0.5 minutes at 16000 x *g*. The flow-through was transferred to a new tube, not discarded. The wash buffer was then added to the column, and the spinning the sample again for 1 minute, 16000 x *g* at RT and discarding the flow-through and repeating the washing step one more time. The spin column was transferred to a new 1.5 mL Eppendorf tube, along with 70 µL of elution buffer to the spin column before centrifugation at 0.5 minutes under the same conditions. The spin-column was discarded, and the purification was complete.

To make sure the purification was optimal, the elution buffer was heated to 60 °C, and the purified DNA was added directly to the matrix one more time before centrifugation one last time for 1 minute, 16000 x *g* at RT.

2.5.2 Ethanol precipitation

Day 1 – Ethanol precipitation was performed by adding 23 µL of 10 M Ammonium acetate to 24 µL of the sample. 52.8 µL Ice cold ethanol (prepared sometime before the precipitation) was added to the DNA solution before mixing well and incubating at -20 °C overnight.

Day 2 – The sample was transferred directly from the freezer to centrifugation at 15000 x *g* for 30 minutes at 4 °C. The supernatant was discarded before washing the pellet with 200 µL of 70% ethanol and spinning it down at 15000 x *g* for 30 minutes at 4 °C. The wash step was repeated. The supernatant was discarded before drying the pellet for 5-20 minutes until all ethanol had evaporated. The pellet was redissolved in 20 µL of buffer AE.

2.5.3 Purification using AMPure XP magnetic beads

A 1:1 ratio of the sample (50 µL) and AMPure XP magnetic beads (50 µL) (Beckman Coulter, Brea, California) were mixed before incubating on a rotator mixer for 5 minutes at RT. The sample containing the beads was spun down briefly before pelleting on a magnet. Keeping the tubes on the magnet, the supernatant was pipetted off and discarded before washing the remaining pellet with fresh 70% ethanol. The washing was repeated before drying the pellet for 30 seconds to ensure no ethanol residue. As a final step before retaining the purified DNA, 40-

50 µL of buffer AE was used to dissolve the pellet before placing it back on the magnet and retaining 40-50 µL of eluate in a new tube and disposing of the pelleted beads.

2.6 Primer preparation and testing

Polymerase chain reaction, or PCR, is a commonly used method for the amplification of a DNA sequence. The amplification relies on using flanking primer sequences. The primer sequences are often 15-20 nucleotides long. One primer sequence is complementary to one end of the target sequence on one strand; the second primer is complementary to the other end of the sequence on the other strand. PCR is a fast and very specific method of amplification, where only a small amount of DNA needs to be present, as long as a few molecules contain the complete target sequence. The PCR method is a three-step cycle that is typically repeated up to 20-35 times.

Denaturation of the double-stranded DNA template is the initial stage in the cycle. This is normally completed at a temperature of 95 °C. After the denaturation, annealing begins, in which the primers anneal to the complementary sequence. The melting temperature of the primers is commonly used to calculate the annealing temperature, which is normally between 50 and 55 °C. The next step is the extension, which involves adding nucleotides to the 3' end of the primers using a DNA polymerase. The temperature to apply in this stage varies depending on the DNA polymerase being used, but it's normally about 70 °C. After the last cycle, the last elongation step of 5-15 minutes is frequently undertaken to verify that single-stranded DNA is completely stretched. To assess the outcome of the PCR, gel electrophoresis is often used to determine the size of the amplicon. Gel electrophoresis is made of a polymer that works as a filter where nucleic acids or proteins can pass through and be separated depending on size, electrical charge, or other properties.

To be able to perform library preparation before sequencing, primers specifically designed for ONT technologies need to be added to each DNA sample, as they contain the unique barcodes for each sample. The ONT-specific primers are constructed to have overhangs on each side of the primer sequences. Both the ONT-specific and -unspecific primers were tested to see if they worked with genetic material suited for 16S- and 18S-specific PCR.

The following primers listed in Table 3 were used for the 16S and the 18S, both with and without the overhang of ONT.

Table 3. Primers used for the different genomic materials. Primers are both specific and unspecific for ONT sequencing. For the ONT-specific primers, the overhangs are highlighted in bold.

Sequence ID	Overhang	Sequence
16S-ONT	Primer with overhang	F: **TTTCTGTTGGTGCTGATATTGC**AGAGTTTGATCMTGGC R: **ACTTGCCTGTCGCTCTATCTTC**TACCTTGTTACGACTT
18S	Primer without overhang	SSU-F: AACCTGGTTGATCCTGCCAGTAGTC SSU-R: TGATCCTTCTGCAGGTTCACCTACG
18S-ONT	Primer with overhang	F: **TTTCTGTTGGTGCTGATATTGC**AACCTGGTTGATCCTGCCAGTAGTC R: **ACTTGCCTGTCGCTCTATCTTC**TGATCCTTCTGCAGGTTCACCTACG

10 different PCR setups were constructed to optimize the conditions for the primers. The primers had to be adjusted according to temperature, concentration, and different sources of genomic material used to perform the PCR.

The initial reaction used DreamTaq PCR master mix (2X) for all reactions, both 16S, and 18S. Later trials used KAPA HiFi HotStart Ready-mix master mix and LongAmp Taq 2X master mix. DNA from sheep lungs, human pancreas, human stool samples, and bacterial colonies was used for the different trials, depending on which primers to use.

The first PCR trial was done using DNA from a bacterial colony and lungs from sheep. Both the bacterial DNA and the sheep DNA were diluted to a final concentration of 50 ng/µL. In a clear 0.2 mL PCR tube, the components for PCR were mixed as shown in Table 4.

Table 4. General overview of reagents used for all PCR reactions. Some volumes may have been adjusted for later trials.

Reagents (for 0.4 µM)	Volume
Master Mix	12.5 µL
Primer Mix (0.4 µM) (forward primer + reverse primer)	1 µL
DNA sample (50 ng/µL)	1 µL
water	10.5 µL
Total volume	**25 µL**

For some of the trials, BSA (bovine serum albumin) was added to enhance PCR yield. In the samples with the addition of BSA,1 µL of BSA was added and 1 µL of water was removed to keep reaction volume at 25 µL. The different trials are all listed in Table 7, which displays an overview of all trials and the conditions used or changed.

For each of the primers, several duplicates and different temperatures were tested to see if the primers were sensitive to different concentrations or temperatures. The gradient PCR provided

useful insight into what annealing temperature would be optimal for 18S (68 °C) and 16S (50 °C). PCR conditions were altered between 16S- and 18S rRNA amplification,

Table ***5*** and Table 6 shows what PCR conditions that were used for 16S and 18S after determining the optimal temperature for the specific primers.

Table 5. PCR setup for 16S rRNA amplification using 16S-specific primers.

Cycle step	**Temperature (°C)**	**Time**	**No. of cycles**
Initial denaturation	95	1 minute	1
Denaturation	95	20 second	25x
Annealing	50	30 seconds	
Extension	65	2 minutes	
Final extension	65	5 minutes	
Hold	4	No limit	-

Table 6. PCR setup for 18S rRNA amplification using 18S-specific primers.

Cycle step	**Temperature (°C)**	**Time**	**No. of cycles**
Initial denaturation	95	1 minute	1
Denaturation	95	20 second	30x
Annealing	68	30 seconds	
Extension	72	1 minutes	
Final extension	72	10 minutes	
Hold	4	No limit	-

After completing initial testing with the primers using different genomic materials, several trials were set up based on the results. Trials 1-10 each have different settings and conditions as shown in Table 7. All the different samples and trials were run on a 1% agarose gel, using GelRed Nucleic Acid Strain, 10000X concentration.

Table 7. Overview of all trials performed to test the different conditions for 16S, 16S-ONT, 18S and 18S-ONT primers.

Trial number	Primers	Primer concentration	DNA (16S/18S)	Control	Polymerase	Additions
1	16S- 16S-ONT 18S-SSU 18S-ONT	0.1 µM	Sheep lungs (18S) & bacterial colony (16S)	Negative control for 18S and 16S	DreamTaq	-
2	16S-ONT 18S-SSU 18S-ONT	0.4 µM & 0.8 µM	Human pancreas DNA (18S) Bacterial colony (16S)	No DNA control No primer control	DreamTaq	-
3	18S-SSU 18S-ONT	0.5 µM	Sheep lungs (18S)	No DNA control No primer control	DreamTaq	+BSA
4	18S-SSU 18S-ONT	0.5 µM & 1.0 µM	Sheep lungs (18S)	-	DreamTaq	+BSA/ -BSA
5	18S-SSU 18S-ONT	0.5 µM	Sheep lungs	-	KAPA HiFi Hotstart ready mix	+BSA/ -BSA
6	18S-SSU 18S-ONT	0.5 µM	Sheep lungs Bacterial colony	-	KAPA HiFi Hotstart ready mix	+BSA
7	18S-SSU 18S-ONT	0.5 µM	Sheep lungs	Negative control	KAPA HiFi Hotstart ready mix	+BSA
8	18S-SSU 18S-ONT	0.5 µM	Sheep lungs	Negative control	KAPA HiFi Hotstart ready mix	-
9	18S-ONT 16S-ONT	0.5 µM	Sheep lungs Human stool DNA	Negative control	KAPA HiFi & LongAmp	-
10	18S-ONT 16S-ONT	0.5 µM	Sheep lungs Human stool DNA	Negative control	KAPA HiFi + LongAmp (mixed)	-

2.7 Controls

ZymoBIOMICS gut microbiome standard was used to evaluate the extraction protocol and to assess the protocol used for NGS. The gut standard is composed of 21 different strains to mimic the human gut microbiome and the microbial composition of the control is shown in Table 8.

Table 8. Microbial composition of ZymoBIOMIC gut microbiome standard. All content is in %.

Species	Theoretical composition (%)			
	Genomic DNA	**16S Only**	**Genome copy**	**Cell number**
Faecalibacterium prausnitzii	14	17.63	14.77	14.82
Veillonella rogosae	14	15.87	19.94	20.01
Roseburia hominis	14	9.89	12.43	12.47
Bacteroides fragilis	14	9.94	8.33	8.36
Prevotella corporis	6	4.98	6.26	6.28
Bifidobacterium adolescentis	6	8.78	8.83	8.86
Fusobacterium nucleatum	6	7.49	7.53	7.56
Lactobacillus fermentum	6	9.63	9.68	9.71
Clostridioides difficile	1.5	2.62	1.10	1.10
Akkermansia muciniphila	1.5	0.97	1.62	1.62
Methanobrevibacter smithii	0.1	0.066	0.17	0.17
Salmonella enterica	0.01	0.009	0.007	0.0065
Enterococcus faecalis	0.001	0.0009	0.0011	0.0011
Clostridium perfringens	0.0001	0.0002	0.00009	0.00009
Escherichia coli (JM109)	2.8	2.53	1.82	1.83
Escherichia coli (B-3008)	2.8	2.53	1.82	1.82
Escherichia coli (B-2207)	2.8	2.29	1.64	1.65
Escherichia coli (B-766)	2.8	2.31	1.66	1.66
Escherichia coli (B-1109)	2.8	2.46	1.77	1.77
Candida albicans	1.5	N/A	0.31	0.16
Saccharomyces cerevisiae	1.4	N/A	0.32	0.16

A volume of 75 µL of the standard was used for DNA extraction, protocol 2.4.1. After DNA isolation, the concentration of the standard was measured using NanoDrop and Qubit fluorometer. Using a mock community like this one from ZymoBIOMICS can be used to evaluate protocols like DNA extraction and library preparation. The control consists of 21 different species, and it was constructed by pooling cells from pure cultures of 21 microbial strains. The cells from each pure culture were quantified before pooling. After mixing, the microbial composition was confirmed using NGS-based sequencing.

2.8 Library preparation – Oxford Nanopore Technologies

For the library prep, the protocols provided by ONT were used as described below in protocols 2.8.1 and 2.8.2.

2.8.1 16S Barcoding kit 1-24 (SQK-16S024)

A total of 10 ng of genomic material was transferred to a new tube, adjusting the total volume to 10 µL with nuclease-free water. For each of the samples, the following reaction was prepared in a 0.2 mL thin-walled PCR tube (Table 9).

Table 9. Components of the PCR for library preparation. Barcodes are specific for each sample.

Reagent	Volume
PCR barcode (one of BC1-BC96, at 10uM)	10 µL
10 µL input DNA (10 ng)	10 µL
LongAmp Taq 2x master mix	25 µL
Nuclease-free water	5 µL
Total volume	**50 µL**

The wells were mixed by pipetting before sealing the plate with adhesive film or a PCR strip cap and spinning the plate down. The samples are amplified using the following cycling conditions (Table 10):

Table 10. PCR cycle used for library preparation for 16S.

Cycle step	Temperature	Time	No. of cycles
Initial denaturation	95 °C	3 minutes	1
Denaturation	95 °C	15 seconds	12-15
Annealing	62 °C	15 seconds	12-15
Extension	65 °C	Dependent on length of fragment	12-15
Final extension	65 °C	Dependent on length of fragment	1
Hold	4 °C	No limit	-

The PCR product was transferred to a clean 1.5 mL Eppendorf tube, before adding 30 µL AMPure XP magnetic beads (Beckman Coulter). The eluate was thoroughly mixed on a rotator mixer for 5 minutes at RT. After pelleting the eluate on a magnet, the supernatant was discarded before washing the pellet with 200 µL of freshly made 70% ethanol.

The ethanol was discarded before repeating the washing with another 200 µL of ethanol. After discarding the ethanol, the pellet was left to dry for 30 seconds before dissolving the pellet in 10 µL of 10 mM Tris-HCl pH 8.0 with 50 mM NaCl. The pellet was once again pelleted on the magnet until the eluate was clean and colorless, and 10 µL of the eluate was retained in a new tube. 1 µL of the eluted sample was quantified using a Qubit fluorometer (Invitrogen).

All barcoded libraries were pooled in a ratio of 50-100 fmoles in 10 µL of 10 mM Tris-HCl pH 8.0 with 50 mM NaCl before adding 1 µL of RAP.

2.8.2 16S Barcoding kit 1-12 (SQK-RAB204)

The 16S barcodes were thawed and spun down before use. 10 ng of genomic DNA was transferred to a DNA LoBind tube and the volume was adjusted to 10 µL using nuclease-free water. In a 0.2mL PCR tube. The PCR was set up using the components shown in Table 11:

Table 11. Reagents used in the PCR for library preparation. Barcodes are specific for each sample and are provided in the kits from Oxford Nanopore.

Reagent	Volume
Nuclease-free water	14 µL
Input DNA (10ng)	10 µL
16S barcode, at 10 µM	1 µL
LongAmp Taq 2X master mix	25 µL
Total	**50 µL**

The PCR cycles were adjusted according to the amount of input material to produce the same yield. The PCR was done to amplify samples using the following cycling conditions (Table 12):

Table 12. PCR cycle used for library preparation for the 16S barcoding protocol.

Cycle step	Temperature	Time	No. of cycles
Initial denaturation	95 °C	1 min	1
Denaturation	95 °C	20 seconds	25
Annealing	55 °C	30 seconds	25
Extension	65 °C	2 minutes	25
Final extension	65 °C	5 minutes	1
Hold	4 °C	-	-

The sample was transferred to a 1.5 mL Eppendorf DNA LoBind tube. The AMPure XP magnetic beads (Beckman Coulter, Brea, California) were resuspended and vortexed before adding 30 µL to the reaction. The samples with the beads were incubated on a rotator mixer for 5 minutes at RT. The samples were spun down on a magnet before the supernatant was removed and the beads were washed using freshly made 70% ethanol. The ethanol was removed using aspiration before repeating the washing one more time. The pellet is centrifuged again to remove any remaining ethanol before drying the pellet for about 30 seconds. 10 µL of 10 mM Tris-HCl pH 8.0 and 50 mM NaCl were added to resuspend the pellet before incubating for 2 minutes at RT. To make the eluate clear and colorless, the beads were pelleted on a magnet and 10 µL of the eluate was removed and retained into a clean 1.5 mL Eppendorf DNA LoBind tube. All the barcoded libraries are pooled to a total of 50-100 fmoles in 10 µL of 10 mM Tris-

HCl pH 8.0 with 50 mM NaCl. For 16S amplicons of about ~1500 bp, 50-100 fmoles equate to ~50-100 ng. After pooling the libraries, 1 µL of the RAP was added to the barcoded DNA before gently mixing and incubating at RT for 5 minutes. The prepared library was used for loading into the MinION flow cells, the library was stored on ice until ready for loading.

2.8.3 Priming and loading the SpotON flow cell

Sequencing buffer, loading beads, flush tether, and flush buffer were thawed before mixing the four components by vortex and stored on ice.

The SpotON flow cell is loaded by first priming the SpotON priming port before loading the sample directly on the sample port. To prepare the flow cell, 30 µL of thawed and mixed Flush Tether were directly added to the tube of the mixed Flush buffer. The sensor array area is sensitive to air bubbles, to ensure there are no air bubbles after opening the priming port, a small volume was drawn back using a P1000 pipette. After making sure there are no air bubbles near the sensor array area, 800 µL of the priming mix was added to the flow cell via the priming port, avoiding any air bubbles. The priming Mix was incubated for 5 minutes, and the following was mixed for the sample library (Table 13):

Table 13. List of reagents used for priming and loading the SpotON flow cell.

Reagent	**Volume**
Sequencing buffer (SQB)	34 µL
Loading beads (LB), mixed immediately before use	25.5 µL
Nuclease-free water	4.5 µL
DNA library	11 µL
Total	**75 µL**

The flow cell priming was completed by lifting the SpotON sample port cover to make the SpotON sample port accessible and loading 200 µL of the priming mix into the flow cell via the priming port, avoiding air bubbles. The prepared library is thoroughly mixed prior to loading, and 75 µL of the sample is added to the flow cell via the SpotON sample port in a drop-by-drop way.

After ending the sequencing, the flow cell is washed according to the washing protocol provided by ONT. The wash protocol can be adjusted accordingly to directly run a second library, or for storing the flow cell for later use.

2.8.4 Flow cell wash protocol

The MinION flow cells can be reused after sequencing. The wash protocol is used for cleaning out DNA/RNA libraries already added to the flow cell. Washing the flow cell aims to remove most of the initial library and prepare the flow cell for loading of a subsequent library.

One tube of wash mix (WMX) was placed on ice and one tube of wash diluent (DIL) was thawed at RT before vortexing, spinning down and storing on ice until further use. In a clean 1.5 mL Eppendorf DNA LoBind tube, the following flow cell wash mix was added (Table 14):

Table 14. Reagents that were used for washing the SpotON flow cell.

Reagent	**Volume**
Wash mix (WMX)	2 µL
Wash diluent (DIL)	398 µL
Total	**400 µL**

The content was mixed well by pipetting and placed on ice. If not already, the sequence run was stopped, and the flow cell was left in the device. Using a P1000 pipette, all fluid was removed through waste port 1 making sure the sample port and priming port are closed so no fluid will leave the sensor array area. The priming port cover was opened, checking for small air bubbles by drawing back 20-30 µL of fluid before loading 400 µL of the wash mix into the flow cell via the priming port, avoiding the introduction of air. The priming port was closed and left for 60 minutes. The fluid was once again removed through waste port 1, making sure the priming port and sample port are closed. It is crucial that no air is drawn across the sensor array area, which would lead to a significant loss of sequencing channels.

From this step, the flow cell can be used straight away by following protocol 2.8.3 Priming and loading the SpotON flow cell or storing the flow cell for later use.

One tube of storage buffer (S) was thawed at RT before loading 500 µL of the storage buffer through the priming port and closing the port. All fluids were drawn back from waste port 1. The flow cell can now be stored at 4-8 °C.

After completing the flow cell washing, the flow cell needed to be checked for remaining pores by performing a flow cell check. The flow cell can be checked before use by running a Flow Cell Check with MinKNOW (the MinION and GridION device software). The number of nanopores available for sequencing will be reported by the Flow Cell Check.

2.9 Sequencing

All flow cells used for sequencing were checked to see the number of available pores prior to sequencing. The sequencing runs were performed and monitored using the MinKNOW software. Data acquisition, real-time analysis, run feedback, local basecalling, and data streaming are all performed by MinKNOW (v 5.1.0), the operating software that operates nanopore sequencing devices.

2.9.1 Illumina sequencing, short-read sequencing

Aliquots of all samples were diluted to a known set concentration (10 ng/µL per sample) and sent to Statens Serum Institut (SSI), Copenhagen, Denmark for Illumina MiSeq sequencing, as a part of their public health screening at the laboratory of Parasitology. The resulting data number of *Blastocystis* positive/negative along with some taxonomic information was sent in return for further analysis. The NGS platform Illumina is currently the golden standard for sequencing used in diagnostics and clinical trials, using well-developed pipelines for sequencing and data analysis. The SSI group has their in-house designed primers used for sequencing of *Blastocystis* and their own data analysis pipeline for taxonomy and diversity studies.

2.10 Bioinformatics/data analysis

For Oxford Nanopore sequencers like the MinION, the FAST5 format is the usual sequencing output. It is based on the HDF5 hierarchical data format, which allows vast and complex data to be stored. A FAST5 file, unlike a FASTA or FASTQ file, is binary and a standard text editor cannot read the file. To convert FAST5 to FASTQ, basecalling is required.

FASTQ is the standard for second-generation sequencing technologies like Illumina sequencers. It's like the FASTA format, except a FASTQ file, additionally holds the sequence's quality ratings in addition to the sequence itself. The process of translating the electrical signals generated by a DNA or RNA strand passing through a nanopore into the strand's matching base sequence is known as basecalling. Figure 6 shows an overview of the bioinformatic workflow used in this study from raw sequences to taxonomic analysis.

Guppy is a data processing toolkit that features basecalling algorithms from ONT as well as numerous bioinformatic post-processing tools. It's accessible as binaries for Windows, OS X, and Linux, and it's also integrated within the MinKNOW (v 5.1.0) software, Oxford Nanopore's device control software. In this study, Guppy basecaller (v 6.1) was used on the raw sequencing files.

After the basecalling is complete, Guppy (v 6.1) will put all FASTQ files into the same folder, so all the different barcoded samples needed to be separated into different folders before further processing. The barcode and adapter sequences will still be attached to your FASTQ files (from when you prepared the library prior to sequencing) to be left with only the sequence from the original sample, the barcodes, and adapter sequences were trimmed off using Guppy (v 6.1). The quality control ensures good coverage and gives a good indication of the distribution of reads. The minimum Q-score is 7, so anything not within the standard will be cut out also using the Guppy (v 6.1) basecalling along with adapters, primers, and barcodes. Guppy is only available to ONT customers via their community site (Oxford Nanopore Technologies – ONT, Oxford, UK).

To ensure even better quality, a software called Porechop (v 0.2.4) was used (Wick, 2017). This removes even more barcodes and finds the best match between the barcode sequences. Porechop is a software that searches for and removes adapters from Oxford Nanopore reads. Reads with adapters on the ends are clipped, while reads with adapters in the middle (chimeric reads) are cut into distinct reads according to user-defined parameters and minimum length thresholds. Porechop uses thorough alignments to identify adapters, even when the sequence identity is low. Nanopore reads that are barcoded using kits provided by Nanopore (Native barcoding kit, PCR barcoding kit or rapid barcoding kit) can be demultiplexed using Porechop. It would be very time-consuming to check the quality of reads manually. That's why tools like NanoPlot and FastQC exist to generate a summary and plots of data statistics. NanoPlot is mostly used for long-read data, such as ONT and PACBIO, whereas FastQC is primarily used for short-read data, such as Illumina and Sanger. The NanoPlot package is developed for use with ONT sequencing and is easily downloaded and the codes are ready to use, the only steps required are to change the output file, directory, and which files or reads to plot.
Another filtration software to use for ONT data is NanoFilt. NanoFilt (v 2.8) was used for trimming and filtering long-read sequencing data and cutting out chimeras (sequences attached together, separated in the middle by an adapter or BC) (De Coster *et al.*, 2018). Reads from stdin, writes to stdout.
After trimming and filtering, all sequence reads were quality checked using NanoPlot (v 1.39) (De Coster *et al.*, 2018). With the trimmed and filtered sequences, Kraken 2 (v 2.1.2) was used to assign taxonomic ranks. Kraken 2 uses a sub database downloaded from NCBI, containing 16S rRNA genes for taxonomic classification. After basecalling, trimming, filtering and taxonomic classification, the files are converted to a .BIOM file, which can be used as an input file for many programs and scripts to assign taxonomy, especially used in phyloseq analyses.

A phyloseq package is a tool for importing, storing, analyzing, and graphically displaying complex phylogenetic sequencing data that has already been clustered into Operational Taxonomic Units (OTUs), particularly when sample data, a phylogenetic tree, and/or taxonomic assignment of the OTUs are available (McMurdie *et al.*, 2013).

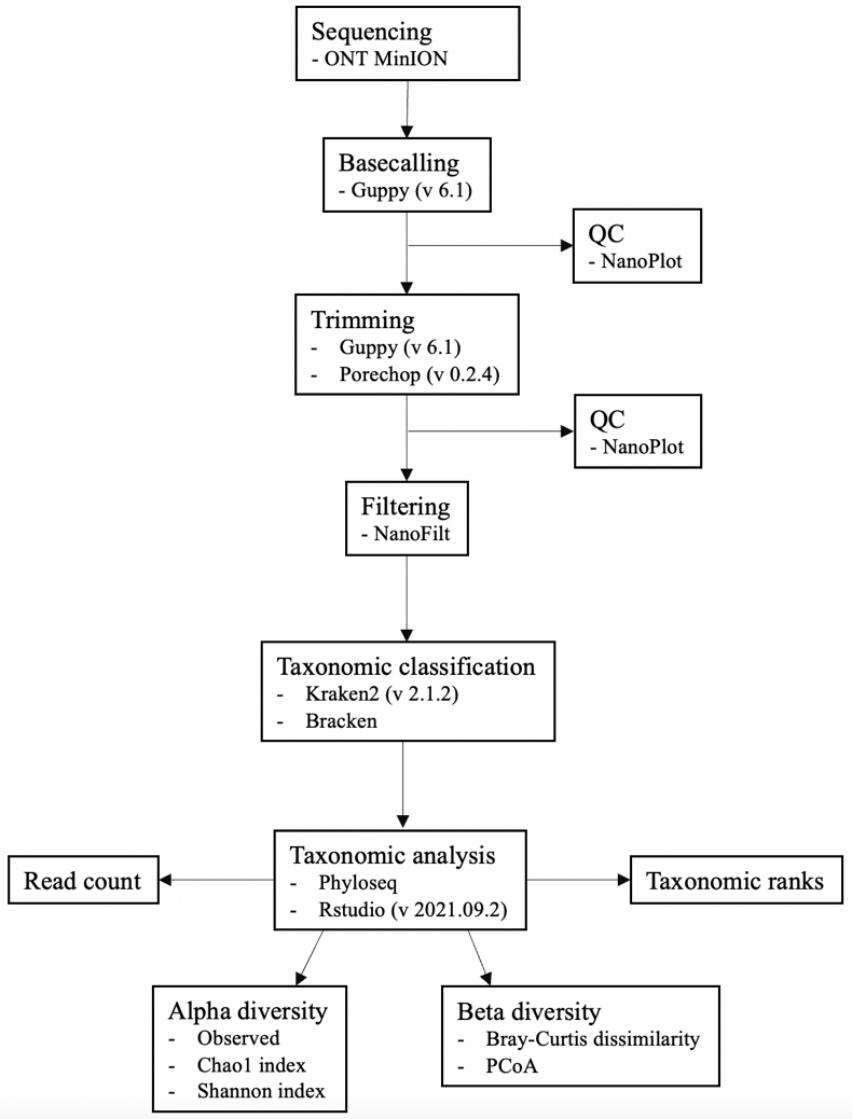


Figure 6. Flowchart of the workflow used for data analysis from raw sequence reads to taxonomic analysis. The raw sequencing reads were basecalled before trimming and filtering. All trimming and filtering were visualized using NanoPlot (v 1.39) as a quality control. After trimming and filtering, taxonomic classification was done using Kraken 2 (v 2.1.2). All taxonomic analyses were done using Rstudio (v 2021.09.2) and phyloseq package for exploring microbiome profiling.

All taxonomy work took place in Rstudio (v 2021.09.2) running R (v 4.1.3). Phyloseq is a library with tools to analyze and plot your metagenomics tables. The Phyloseq package was installed to Rstudio before loading in the necessary libraries to do the taxonomy work. The data files with numbers of reads per OTU and taxonomic labels for each OTU were loaded into the program. After loading in the data, one can inspect the object and see what is created. It might be necessary to remove or replace unnecessary characters in the data matrix.

Alpha diversity, beta diversity, and a general overview of the taxonomy is then created using libraries and packages included in the phyloseq pipeline.

2.10.1 Statistical analysis

The statistical analyses were performed using RStudio (v. 2021.09.2) and scripts used for NGS sequence analysis. Kruskal-Wallis test is done in a way where all the data are pooled and ranked from smallest to largest, then the sums of ranks in each subgroup are added up. The probability is then calculated (Hoffman, 2019). A PERMANOVA test is best described as a geometric partitioning of multivariate variation in the space of a chosen dissimilarity measure. PERMANOVA is often used to compare groups of objects and to test the null hypothesis that all groups' centroids and dispersion, as described by measure space, are equal (Anderson, 2017). PERMANOVA is for testing if the samples differ between groups, while the Kruskal-Wallis is used to find which groups differ from which.

3 Results

3.1 DNA quality after extraction

Genomic material used for sequencing needs to be of good quality to ensure good quality reads. The stool samples were collected in a stool sample collection tube, and in order to sequence the samples, the DNA needed to be extracted. The DNA was isolated using Qiagen fast DNA stool kit and protocol Q provided by Costea *et al.* (2017), resulting in a large spread in the amount of DNA measured as shown in Figure 7. The values for all DNA qualities for all the samples are also listed in Table 15. All samples were measured using a Nanodrop One microvolume UV-Vis spectrophotometer. The purity of the DNA samples also varied from poor to excellent purity.

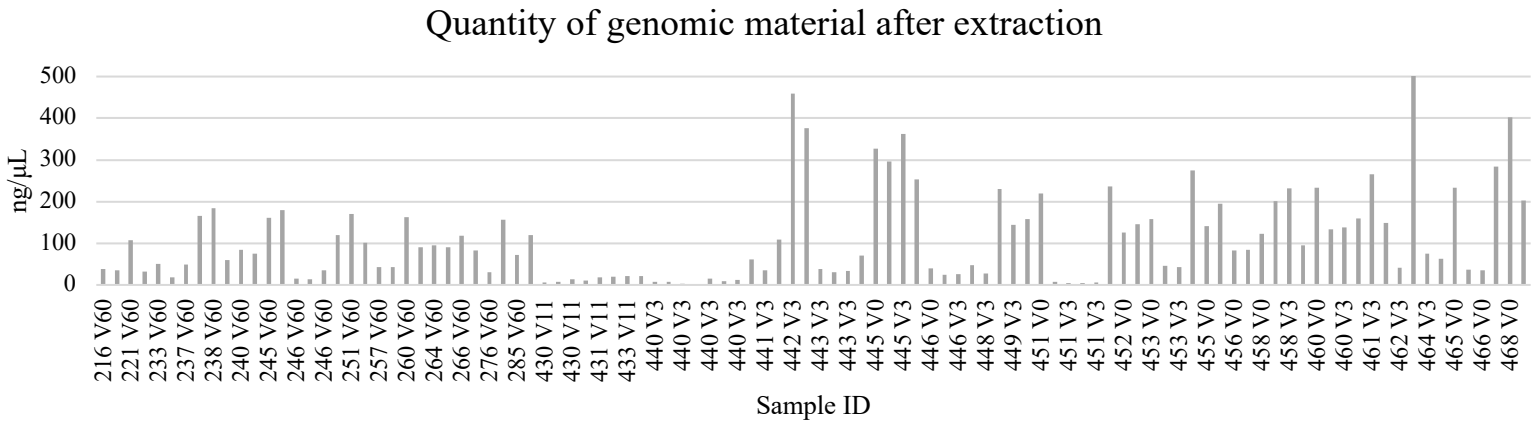


Figure 7. Measured DNA yield after DNA extraction of stool samples. The plot shows the overall quantity. The y-axis is ng/µL, where the maximum value is set to 500 ng/µL. Only one sample exceeded this range. The x-axis shows the sample names.

About 27% of the extracted samples had a DNA concentration of under 30 ng/µL, and 21% of the samples had absorbance below the desired 1.8 (A260/280). 41% of the 96 extracted samples fell below the desired absorbance for A260/230 (2.0).

3.2 DNA purification and clean-up

For sequencing, it is crucial to do constant quality assessments during pre-sequencing steps to ensure high quality to determine whether the DNA sample should be proceeded with for sequencing. If the quality is not up to par, additional measures like re-extraction or purification should be performed. After DNA extraction and measurement of the quantity and quality of the genomic material, DNA purification was performed on samples with low quality or poor purity. Three different purification protocols were tested, ethanol precipitation, Zymo DNA concentration and clean-up kit and clean-up using magnetic beads (AMPure XP beads). The

purification using AMPure XP beads was the most efficient method for clean-up while samples with lower quality were purified using the magnetic beads or re-extracted.

3.3 ZymoBIOMICS mock community – control

The ZymoBIOMICS mock community was used to determine the quality and outcome of the DNA extraction protocol used in this study. The mock community was extracted using protocol Q and the sample was included in library preparation and sequencing along with samples from patients. The control was run through the same pipeline for data analysis, and the sequencing results is shown in Figure 8, which displays the results on a species-level.

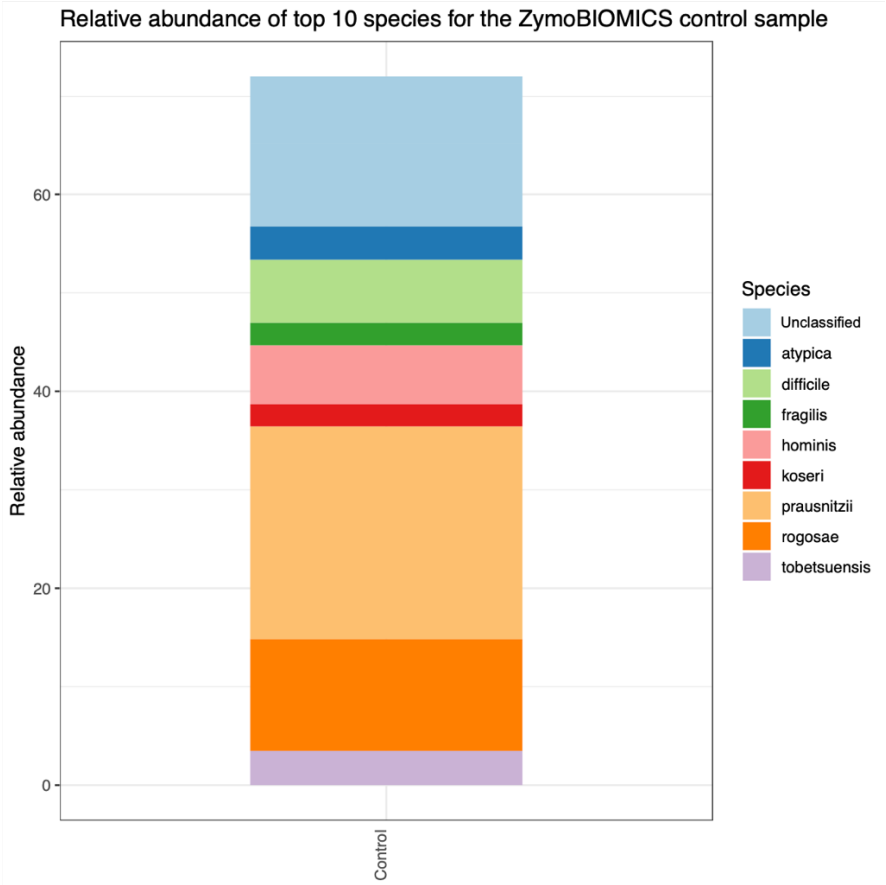


Figure 8. The ZymoBIOMICS mock community control's composition of species, sequenced by MinION. The plot displays the distribution of bacterial species in the mock community. The colors are representing the different species, as shown.

The relative abundance of the top 10 species for the ZymoBIOMICS shows that the most abundant species is *Faecalibacterium prausnitzii*, *Veillonella rogosae* and *Clostridioides difficile.* The light blue section marked "unclassified" is unclassified species.

3.4 PCR

ONT uses primers especially designed for NGS. These primers differ from regular primer sequences in that they have an overhang at the ends of each primer sequence. This overhang allows for barcoding the samples during the PCR part of the library preparation. The most critical aspect in generating good PCR performance data is the precision of a primer pair's design and synthesis. The results from the testing of primer conditions. The first trial included control DNA that was available in the lab from another project. The DNA originated from lungs of sheep and bacterial colonies, isolated from the nasal passage of sheep. The PCR was conducted using DreamTaq PCR Master Mix (2X). Both specific- and unspecific ONT-primers were used with both DNAs, the optimal temperature for 16S ended up being 50 °C and 68 °C for 18S. Figure 9 shows the 1% agarose gel from the first trial.

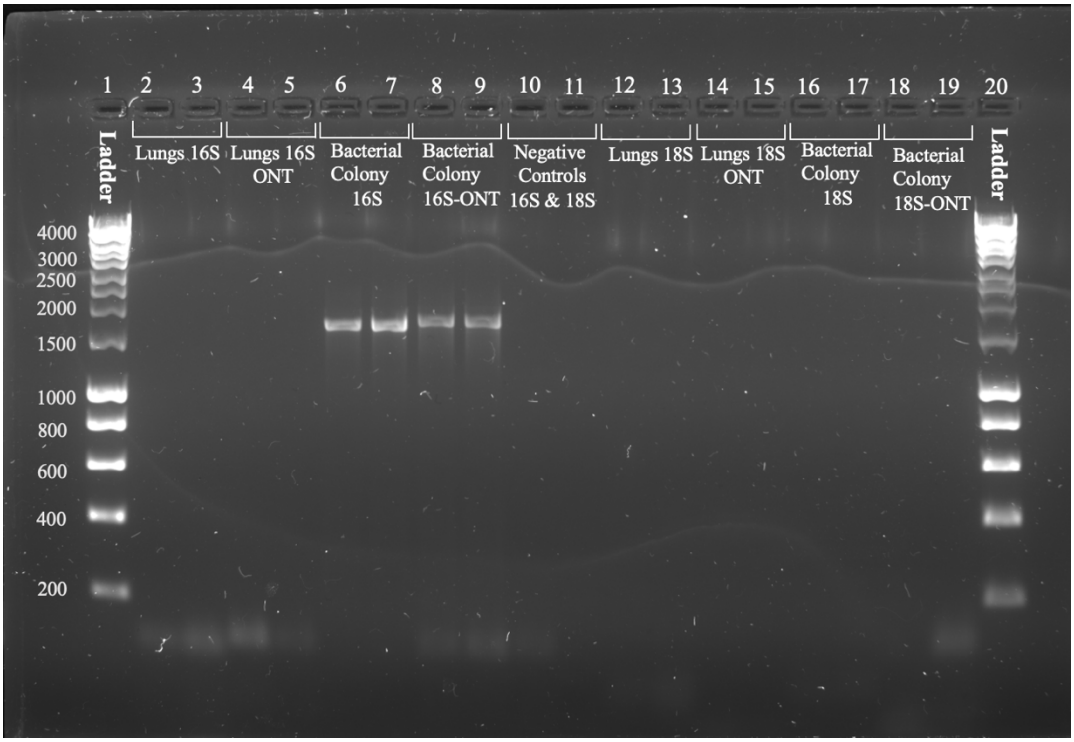


Figure 9. Gel electrophoresis of PCR conducted on 16S and 18S (both ONT-specific and unspecific primers) using DNA from lungs of sheep and bacterial colony from nasal swabs from sheep. The first gel electrophoresis, using a 1% agarose gel, with GelRed staining. All samples were loaded with a volume of 10 µL, whereas 2 µL were loading dye, 3 µL of the genomic sample, and 3 µL of nuclease-free water. Well 1 and 20 are 1 kb HyperLadder, well 10 is the 16S negative control, and well 11 is the 18S negative control. Wells 2, 3, 12, and 13 show Lung DNA with 16S and 18S primers. Well 4, 5, 14, and 15 show lung DNA with 16S-ONT- and 18S-ONT-specific primers. Wells 6, 7, 16, and 17 show bacterial colonies with 16S and 18S primers. Lastly, wells 8, 9, 18, and 19 show bacterial DNA with 16S-ONT and 18S-ONT primers.

The first gel shows four clearly visible bands. The first and last well contains the 1 kb HyperLadder, followed by four wells with DNA from sheep lungs with 16S- and 16S-ONT primers. The visible bands in well 6 and 7 are bacterial DNA with 16S primers, and well 8 and 9 are bacterial DNA with 16S-ONT primers. The rest of the wells all contained 18S- and 18S-ONT primers, none of them showed any bands. The primer concentration was 0.1 µM for all samples. Only the bacterial DNA combined with 16S- and 16S-ONT primers showed visible bands. The next trials were tweaked until both 16S and 18S (both ONT-specific and unspecific) showed bands.

16S and 16S-ONT primers worked for almost all trials, so further testing is needed to make 18S and 18S-ONT work. The DreamTaq Master Mix was replaced by KAPA HiFi Hotstart Ready-mix and LongAmp master mix, which had positive results for the 18S primers.
The final trial of the primer testing shows that both the LongAmp Taq 2X master mix and KAPA HiFi master mix both work for sheep lung bacterial DNA (18S) while the 16S-ONT specific primer works for human stool DNA. Figure 10 shows the last gel for the final trial. Bands are located at 1500 bp, the no DNA controls show no bands, as expected.

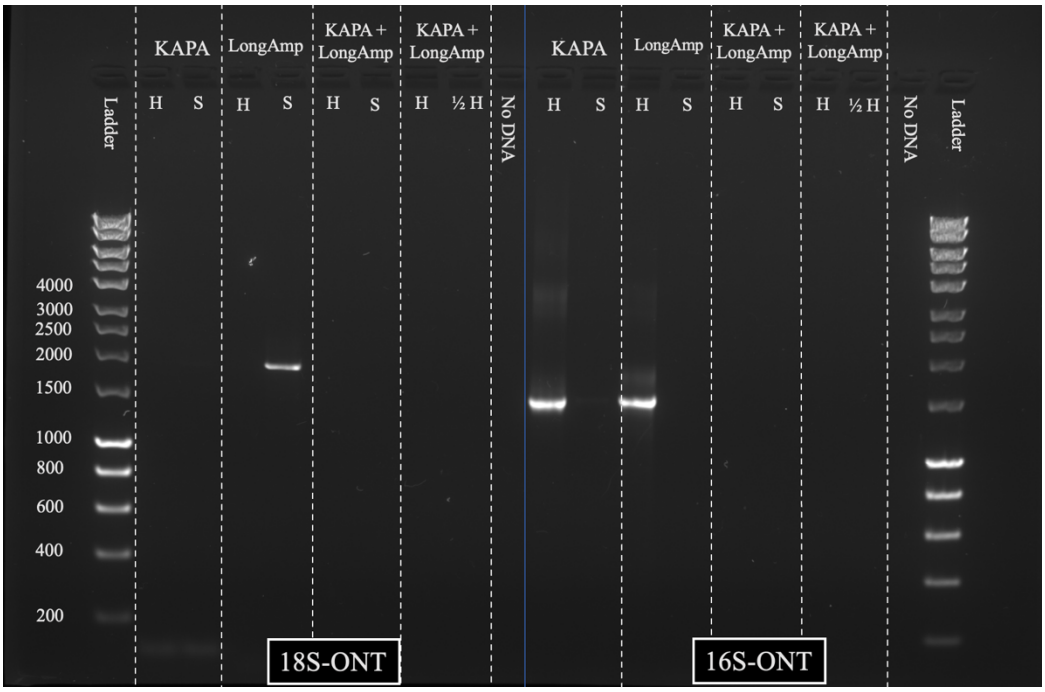


Figure 10. Gel electrophoresis of last primer trial for 16S-ONT and 18S-ONT specific primers. The first half of the gel is 18S-ONT primers, H = DNA from human stool samples, S = DNA from sheep lungs. Both the KAPA HiFi Hotstart Ready-mix and LongAmp Taq 2X master mix were used for this last trial. The right side of the gel is 16S-ONT, with the same setup as the left side. The two last wells on each side contained ½ the amount of human stool DNA.

This study ended up focusing on sequencing only 16S due to time limitations, but the knowledge of which master mix worked best for both 16S and 18S was used for the rest of the study and for future prospects.

3.5 Sequencing results

3.5.1 Sequencing run analysis

The importance of assessing the quality of sequencing data cannot be emphasized enough. It can also aid in determining if the sequence data requires any additional treatments before going to downstream analysis, depending on the sequencing purpose.

The sequencing run was monitored in real-time using ONT's published software, MinKNOW, which was created exclusively for ONT's nanopore sequencing machines. The final output of the run can be predicted with significant reliability by observing the start of the run; and if it does not appear to be promising, the run can be terminated rather than discovering the poor result after the run, as is the case in ONT technologies. The sequence data generated at the end of the run can also be used to assess the run's quality in addition to the real-time feedback.

All sequencing runs began with a pore occupancy that exceeded 70%; this pore occupancy usually is maintained for about 5 hours before the occupancy starts to decline.

The pore occupancy decreases with time in a steady, slow manner. As the number of inactive pores grows, the pore occupancy decreases, and ultimately the quality of the run declines. The pore occupancy can affect the number of reads per sample, as shown in Figure 11. The sequencing runs were usually stopped after approximately 24 hours, and the flow cell was re-used unless the pore occupancy was too low to finish a second run using the same flow cell.

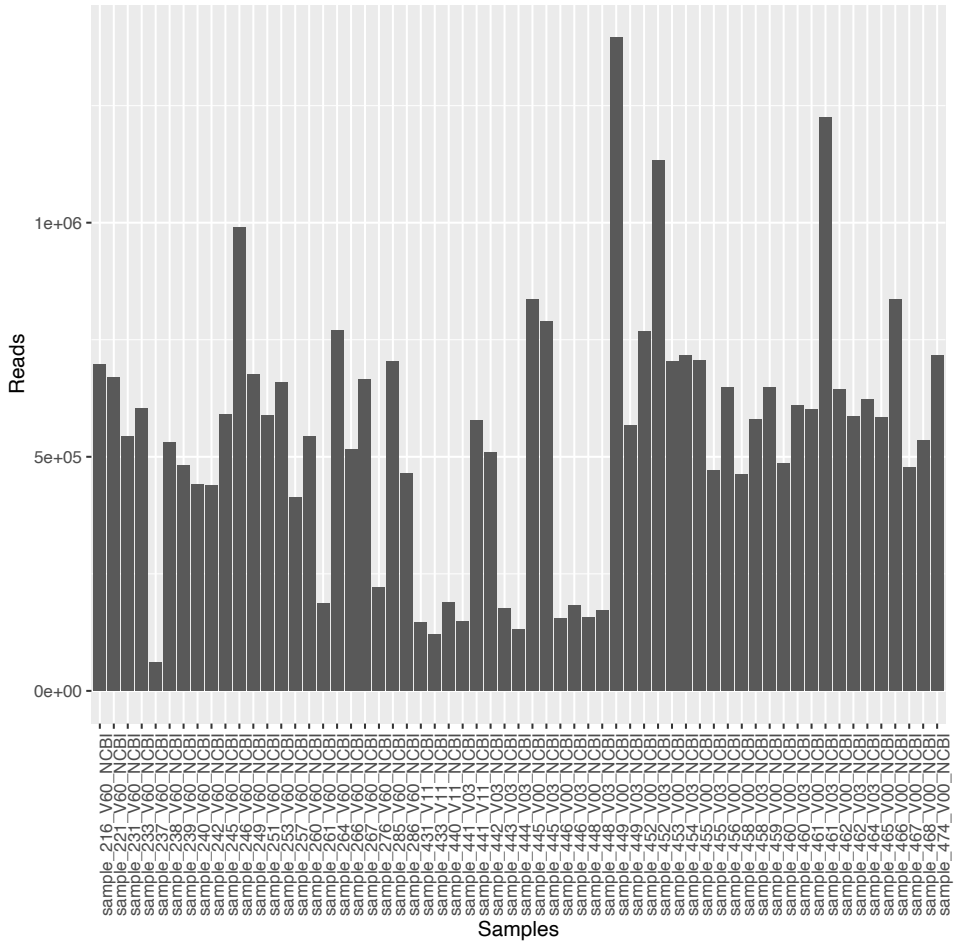


Figure 11. Total read count for all samples after sequencing with the MinION. This plot shows all samples and their read counts. The y-axis is number of reads; the x-axis is sample names. The samples are not categorized based which runs they were in. This plot was made using phyloseq microbiome analysis tool In Rstudio (v 2021.09.2).

The last sequencing run (SUS_8) was performed on a used flow cell, where the quality of the DNA was poor, resulting in a high number of pores being blocked during the first run. The quality of the second run was affected by this, and the total reads and bases were not as good as the previous sequencing run. The samples with the lowest read count in Figure 11 all belonged to SUS_8.

3.5.2 Quality assessment of sequencing

In *de novo* genome assembly construction, the length of the reads is of importance. The longer the reads, the better it is for the assembly's continuity, and for resolving repetitive regions in the genome. This means that long reads in the sequence data are a good indication of good quality DNA samples and vice versa. Another metric used to evaluate the quality of the sequencing was to assign Phred Quality score (Q-score or value) to the read during basecalling and filtering. Assigning a Q-score to reads allows to gauge the quality of the sequencing by actively sorting out the more accurate reads and it is primarily dependent on the method of

sequencing rather than the quality of the input DNA sample. The Q-score used for these sequencing runs was set to 7. For filtering and trimming post-sequencing, Guppy, Porechop and NanoFilt (v 2.8) was used to trim off barcodes, adapters, primers, and splitting chimeras. The average quality of the reads (Q-score) was set to 7, and the desired read length for long-read sequencing is 1200-1800 bp. All trimming and filtering were visualized using Nanoplot (v 1.39), as shown in Figure 12.

Figure 12. The read length vs average read quality before and after trimming. The plot showing read length and quality before trimming (left) has several points below the desired length, which is a minimum of 1200 bp. The plot showing the quality after trimming and filtering (right) has fewer points, but more even distributed in the length of 1200-1800 bp (x-axis) area. The quality (y-axis) is also set to a minimum Q-score of 7.

After trimming and filtering, the reads left are of good quality (above 7) and of good length (between 1200-1800 bp). This is crucial to proceed with data analysis and obtain good quality results.

3.5.3 Alpha diversity

The phyloseq package estimates the alpha diversity, which is the diversity within the samples. Observed species, Chao1, and Shannon-Wiener indexes were calculated and are presented in Figure 13 as a), b) and c) respectively. The Chao1 index is a qualitative measure of alpha diversity that, in addition to species richness, includes the ratio of single observations (n = 1) to double observations (n = 2), weighting unusual species more heavily. However, when it comes to diversity, it's important to consider not just the number of species, but also their abundance.

"Evenness" refers to the relative abundances of the many species that make up the samples richness. Both OTU richness and evenness are correlated by the Shannon-diversity index.

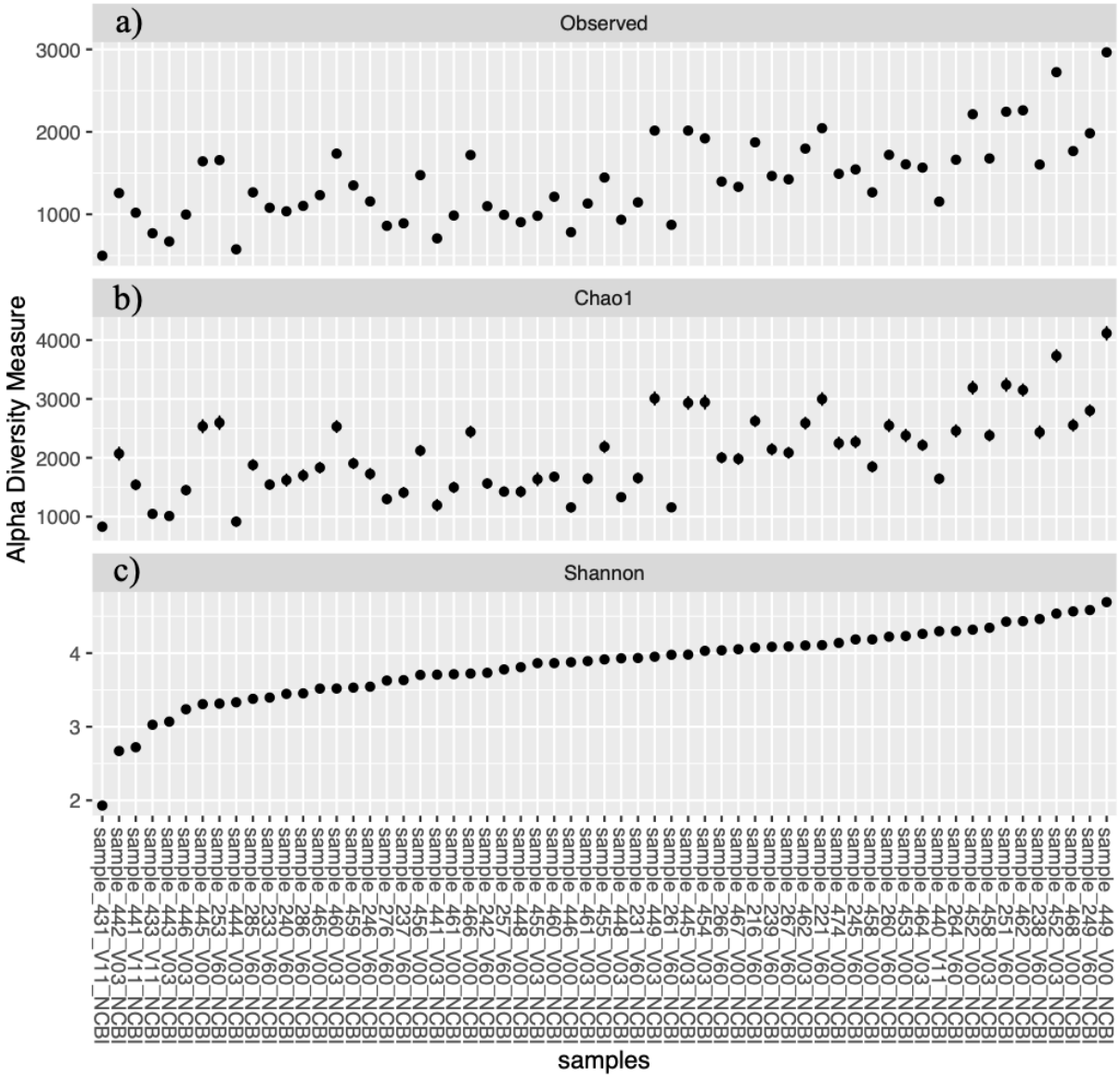


Figure 13. Alpha diversity indexes for sequencing all samples included in the study. The alpha-diversity indexes shown are species richness; observed species) in a), Chao1 index in b) and Shannon index in c). The y-axis a) shows the number of unique species observed within the samples, while b) it shows the Chao1 index, and c) shows the Shannon-Wiener index. The x-axis shows the sample names.

The observed diversity is ranging from below 1000 observed species to about 3000 in unique species observed within the different samples. The Shannon diversity index show that the samples varied from a Shannon index of 2 to about 4.5, meaning that the diversity within the samples varies a lot, from the sample with the lowest diversity to the sample with the highest diversity ($p > 0.05$, Kruskal-Wallis).

3.5.4 Beta diversity

The diversity between the samples was determined using beta-diversity indexes for the communities. The beta-diversity indexes presented in Figure 14 are derived from the phyloseq pipeline. Non-metric multidimensional scaling (nMDS) ordination plot was generated to display dissimilarities in the samples, the ordination used Bray-Curtis dissimilarity, the results are presented as Principal Component Analysis (PCoA) plots.

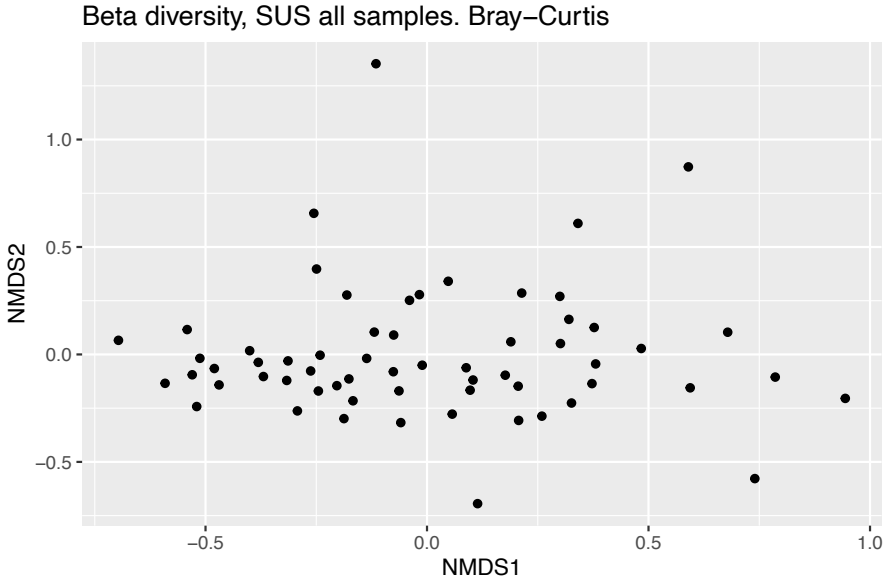


Figure 14. PCoA plot of Bray-Curtis dissimilarity index for all samples sequenced by MinION. The samples are scattered as black points. The y-axis is NMDS2; the x-axis is NMDS1 (Non-metric multidimensional scaling).

The samples are ordinated with some distance, indicating some dissimilarity. A Bray-Curtis analysis is based on dissimilarity, and there are some samples that differ a lot (those who are closer to 1.0 on both axes) while there is somewhat of a clustering between -0.5 and 0.

3.5.5 Taxonomy

We analyzed which groups of bacteria that were dominating in the stool samples. The taxonomic assembly was done using the phyloseq pipeline and NCBI database for 16S classification, and the different phyla are shown in Figure 15.

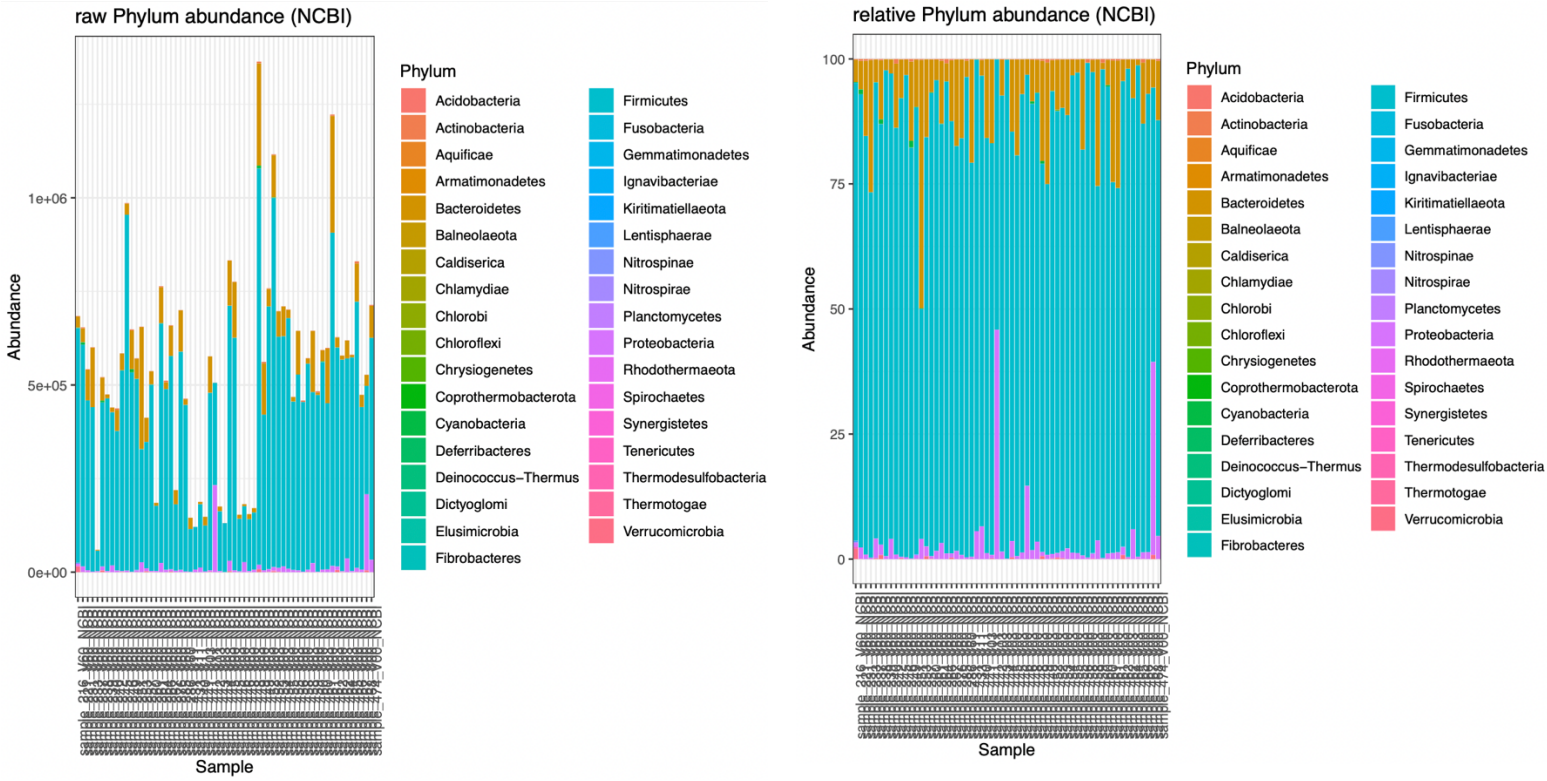


Figure 15. Phyla from the stool samples of sequencing runs using the MinION. The panels show raw phylum abundance (left) and relative phylum abundance (right). Y-axis is the abundance; the x-axis is the different samples.

The sequencing shows that Firmicutes are the dominating phyla among all the samples. The amount varies between the samples, Bacteroidetes is the second most common phylum, followed by a smaller amount of Proteobacteria and Verrumicrobiota. There are some variations in the number of phyla present in the samples, but Firmicutes and Bacteroides made up a relatively large proportion of the phyla detected for all samples. A total of 2237 OTUs were assigned to the phylum Firmicutes, 625 OTUs to Bacteroidetes, and 2015 OTUs for Prevotella. Using the MinION sequencer along with the data analysis pipeline and NCBI database makes it possible to get classification down to genus level, as shown in Figure 16. The plot shows the top ten abundant genera for all the sequenced samples.

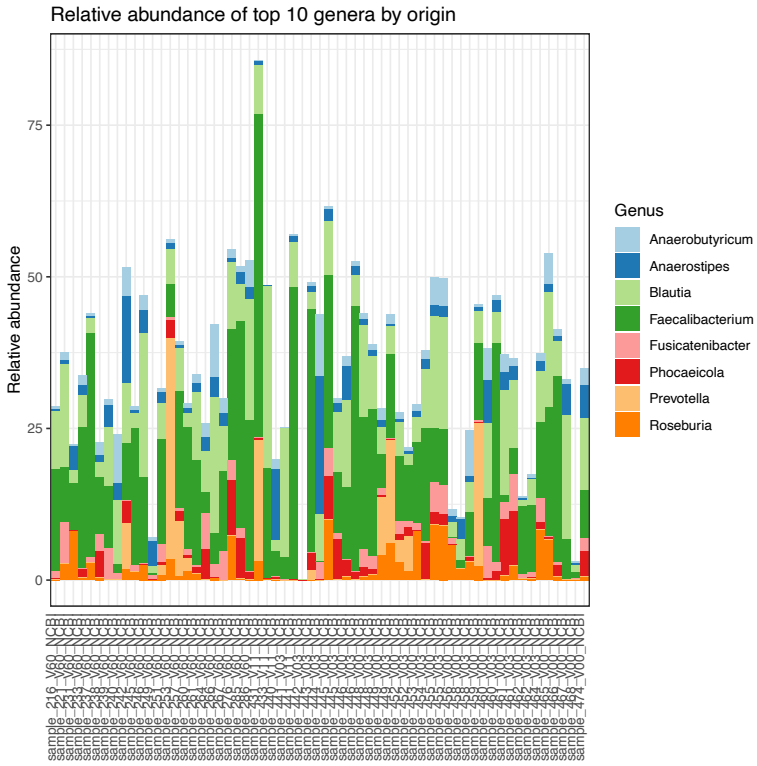


Figure 16. Top ten genera by origin. The different genera are shown in the different colors displayed next to the plots. The y-axis is relative abundance, while the x-axis is the samples.

The top abundant genus is *Faecalibacterium*, followed by *Blautia*, *Roseburia* and *Prevotella*. The distribution between the genera is varying between all patients, some have higher abundance of different genera, while others have large quantities of fewer genera. In order to understand the diversity of genera, the patients can be separated into groups based on metadata like CD vs UC, gender, and treatment, which could provide with more specific taxonomic information. The top ten most abundant species is also shown in Figure 25 (Appendix).

3.5.6 UC *vs* CD

By dividing the patients into groups based on whether they are diagnosed with UC or CD, the difference in the microbiome can be inspected. The raw and relative abundance of phyla for patients diagnosed with CD is shown in Figure 17.

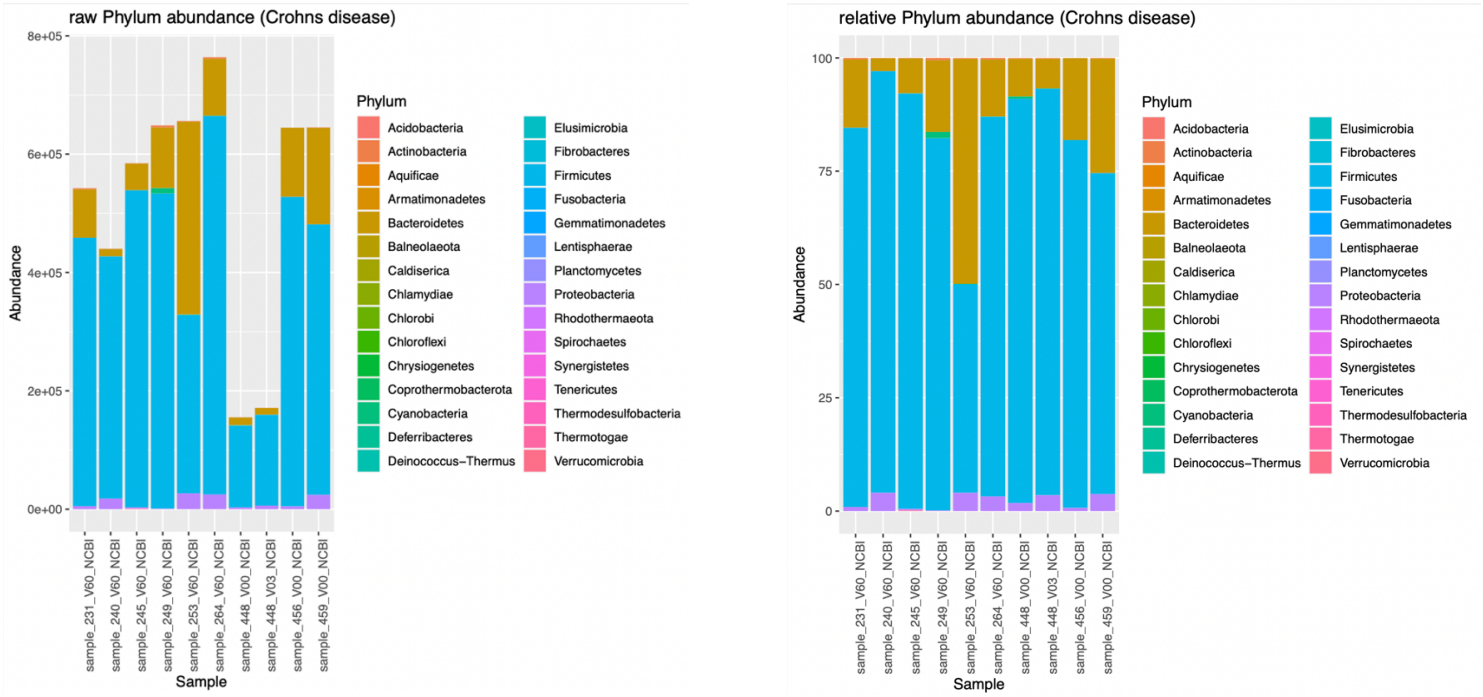


Figure 17. Phyla from the stool samples of patients diagnosed with CD. The panels show raw phylum abundance (left) and relative phylum abundance (right). Y-axis is the abundance; the x-axis is the different samples.

The most abundant phylum for CD patients is Firmicutes, Bacteroidetes and Proteobacteria. For some of the samples there is a clear variation in the amount of each of the top phyla, showing that the F/B ratio varies between the samples. Visibly, there is a difference in the amount of Bacteroidetes compared to UC. The raw and relative abundance of phyla for patients diagnosed with UC is shown below in Figure 18.

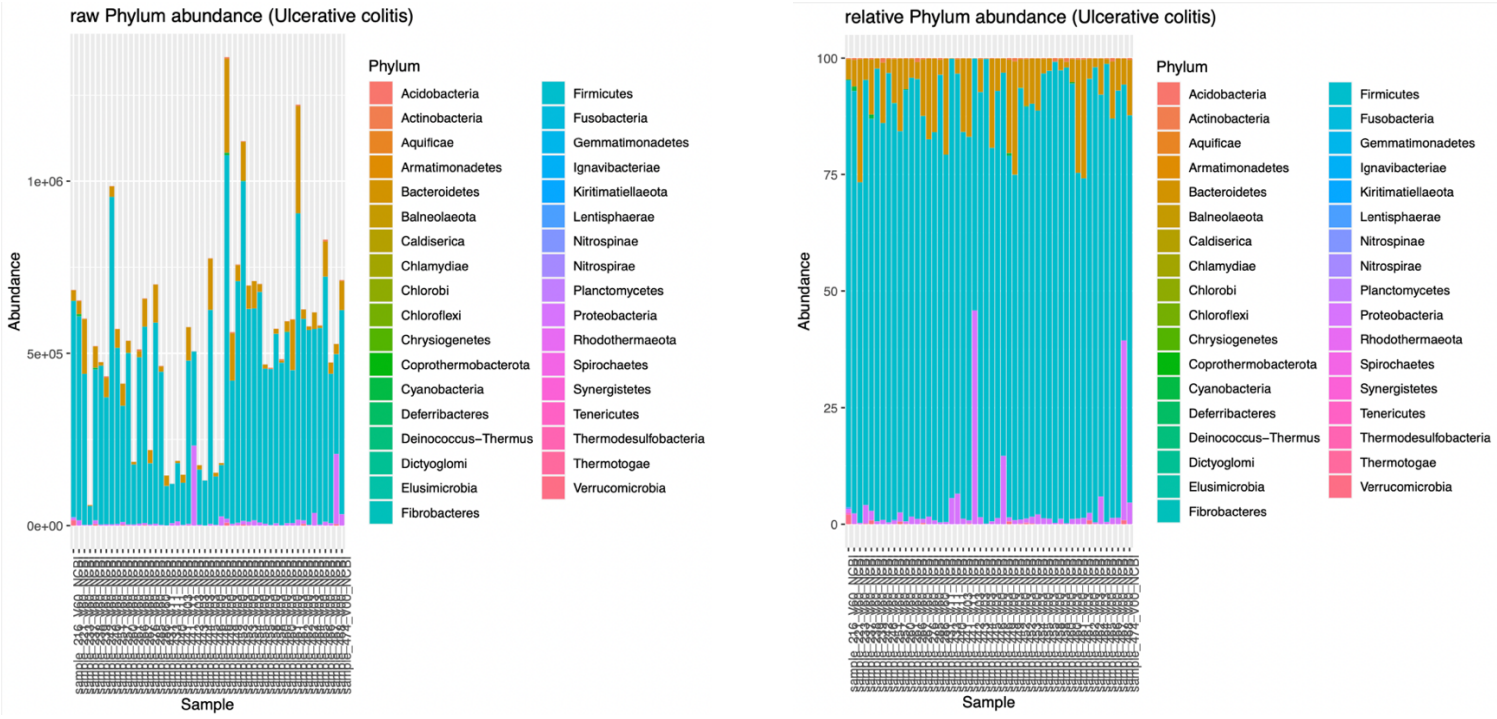


Figure 18. Phyla from the stool samples of patients diagnosed with UC. The panels show raw phylum abundance (left) and relative phylum abundance (right). Y-axis is the abundance; the x-axis is the different samples.

For the UC patients, Firmicutes made up the largest portion of the phyla detected with MinION. The other abundant phyla are Bacteroidetes, Proteobacteria and a small amount of Verrumicrobia. Compared to CD, there is a greater variety in the amount of Proteobacteria for the UC patients. Due to the F/B ratio being so dominating, getting a more detailed representation of other phyla or even genera is difficult. A plot showing the top ten genera for UC and CD patients were created in order to look more closely at the taxonomic variation between the two groups (Figure 19).

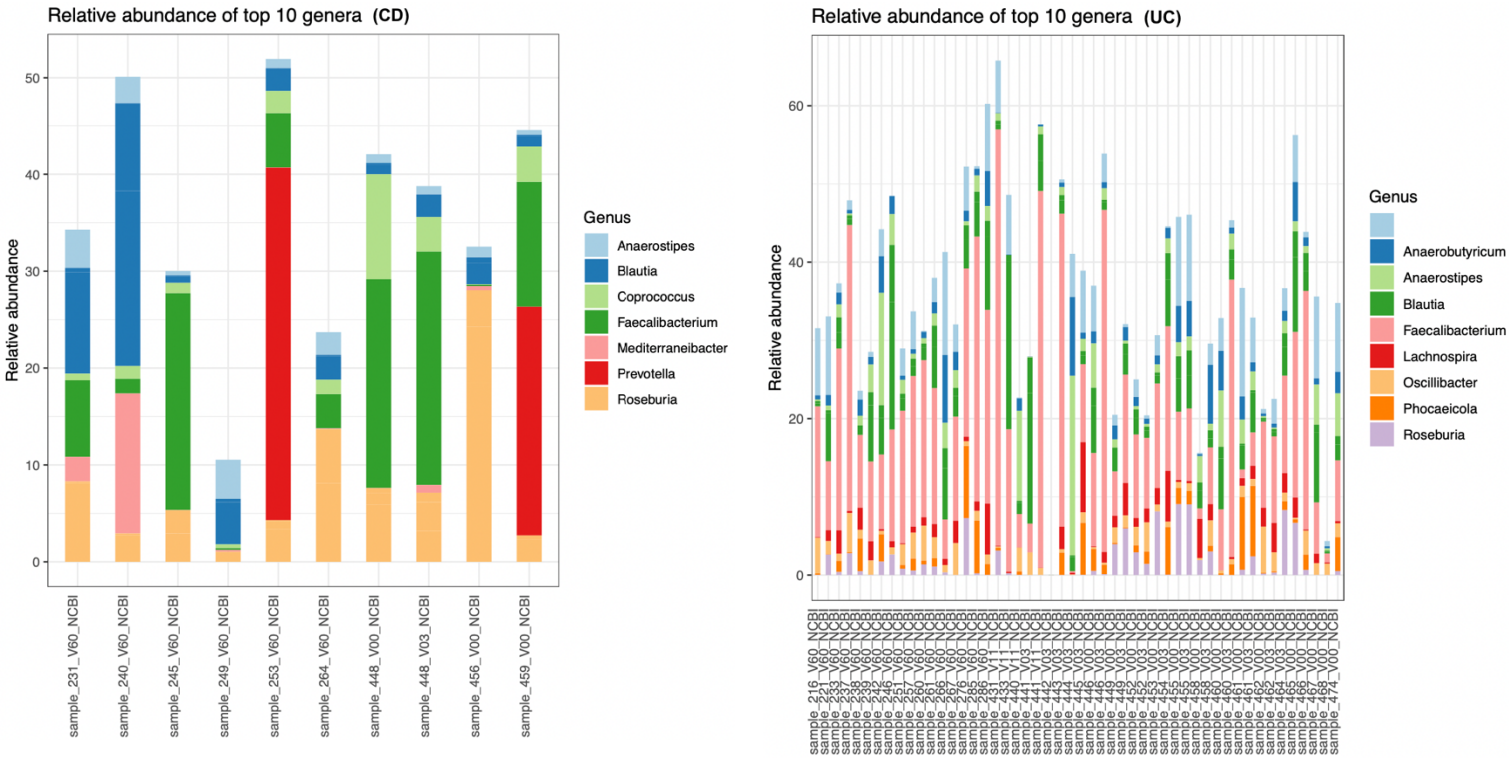


Figure 19. Top ten genera by origin by MinION sequencing. The two boxes represent the two groups; CD patients (left) and UC patients (right). The different genera are shown in the different colors displayed next to the plots. The y-axis is relative abundance, while the x-axis is the samples.

The genera abundance varies a lot between the CD patients, sample 231 V60 had most abundancy of *Roseburia*, *Faecalibacterium* and *Blautia*. Sample 240 V60 had a larger amount of *Blautia* and *Mediterraneibacter* and smaller amounts of *Roseburia*, *Faecalibacterium*, *Coprococcus* and *Anaerostipes*. Samples 253 V60 and 459 V0 had a larger number of *Prevotella*, *Faecalibacterium* and a small amount of *Roseburia*. Sample 448 V0 had a *Faecalibacterium* dominating composition at the time of diagnosis, but for 448 V3 (3 months after diagnosis), the amount of *Coprococcus* had decreased, while a small amount of *Mediterraneibacter* had appeared, causing a small change in the composition. Sample 249 V60 had low diversity compared to the rest of the samples, both in abundance and population.

For UC patients, the dominating genus is *Faecalibacterium*, followed by *Anaerostipes*, *Blautia* and *Roseburia*. The light blue is unclassified genera. UC patients had larger prevalence of *Phocaeicola*, *Lachnospira* and *Oscillibacter*.

3.6 *Blastocystis* results

About 69 samples from patients were sent to the Staten Serum Institut in Copenhagen for Illumina MiSeq sequencing. The sequencing results shows that 14% of the 69 samples were *Blastocystis* positive. Provided by SSI were the sequencing results, metadata, and a script to provide information about alpha diversity, beta diversity top-ten most genera and patterns in samples that were *Blastocystis* positive and negative.

3.6.1 Alpha diversity

The diversity within the samples, observed richness, and the total number of different species in the sample. The diversity between the two groups (*Blastocystis* positive/negative) was determined using alpha-diversity indexes. Species richness and evenness within the two groups can be visualized in Figure 20, which displays both the Shannon index and the observed index.

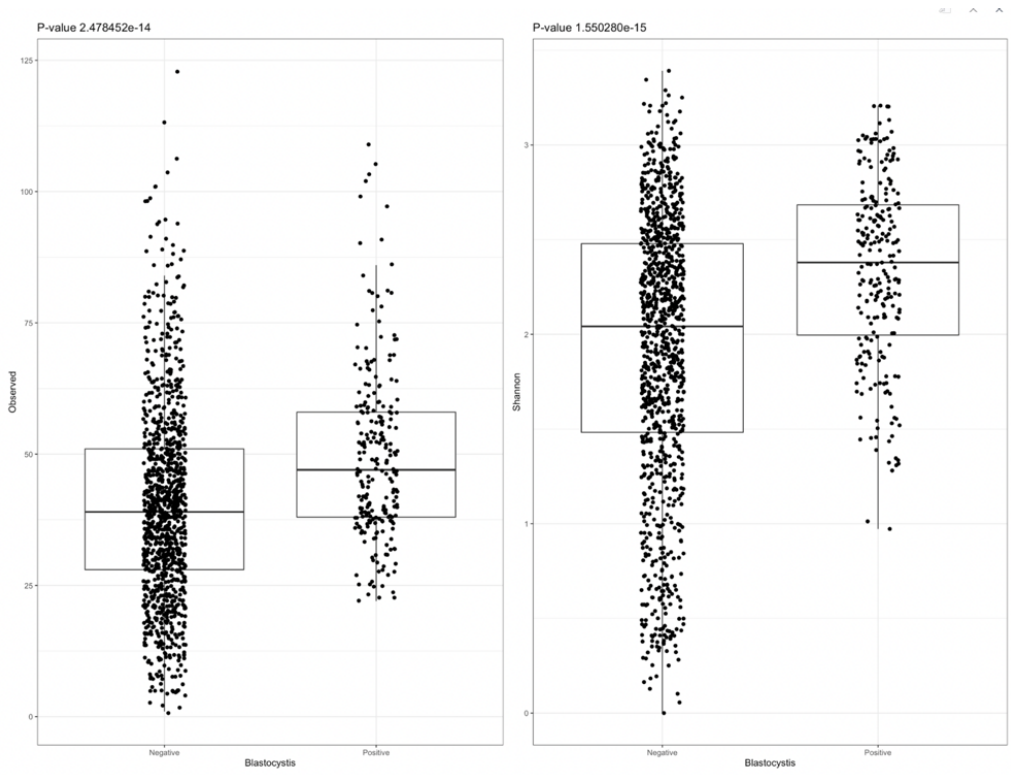


Figure 20. Alpha-diversity indexes for *Blastocystis* positive and negative samples from Illumina MiSeq sequencing. The alpha-diversity indexes illustrated are species richness (observed species) in a), Shannon-Wiener index in b). The y-axis in a) represent the number of unique species observed within the groups, while for b) it shows Shannon-Wiener index. The x-axis is showing both groups, *Blastocystis* positive and negative.

There is a higher number of observed species in patients that are *Blastocystis* positive, while the *Blastocystis* negative shows fewer unique species ($p < 0.05$; Kruskal-Wallis). The highest number of observed species is about 125, but the clustering is more prevalent at 25-100.
The Shannon diversity index for the positive group is generally higher than the for the negative. The negative group has diversity ranged from 0 to 3.5. The positive group varied from 1-3.5 but more evenly distributed at a higher value than the negative. The Shannon-Wiener index represents both unique species and their evenness ($p < 0.05$; Kruskal Wallis).

3.6.2 Beta diversity

For the beta diversity, we measure the differences between the two groups (*Blastocystis* positive/negative). The beta-diversity indexes presented in Figure 21 derived from the pipeline provided by the lab from Copenhagen (SSI) described previously. They are presented as Principal Component Analysis (PCoA) plots. The Bray-Curtis dissimilarity index is used as a measure of dissimilarity, and it is not a distance measure.

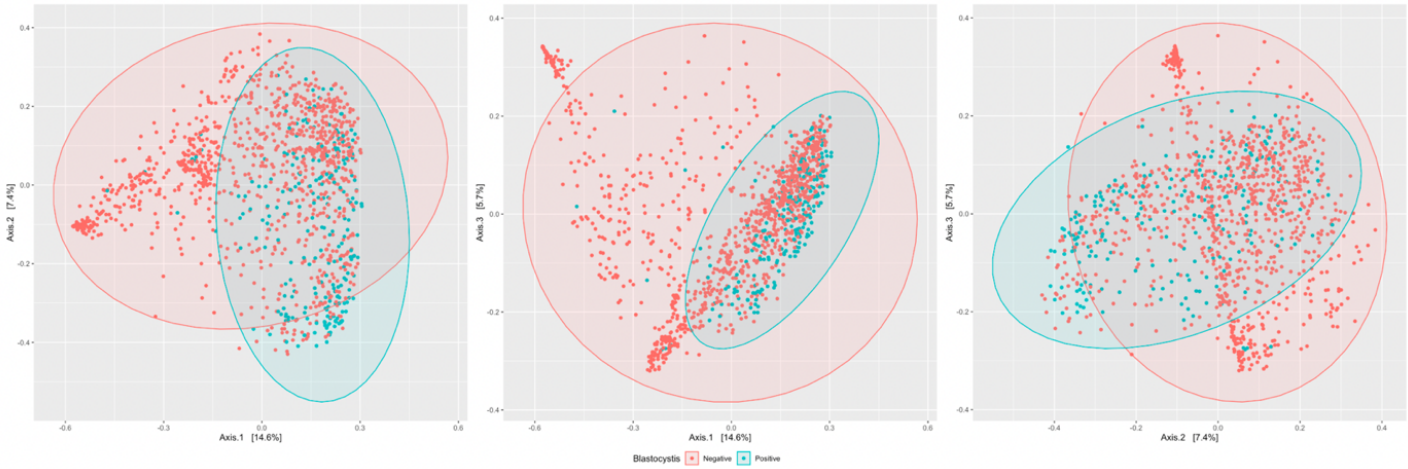


Figure 21. Beta-diversity indexes of *Blastocystis* negative/positive samples, sequenced with Illumina MiSeq. The figure illustrates the beta-diversity index Bray-Curtis in a PCoA plots. The data points represent the X samples, and the different colors represent the two groups: Red = *Blastocystis* negative, Blue = *Blastocystis* positive. For each axis, in square brackets, the percent of variation explained is reported.

The PCoA plots shows that there is not too much variation between the *Blastocystis* positive and negative samples ($p < 0.05$, PERMANOVA). The larger number of negative samples is more spread out, showing some dissimilarity, while the Blastocystis positive ones are less spread out, indicating less dissimilarities between the samples.

3.6.3 Dominating groups of bacteria – Illumina MiSeq

The relative abundance of top ten phyla by origin for the two groups is showed in Figure 27 (Appendix). The plot showing the total top ten phyla for all samples by origin, displays a wide array of colors representing the different phyla. From only inspecting the color variations, the top phylum for both groups are Firmicutes, Bacteroidetes and Proteobacteria.

We wanted to see which top ten genera for the *Blastocystis* positive and negative groups. This is shown in Figure 22. The figure shows the variation of most abundant genera in correlation with *Blastocystis*.

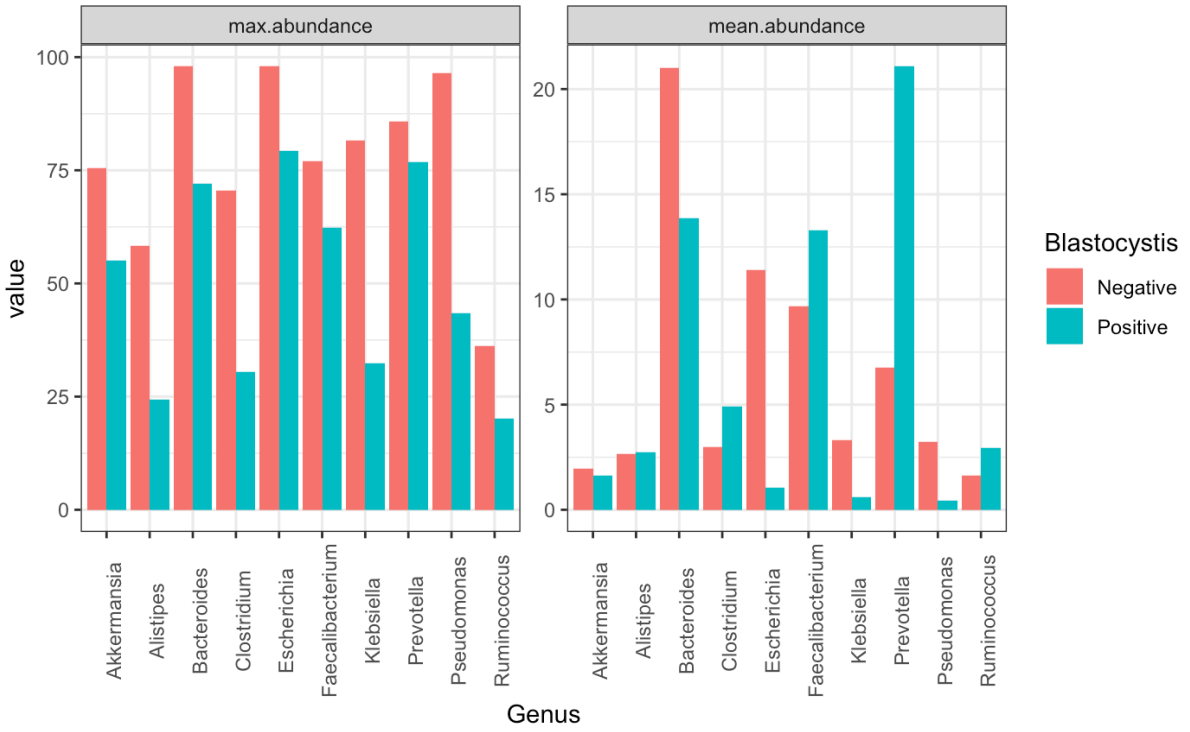


Figure 22. Top ten most abundant genera for the *Blastocystis* negative and positive groups by Illumina sequencing. The two colors represent the two groups of interest. Red = *Blastocystis* negative, Blue = *Blastocystis* positive. Plot a) shows max. abundance and plot b) shows mean abundance on genus level. The y-axis is the amount of each genus, the x-axis shows the genus names.

Max. abundance shows that *Bacteroides*, *Escherichia* and *Prevotella* are the dominating genera for the negative and positive *Blastocystis* samples. The *Blastocystis* positives have lower diversity and wider spread among the different genera. For the mean abundance, *Bacteroides* is the most abundant genus for the negative samples, followed by *Escherichia* and *Faecalibacterium*. For the positive *Blastocystis*, *Prevotella* is the dominating genus by far, followed by *Bacteroides* and *Faecalibacterium*. In the plot for mean abundance, the *Blastocystis* positive have a lower occurrence of all genera, while the *Blastocystis* negative have a higher occurrence. The plot for mean abundance shows that there is higher prevalence of *Bacteroides* and *Escherichia* but a lower amount of *Faecalibacterium* for the *Blastocystis*

negative, while for the *Blastocystis* positive, there is a higher prevalence of *Prevotella* and *Faecalibacterium*. The rest of the phyla are evenly distributed.

3.6.4 Dominating groups of bacteria – ONT MinION

The two groups of interest, *Blastocystis* positive/negative, were also analyzed using ONT sequencing technologies. The most abundant phyla for the two groups are shown in Figure 23. The top ten genera was classified using the data from MinION sequencing, as shown in the figure below.

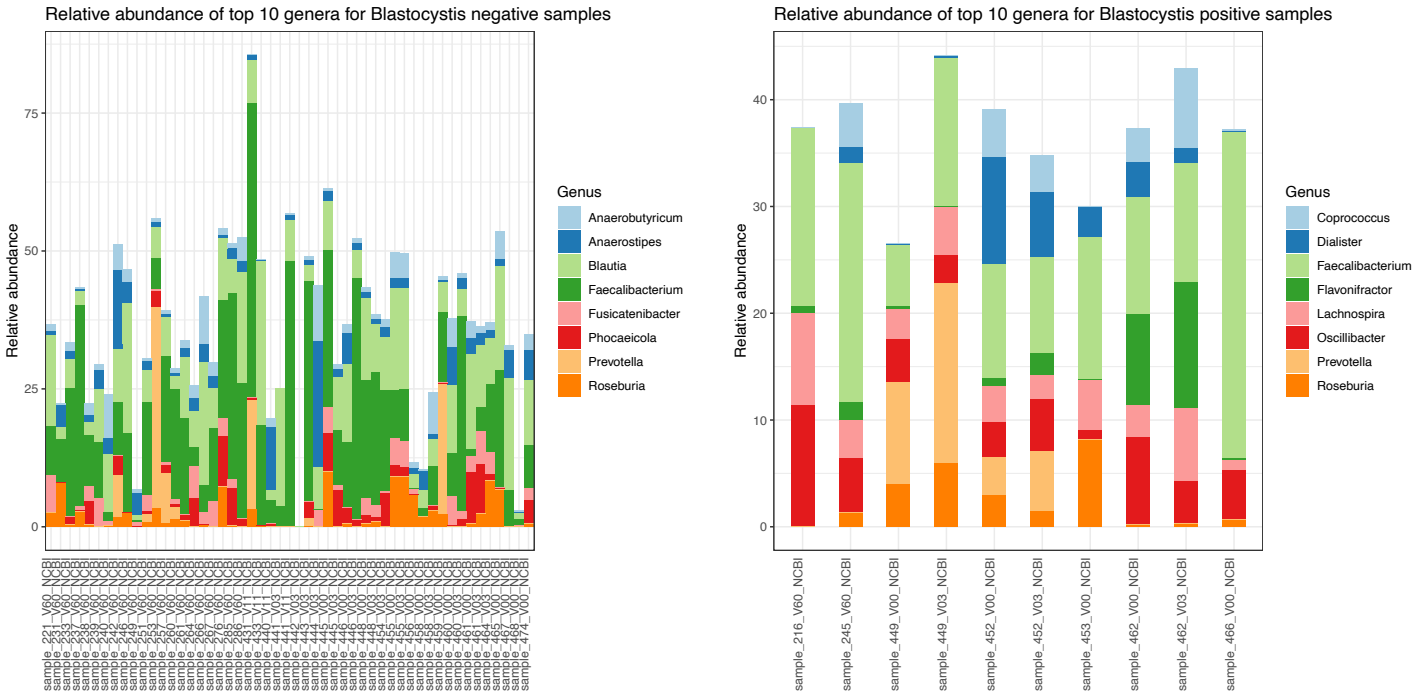


Figure 23. Top ten genus abundance for *Blastocystis* positive and negative, sequenced with MinION from ONT. The different colors represent the different genera. Plot a (left) is *Blastocystis* negative samples, while plot b (right) is *Blastocystis* positive. Y-axis is the abundance; x-axis is sample names.

For the *Blastocystis* negative, the most abundant genera is *Faecalibacterium* and *Blautia*. The other genera differ in distribution, some samples have a high number of *Prevotella*, while others have increased numbers of *Roseburia* and *Anaerostipes*. The most abundant genera for *Blastocystis* positive is *Faecalibacterium*, *Oscillibacter*, *Prevotella* and *Roseburia*. Samples 216 V60 and 462 V0 had a larger amount of *Oscillibacter* compared to the rest, while sample 449 V0 had an increase in *Roseburia*, *Prevotella* and *Faecalibacterium* from V0 to V3. The Blastocystis positive samples had a total of 9 patients diagnosed with UC and 1 patient with CD.

4 Discussion

IBD, or inflammatory bowel disease, refers to a group of diseases that affect the intestines and cause chronic inflammation of the bowels and digestive tract (Khor *et al.*, 2011). The increasing prevalence of IBD causes a substantial cost for healthcare as well as the general life quality of the affected patients. The etiology of IBD is somewhat unknown, there are thought to be several factors contributing to rising number of IBD incidents, these factors include genetic, bacterial, and environmental factors. A change in one of these factors can accelerate the prevalence of IBD (Khan *et al.*, 2019). The composition of the gut microbiota is a diverse community, and its effect on gut health is believed to have a great impact. Using NGS technologies, screening of the intestinal microbiome is possible even down to species level. The major aim of this study was to sequence the gut microbiome of patients diagnosed with IBD and to evaluate the presence of microbes in the gut of patients diagnosed with ulcerative colitis or Crohns disease. The samples were sequenced after DNA extraction and library preparation, and the sequencing data was analyzed using pipelines designed for NGS. The samples were also examined for the presence of *Blastocystis*, where 14% of all the samples were positive for the parasite. One important subgoal to this study was to assess the extraction method used for DNA extraction, where the quality of the DNA was measured using two different quantitation methods as well as adding a mock community to the sequencing pipeline. A larger amount of the samples had a satisfactory quality of genomic material for sequencing, while the samples with lower quality and yield was purified or re-extracted before sequencing. The sequencing results shows a wide diversity of bacteria, and the gut microbiome of the patients with IBD does vary regarding the composition and diversity between the different samples.

4.1 Genomic material used for sequencing

NGS has revolutionized genomic research by allowing complete genomes to be sequenced in a single day (Park *et al.*, 2016). This has resulted in significant improvements in disease diagnosis, prognosis, therapy, and solutions to genetic questions from various applications and biological systems (Boers *et al.*, 2019). The rising demand for NGS often puts pressure on upstream systems to process more samples and provide high-quality DNA for library prep and analysis (Phillips *et al.*, 2018). Low-quality genomic material can lead to poor performance and even failed sequencing runs. As a result, it's critical to improve the DNA extraction procedure such that it consistently produces reliable and reproducible DNA quality. Some of the

requirements for good quality genomic material used for NGS involve intact genomic DNA, no contamination of RNA or proteins, and a decent amount of genomic material. The genomic DNA used in downstream next-generation or third-generation sequencing should be intact and unsheared, with a length of at least 50 kB and no significant smears across the lanes/wells indicating the existence of smaller fragments in the sample. Some of the DNA samples were tested on an agarose gel to see if the genomic DNA had been fragmented during DNA extraction. All samples were tested on a gel, and showed no fragmentation, indicating that the genomic material was intact and at the correct size, as shown in Figure 10, which shows a gel electrophoresis including DNA from human stool samples. The concentration of RNA in genomic DNA used for sequencing should be as low as possible, preferably none. While RNA may not hinder the workflow or the actual NGS sequencing process, it absorbs UV light at the same wavelength as DNA, resulting in errors in spectrophotometric estimations of the amount of DNA. All samples were quantified using NanoDrop and Qubit fluorometer which gave an estimate of contaminants and the relative quality. The amount of DNA is important for library preparation prior to sequencing, so a wrong estimate of quantity can impact the sequencing results.

Because proteins absorb light at around 280 nm and nucleic acids at around 260 nm, the ratio of absorbance at these two wavelengths may be used to determine DNA purity. A A260/280 ratio of 1.8 or higher is generally considered to indicate good DNA purity, about 21% of the extracted samples fell under the value of 1.8, some of them were of decent quality but most of them were purified or re-extracted. Pure RNA has an A260/280 ratio of 2.0. Lower than optimal ratios can suggest the presence of leftover phenol or another reagent from the extraction procedure, or an unsuitably low nucleic acid content (less than 10 ng/µL). Some of the samples measured on the NanoDrop had an A260/280 value over 2.0 but these samples were either re-extracted or purified to try to get the value to around 1.8. About 41% of the samples had a A260/230 value over 2.0, while 66% had a value below 2.0. Samples with a dissatisfactory value on either A260/280 or A260/230 were not used for sequencing as any contaminants and impurities could block the pores in the flow cell and disrupt and affect the outcome of the sequencing.

The last sequencing runs performed (SUS_7 and SUS_8) used the samples with the lowest quality. The flow cells health decreased faster than with earlier runs. The pores were blocked from the poor genomic materials of SUS_7, resulting in poor coverage of SUS_8. This could be avoided by not using the DNA with poor quality for sequencing or to shorten the sequencing time in the first run (SUS_7). To provide successful results in subsequent stages of the NGS

workflow or other downstream applications such as third-generation sequencing and genotyping, extraction protocols must yield sufficiently high DNA yields and concentrations, regardless of the technique.

4.2 Technical considerations and evaluations

4.2.1 Extraction protocol evaluation

Samples delivered to SUS are collected in a feces sample collection tube. It is up to each patient to collect their own stool samples; instructions are given in the collection kit. One setback using this method for collection is that a symptom of many gastrointestinal diseases is diarrhea, which can make collection difficult. Some of the samples have visibly a larger amount of feces in the collection tube while some samples have very little feces, making the sample tube grey/transparent as the tube consists mostly of preservation liquid. At first, the grey/transparent samples were thought to have lower quality after DNA extraction, but that was not always the case. The samples with smaller amounts of visible feces were thawed longer and homogenized thoroughly before DNA extraction. This extra step worked for most of the transparent samples, but not all, indicating that there is a low number of bacteria present in the sample.

The ZymoBIOMICS mock community control had a lower yield than desirable, indicating that there might be steps in the extraction protocol that might not be as efficient as described in the next part. The MinION was able to pick up a wide array of bacteria, all the way down to species level, meaning that some DNA was still intact, and after performing PCR, the genomic material was good enough to produce a decent result. The amount of reads for the sequencing run that included the control was low, due to a flow cell with reduced quality and this might have affected the outcome of the sequencing.

In a study conducted by Costea *et al.* (2017) they looked into 21 different DNA extraction protocols using aliquots of the same fecal sample to look at differences in the observed microbial community composition. They compared the composition with differences due to library preparation, sample storage and extraction, which they contrasted with the observed variation within the same sample over time. As a result, they concluded that DNA extraction had the greatest effect on the outcome of the metagenomic analysis. After reviewing the 21 protocols, they concluded that Protocol Q (IHMS) to suit most applications and to be the most reproducible (Costea *et al.*, 2017). Based on the results from sequencing of the ZymoBIOMICS control mock community, the taxonomic dispersion is good for the sequenced sample. There are similarities in the distribution between the sequenced control and the mock community described by ZymoBIOMICS in Figure 24. This suggests that

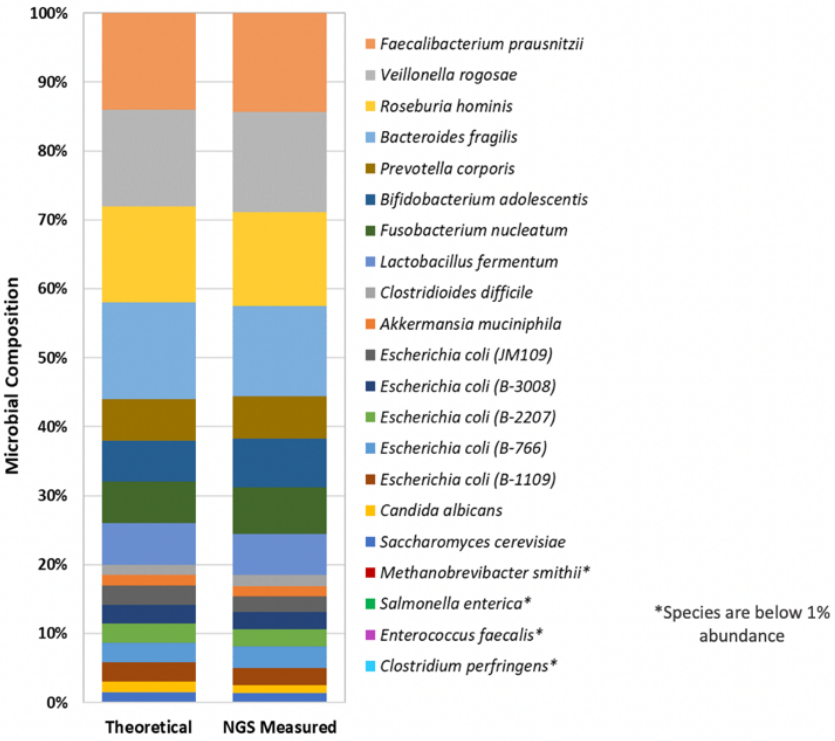
extraction using protocol Q, used in this study, might support the findings done by Costea *et al.* (2017) stating that the protocol is reproducible and give a good extraction yield. The beat beating process was altered by using a different bead beater, but not other modifications were done to the protocol.

4.2.2 Control evaluation

In microbiomics and metagenomics research, microbial composition profiling techniques based on NGS are becoming more prevalent. It is usually determined that every step of the workflow, including DNA extraction, library preparation, sequencing, and bioinformatics analysis, can be prone to bias and errors. There is an urgent demand in the area for accurate reference materials, such as a mock microbial community with a predetermined composition, to test the efficacy of different metagenomic studies.

The ZymoBIOMICS gut microbiome standard consists of 18 different bacterial strains, 2 fungal strains, and 1 archaeal strain in mixed abundances, as shown Figure 24. This is to mimic the true gut microbiome. The standard represents multiple challenges for NGS pipelines, such as tough-to-lyse Gram-positive bacteria (like *Roseburia hominis*) to test lysis efficiency, genomes with a wide range of GC content to test sequencing coverage bias, low-abundance pathogenic organisms for detection limit assessment. The control also contains 5 different strains of *E. coli* to test the taxonomic resolution. Serving as a defined input, this standard can be used to guide the construction and optimization of entire workflows as a quality control. The microbial standard is accurately characterized and contains a low impurity (< 0.01%).

The expected yield for one prep of the ZymoBIOMICS standard is around 1 µg. Yields significantly lower than 1 µg may suggest inefficient lysis during DNA extraction or other insufficient steps during the extraction. The measured quality after DNA extraction was 4.7 ng/µL and the final elution volume was 140 µL. So, the standard is under the desired yield for an efficient standard but was still included in a sequencing run. The amount of DNA was significantly increased after the barcoding PCR, with a final concentration of 60 ng/µL.



Species
Faecalibacterium prausnitzii
Veillonella rogosae
Roseburia hominis
Bacteroides fragilis
Prevotella corporis
Bifidobacterium adolescentis
Fusobacterium nucleatum
Lactobacillus fermentum
Clostridioides difficile
Akkermansia muciniphila
Methanobrevibacter smithii
Salmonella enterica
Enterococcus faecalis
Clostridium perfringens
Escherichia coli (JM109)
Escherichia coli (B-3008)
Escherichia coli (B-2207)
Escherichia coli (B-766)
Escherichia coli (B-1109)
Candida albicans
Saccharomyces cerevisiae

Figure 24. ZymoBIOMICS gut microbiome composition. The plot (left side) shows the distribution of the control species, the table (right side) show the different species in the ZymoBIOMICS control, which is made to mock a real gut microbiome.

Below 1 µg indicates that the extraction method or the cell lysis is inefficient. The performance of bead beating in fecal DNA extraction is improved by concurrent homogenization of the material. This allows the lysis buffer to permeate the entire fecal sample, regardless of consistency. However, the mechanical disruption does have the drawback of shearing DNA, which restricts its usage in applications that require intact genomic material. After extraction, the measured quality was 4.7 ng in 140 µL (final elution volume after DNA extraction). After performing sequencing on the control sample, the different taxonomic ranks were assigned using the same pipeline as for the samples from patients. During taxonomic analysis in Rstudio (v 2021.09.2), the final output was set to species level, this is shown in Figure 8. The figure shows that the MinION sequencing was able to pick up a wide array of species, and compared with the standards known composition, as shown in Figure 24. The most abundant species was *Faecalibacterium prausnitzii*, *Veillonella rogosae* and *Clostridioides difficile.* This correlates to the known composition of

the mock community, and a total of 5 species are the same as in the control. Some species remained unclassified; this could be due to the lack of taxonomic information from the database used in the data analysis. It is a known problem that identification at species level may have a high error rate. Even though the concentration of DNA after extraction was poor, the PCR was proven to be efficient, as the total ng before sequencing was at 60 ng/μL. This might indicate that the lysis part of the extraction might not be sufficient, but that some DNA remains unsheared and complete through the DNA extraction. This control sample was sequenced on the very last run (SUS_8), using a flow cell with low pore occupancy, resulting in fewer reads than desirable. More sequencing runs including a control sample could have been performed to get more coverage and less room for error by using a flow cell of higher quality. Negative controls were not used for sequencing to evaluate DNA extraction or library preparation as the protocols for library preparation require genomic material with at least 10 ng. Samples with lower amount and quality will block pores during the sequencing run, and ultimately affect the other samples included in the same run. As a negative control, the preservation liquid from the stool sampling kit or preferably the elution buffer used for DNA extraction and purification could be good options for a negative control. Both liquids were measured using both NanoDrop and Qubit and had very low amounts of DNA/other contaminants, and the low amount of genomic material was not suited for the library preparation protocols and sequencing.

4.3 Sequencing results

Since the launch and beta-release of the MinION in 2012 and 2014 respectively, numerous studies and evaluations of the sequencing platform have been performed (Ip *et al.*, 2015). Both the chemical and technical aspects of the MinION have undergone some major changes to improve the performance of the platform (Jain *et al.*, 2018; Lu *et al.*, 2016). Laver *et al*. (2015) conducted a study, assessing the performance of the MinION where the estimated sequencing error rate was 38.2% after base calling. This is regarded as highly inaccurate sequencing (Laver *et al.*, 2015). Another study by Jain *et al.* (2016), concluded the error rate to be <8%, which is a major improvement and shows the positive influence of the improvements made to both the chemical part of library preparation and the software improvements made by ONT developers. This strengthens the MinION's potential (Jain *et al.*, 2016). Even though the rapid and continuous developments and improvements of sequencing technologies and associated software keeps coming, incorrect reads during the sequencing are still an issue for NGS

technologies, especially Nanopore reads (Ma *et al.*, 2017). Despite the incorrect reads and error rates, the constant improvements have made it possible to sequence a human genome and to obtain a complete and contiguous *de novo* assembly, as shown by Jain *et al.* (2017). One other major setback with NGS is the challenge to receive sequencing results and classification at species level, this is more due to the access to databases containing detailed data on species-level for taxonomic identification. The trimming and filtration of sequencing reads had a good effect, resulting in reads with a Q-score above 7, and only reads within 1200-1800 bp in length as shown in Figure 12. The results of the alpha diversity analyses showed that there was some variation in the diversity, with the number of observed species was varying greatly from below 1000 observed species to the top sample which had about 3000 unique observed species. The taxonomic analysis was done at phyla-level, but there was good coverage all the way down to species-level using the MinION.

4.3.1 Sequencing run analysis

The technology behind ONT sequencing allows for real-time sequencing with a live view of the sequencing quality. Pore occupation, run feedback, and data acquisition. The run feedback includes flow cell health, pore occupation, read length histogram, and channel status. The read depth and quality may vary for each sequencing run, as the different variables affecting the run can be anything from the input genomic material to the pore occupancy of the flow cell. All flow cells used for this study were checked before use to ensure they were of good quality and > 1000 pores were available at the time of sequencing. Flow cells will deteriorate slowly over time, resulting in fewer pores. All flow cells were kept unopened in the fridge (4 °C) for up to 12 weeks, which is recommended by ONT.

4.3.2 Quality control

Based on the Nanoplots shown in Figure 12, the trimming and filtering of the raw sequences proved to have good effect. The number of sequences below 1200 bp was cut out, along with anything over 1800 bp. The Q-score was set to 7, even though the quality was not under to begin with, the figure showing the quality after trimming and filtering ensured that all sequences were of good quality. The Guppy basecaller trims off barcodes, adapters, and primers, but to ensure that everything was trimmed off properly, Porechop was also used for further trimming to ensure that all chimeric reads were split up, as well are ensuring no left-over adapter, barcode or primer sequences.

4.3.3 Data acquisition and analysis

Taxonomic assessments of samples or communities is a key step in metagenomic studies. Assigning taxonomic ranks and labels to the sequenced reads and the composition of the microbial community are increasing in demand due to growing use of more modern sequencing technologies which demands more accurate and efficient tools for these types of metagenomic analyses. Data created by technologies like ONT, long read sequences, enables for better taxonomic resolutions due to the high input of information the sequences contain. The lack of well-established tools and pipelines for long-read sequencing causes the users to rely on databases, pipelines and other tools designed for short reads, which do not work well with the long-read data (Ciuffreda *et al.*, 2021). Despite this setback, long-read data analysis tools are constantly being developed, including error correction and extended databases more suited for species-level classification. The more traditional classification methods are based on similarities between the newly sequenced genes and aligning it with already existing databases, containing taxonomic information about a variety of organisms, like NCBI or Greengenes. The database selection is important for metagenomic workflows, as these databases contain both partial- and full-length 16S rRNA gene sequences. Databases are constantly growing and expanding every year, this does not guarantee successful classification of the generated reads (Ciuffreda *et al.*, 2021; Santos *et al.*, 2020).

Data analysis using the pipeline shown in Figure 6 provided sequence reads with a desired length of 1200-1800 bp, and the coverage of the sequences were ideal. Kraken 2 created a sub database from NCBI, which was the foundation for taxonomic classification. After running all samples through the Phyloseq taxonomic analysis pipeline, taxonomic classification down to species level was possible and gave a good abundance at species level. However, due of MinION's high error rates, it's impossible to say how accurate the identification was (Rang *et al.*, 2018).

4.4 The gut microbiome of IBD patients

The microbial profiles in inflammatory conditions have been explored in a great number of studies. The reduction of microbiota diversity is a common incidence in inflammatory bowel disease (IBD), and studies show that microbial diversity is negatively correlated with disease severity in IBD. Microbial diversity is known to increase in disease remission. Species diversity is critical for sustaining the intestinal ecosystem's stability as well as appropriate ecological

function. A decrease in microbial diversity correlates with a decrease in ecosystem stability, which can affect ecological function. Several other studies have also discovered that IBD patients' fecal microbial communities were distinct from those of healthy individuals. The gut microbiota of patients with UC in remission was found to be similar to that of healthy people, indicating that the fecal microbiota plays distinct roles in the pathophysiology of UC and CD (Gong *et al.*, 2016). In another study by Hedin *et al*., samples from 21 patients with CD, along with 17 of their apparent healthy siblings and 19 unrelated controls were sequenced using pyrosequencing. The healthy siblings (aged 16-35 years) who volunteered and did not meet the exclusion criteria, which included previous diagnosis of IBD, and symptoms related to IBD were included to limit bias. At the time of the study, only one patient was living together with one of the included siblings. The results showed that the core microbiota of patients with CD and their healthy siblings was less diverse than the 19 unrelated controls. Healthy siblings with a higher risk of developing CD exhibited lower core microbial diversity than the 19 unrelated healthy controls (though the siblings' microbial diversity was higher than that of CD patients), suggesting that loss of core microbial diversity could be a critical step in the pathogenesis of CD. It's also apparent that sibling risk extends beyond genetics and that non-genetic variables in families have a role in the development of an at-risk microbiota (Hedin *et al.*, 2016). This study highlights the importance of using control groups in order to compare the IBD patients to apparently healthy individuals. Another interesting theory around the decrease in gut microbiome is the hygiene hypothesis. This theory states that the increasing number of chronic inflammatory disorders, hereunder IBD, is due to the lack of exposure to microorganisms that could potentially play an essential role in the immune system. The immunological functions of gut microbiota, as well as the existence of alterations in the microbiota because of diet, hygiene, and antibiotics, are well known. Due to the world progressing into a more modern lifestyle, the exposure to potentially essential microbes is decreasing. These potentially essential microbes can aid in the prevalence of IBD by interacting with the immunoregulation, leading to an increase in chronic inflammatory diseases (Rook, 2010). Some of the measures taken to restore this exposure is fecal microbiota transplantation (FMT), probiotics and prebiotics (Gong *et al.*, 2016). FMT therapy is currently being used to treat IBD. Norway is prevalent in the number of IBD cases worldwide (Lirhus *et al.*, 2021), as a country well adapted to a modern lifestyle, the hygiene hypothesis fits well with the ever-increasing number of IBD (and other chronic inflammatory diseases). The sequencing results showed that Firmicutes were the dominating phyla among all the samples. The amount of Firmicutes varied between patients, some had only a small number of other phyla while others

had mostly Firmicutes. Bacteroidetes was the second most common phylum, followed by a smaller amount of Proteobacteria and Verrumicrobiota. There are some variations in the number of phyla present in the samples, but Firmicutes and Bacteroides made up a relatively large proportion of the phyla detected for all samples.

4.4.1 UC *vs* CD

In a study conducted by Andoh *et al*. (2011), the microbial composition of the gut was investigated, and it was found that the abundance of Clostridium phylum was decreased in patients with UC and CD, while Bacteroidetes significantly increased in patients with CD (Andoh *et al.*, 2011). Chen *et al.* (2014) compared the intestinal microbiota between patients diagnosed with either CD or UC with healthy individuals. CD patients had significantly higher populations of *Streptococcus* and *Enterococcus*, but a decrease of *Roseburia* and *Faecalibacterium* compared to healthy controls. The abundance of *Bacteroides*, *Enterococcus*, *Blautia* and *Escherichia* genera was increased in UC patients, along with a decrease in the abundance of *Coprococcus*. The sequencing results from the MinION shows that UC patients has the highest prevalence of *Faecalibacterium*, *Blautia* and *Anaerostipes*, which in some way correlates with the finding Chen *et al.* did. The patients with CD had larger population of *Prevotella* and *Faecalibacterium*, while the other genera varied some between the samples. The UC patients had some unclassified genera as well, this could be because of the database used for taxonomic identification.

4.5 *Blastocystis*

About 14% of the samples submitted to the Illumina sequencing group in Denmark were *Blastocystis* positive. Although there is plenty of evidence of mixed subtype infections in the literature, the degree of *Blastocystis* diversity inside the host is mostly unknown. This is mostly owing to a scarcity of sensitive molecular methods capable of identifying *Blastocystis* mixed infections in a sample. Maloney *et al.* (2019) conducted a study to develop a next-generation amplicon sequencing protocol, this protocol would target a fragment of the SSU rRNA gene that includes a pipeline for analysis to detect infection and subtypes. They also compared Sanger sequencing to NGS (Illumina) in detecting *Blastocystis* infection and subtypes. They compared NGS and Sanger sequencing with cloning in 75 *Blastocystis* positive fecal samples to detect *Blastocystis* subtypes and within-host genetic variability. Their findings revealed that NGS was as accurate as Sanger sequencing for subtype detection, and NGS is a considerably

more sensitive method for identifying mixed infections and detecting low abundance subtypes (Maloney *et al.*, 2019).

In a study conducted by Krogsgaard *et al*. (2015), they established that the prevalence of intestinal parasites was not greater amongst individuals with IBS. *Blastocystis* has been detected in feces from patients diagnosed with IBS, and parasites like *Blastocystis* is believed to be involved in the pathogenesis of IBS, as well as other GI diseases, hereunder IBD. They established that a greater proportion of the controls carried parasites rather than the diagnosed patients (Krogsgaard *et al.*, 2015). Given the current opportunities for thorough gut microbiota profiling using NGS, studying intestinal parasites in relation to their ecological niche, like relationships with gut microbiota, is an important step toward fine-tuning our clinical and public health understanding of colonization by intestinal parasites. These efforts are already being performed; nevertheless, it is equally critical to establish hypotheses that might explain these relationships (Stensvold *et al.*, 2018).

It is difficult to establish any significance to the number of patients with *Blastocystis* in this study, as controls of healthy people would be necessary. In a study by Petersen *et al*. (2013), they compared the prevalence of 100 patients diagnosed with IBD along with 96 samples from healthy controls. *Blastocystis* was detected by culturing and PCR, with the results of 19% prevalence in the healthy controls and only 5% in the IBD patients. The IBD analysis showed that *Blastocystis* was primarily found in the group of IBD patients with inactive UC (Petersen *et al.*, 2013). The IBD patients from SUS had a 14% prevalence of *Blastocystis*, which is significantly higher than what was found in the study by Petersen *et al*. (2013). Only one of the 10 *Blastocystis* positive had CD, meaning that most of the positive patients are diagnosed with UC.

Andersen *et al.* (2015) conducted a study in which the prevalence and distribution of subtypes of *Blastocystis* were assessed between various cohorts of healthy and diseased individuals. They also explored the link between the gut microbiota and *Blastocystis*. Their results showed that *Blastocystis* carriage was less common in individuals with a *Bacteroides*-dominating enterotype than in those with a higher number of *Ruminococcus*- or *Prevotella*-dominant enterotype (Andersen *et al.*, 2015). Based on the results from this study, as seen in Figure 22 that exploits the top 10 genera for *Blastocystis* positive and negative, there is a much higher prevalence of *Prevotella* in the *Blastocystis* positive subjects. The *Prevotella*-driven enterotype shows less correlation with *Blastocystis* positive samples. In the same study, it was observed that *Blastocystis* positive patients had a lower amount of *Bacteroides* in the gut. The 14% positive *Blastocystis* from this study also had a lower abundance of *Bacteroides*. When *Blastocystis*-

negative samples were compared to *Blastocystis* positive samples, the relative abundance of Bacteroides in the *Blastocystis* negative samples was significantly greater. This can also be viewed in Figure 26 (Appendix) as the *Blastocystis* negative box shows greener parts of the plots (*Bacteroides*) than the *Blastocystis* positive which has more *Prevotella* (orange). Since Illumina MiSeq sequencing is the current golden standard for sequencing due to high sequencing accuracy, it can be used as a reference point to assess the MinION reads.

4.5.1 Short-read *vs* long-read sequencing

The taxonomic identification of bacteria at phyla level showed some differences and similarities between the two sequencing techniques. First, the analysis pipelines are different; the Illumina MiSeq results show the top 10 phyla, while the ONT MinION results show all phyla detected in relation to abundance and amount of each phylum; when finding the top 10 phyla with the MinION reads, the results were more similar. In particular, Actinobacteria had greater abundance when analyzed with MiSeq, while Proteobacteria picked up similarly between the two technologies. The F/B ratio was also even between Illumina MiSeq and ONT MinION, both phyla showing the greatest abundance, with equal parts evenly distributed between the samples. One of the reasons why there is a difference in the abundance between the phyla may be due to the different reference databases used for analysis. For MinION analysis, we used NCBI sub-database downloaded from the NCBI database (Kraken 2). The database used for the Illumina sequencing is most likely a custom database they use for their public health screening. This probably caused some of the differences in the sequencing results, especially at a deeper taxonomic level.

4.6 Future work/prospects

The level of interest in this topic varies significantly across European countries and the United States. Promoting more studies on the relationship between gut microbiota and IBD is necessary (Kaur *et al.*, 2011). There are limitations to this study, one example being the lack of control samples to correlate the patients' microbiome to healthy individuals. The future of IBD research includes studying the role of reduced/increased gut microbial diversity, the attempts to restore the "normal" diversity to alleviate IBD, look at the potential factors contributing to a lowered microbial diversity, like the role of antibiotics, diets, and environments as well as genetic factors. One strength of this study is the number of patients available in the study, with samples taken at different time points to establish differences in their microbiome over time. The

metadata available for each patient is also something to dive into, from age and gender to more clinical aspects like treatment or antibiotic usage.

Further work could include clinical information about patients (age, gender, disease, dietary information) to compare the sequencing results. The metadata could provide information about groupings to potentially see patterns within patients undergoing treatment or not. The most cited drawback of NGS platforms, including MinION, is the high error rate. The MinION and the analysis tools used for TGS platforms are constantly developing, and software updates are posted frequently, reducing the error rate with each update. This is one of the reasons why the use of TGS sequencing platforms is increasing. Future work would benefit from the development of databases designed for TGS. The taxonomic identification is poor at the species level; increasing the coverage of information at the species level would greatly benefit the future of TGS. Long-read amplicon sequencing is rising within Metagenomic studies, and ONTs' MinION will eventually be a viable rival to SRS systems due to advancements in nanopore technology. Developing more databases covering species level designed for ONT sequencing will be of major significance to genomic sequencing research.

5 Conclusion

The main goal of this study was to assess the gut microbiome of patients diagnosed with inflammatory bowel disease by TGS using ONTs' MinION sequencer, as well as establish the efficacy and accuracy of TGS. Genomic material was extracted from stool samples of patients before library preparation and sequencing. The essential functions of bacteria and evidence of dysbiosis in IBD presented in this study suggest that dysbiotic intestinal microflora may play a role in the development of IBD.

The sequencing results showed that Firmicutes, Bacteroides, Proteobacteria, and Verrumicrobiota were the most abundant phyla among all samples. In addition, it was established a difference in the composition of the microbiome for patients with UC and CD down to genus level, showing some differences in the abundance of Faecalibacterium, Prevotella, and Roseburia, indicating that dysbiosis may be involved in the activity of IBD and that there may be differences between patients with CD and UC. Although the changed microbial profiles had no consistent findings across some of the previous studies, a common trait, a lower bacterial diversity, surfaced in most of the IBD patients.

The samples from the same patients were also sequenced using Illumina MiSeq to evaluate the presence of *Blastocystis* in the patients. Illumina MiSeq sequencing is the current golden standard due to its high accuracy sequencing. Results from the MinION sequencing was compared to the Illumina, showing some similarities in taxonomic classification at both phylum and genus level. Alpha- and beta diversity analyses showed that there is a difference in diversity between the samples, both within one sample and between the different samples. The SGS using Illumina MiSeq showed that 14% of the 69 patients were *Blastocystis* positive before further taxonomic analysis, where it was found that the top phylum for both *Blastocystis* positive and negative are Firmicutes, Bacteroidetes, and Proteobacteria. The *Blastocystis* positive samples had a lower population of Bacteroides and Faecalibacterium compared to the *Blastocystis* negative. It was concluded that there was little difference in the taxonomic resolution between Illumina MiSeq and Oxford MinION based on a phylum level analysis. The differences may be due to the databases available for the two sequencing technologies.

This is just a preliminary study to establish the best procedures to start analyzing a larger number of patients from the IBD cohort at SUS. In order to correlate the gut microbiome with the disease outcome, a large cohort study where patients will be monitored for a more extended period of time during treatment would be ideal.

References

Adewale. (2020). Will long-read sequencing technologies replace short-read sequencing technologies in the next 10 years? *African journal of laboratory medicine, 9*(1), 1340-1340. doi:http://dx.doi.org/10.4102/ajlm.v9i1.1340

Alfellani, Stensvold, Vidal-Lapiedra, Onuoha, Fagbenro-Beyioku, & Clark. (2013). Variable geographic distribution of Blastocystis subtypes and its potential implications. *Acta Trop, 126*(1), 11-18. doi:http://dx.doi.org/10.1016/j.actatropica.2012.12.011

Ancona, Petito, Iavarone, Petito, Galasso, Leonetti, et al. (2021). The gut-brain axis in irritable bowel syndrome and inflammatory bowel disease. *Dig Liver Dis, 53*(3), 298-305. doi:http://dx.doi.org/10.1016/j.dld.2020.11.026

Andersen, Bonde, Nielsen, & Stensvold. (2015). A retrospective metagenomics approach to studying Blastocystis. *FEMS Microbiology Ecology, 91*(7). doi:http://dx.doi.org/10.1093/femsec/fiv072

Anderson. (2017). Permutational Multivariate Analysis of Variance (PERMANOVA). 10.1002/9781118445112.stat07841, 1-15. doi:http://dx.doi.org/10.1002/9781118445112.stat07841

Andoh, Imaeda, Aomatsu, Inatomi, Bamba, Sasaki, et al. (2011). Comparison of the fecal microbiota profiles between ulcerative colitis and Crohn's disease using terminal restriction fragment length polymorphism analysis. *J Gastroenterol, 46*(4), 479-486. doi:http://dx.doi.org/10.1007/s00535-010-0368-4

Bernstein. (2015). Treatment of IBD: Where We Are and Where We Are Going. *Official journal of the American College of Gastroenterology | ACG, 110*(1). doi:http://dx.doi.org/10.1038/ajg.2014.357

Black, & Black. (2014). *Microbiology: Principles and Explorations, 9th Edition*: Wiley.9781118934753

Bleidorn. (2016). Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. *Systematics and Biodiversity, 14*(1), 1-8. doi:http://dx.doi.org/10.1080/14772000.2015.1099575

Boers, Jansen, & Hays. (2019). Understanding and overcoming the pitfalls and biases of next-generation sequencing (NGS) methods for use in the routine clinical microbiological diagnostic laboratory. *Eur J Clin Microbiol Infect Dis, 38*(6), 1059-1070. doi:http://dx.doi.org/10.1007/s10096-019-03520-3

Boža, Brejová, & Vinař. (2017). DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads. *PloS one, 12*(6), e0178751-e0178751. doi:http://dx.doi.org/10.1371/journal.pone.0178751

Buttó, & Haller. (2016). Dysbiosis in intestinal inflammation: Cause or consequence. *International Journal of Medical Microbiology, 306*(5), 302-309. doi:http://dx.doi.org/https://doi.org/10.1016/j.ijmm.2016.02.010

Campbell. (2015). *Biology : a global approach*.9789810638795 9810638795

Chen, Wang, Zhou, Ng, Li, Huang, et al. (2014). Characteristics of fecal and mucosa-associated microbiota in Chinese patients with inflammatory bowel disease. *Medicine (Baltimore), 93*(8), e51. doi:http://dx.doi.org/10.1097/md.0000000000000051

Ciuffreda, Rodríguez-Pérez, & Flores. (2021). Nanopore sequencing and its application to the study of microbial communities. *Computational and Structural Biotechnology Journal, 19*, 1497-1511. doi:http://dx.doi.org/https://doi.org/10.1016/j.csbj.2021.02.020

Clark, van der Giezen, Alfellani, & Stensvold. (2013). Recent developments in Blastocystis research. *Adv Parasitol, 82*, 1-32. doi:http://dx.doi.org/10.1016/b978-0-12-407706-5.00001-0

Cole, Wang, Fish, Chai, McGarrell, Sun, et al. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res, 42*(Database issue), D633-642. doi:http://dx.doi.org/10.1093/nar/gkt1244

Costea, Zeller, Sunagawa, Pelletier, Alberti, Levenez, et al. (2017). Towards standards for human fecal sample processing in metagenomic studies. *Nat Biotechnol, 35*(11), 1069-1076. doi:http://dx.doi.org/10.1038/nbt.3960

De Coster, D'Hert, Schultz, Cruts, & Van Broeckhoven. (2018). NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics, 34*(15), 2666-2669. doi:http://dx.doi.org/10.1093/bioinformatics/bty149

Deamer, Akeson, & Branton. (2016). Three decades of nanopore sequencing. *Nature Biotechnology, 34*(5), 518-524. doi:http://dx.doi.org/10.1038/nbt.3423

DeSantis, Hugenholtz, Larsen, Rojas, Brodie, Keller, et al. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and environmental microbiology, 72*(7), 5069-5072. doi:http://dx.doi.org/10.1128/aem.03006-05

Dilthey, Jain, Koren, & Phillippy. (2019). Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps. *Nature Communications, 10*(1), 3066. doi:http://dx.doi.org/10.1038/s41467-019-10934-2

Fang, Fu, & Wang. (2018). Protocol for Fecal Microbiota Transplantation in Inflammatory Bowel Disease: A Systematic Review and Meta-Analysis. *BioMed Research International, 2018*, 8941340-8941340. doi:http://dx.doi.org/10.1155/2018/8941340

Feng, Zhang, Ying, Wang, & Du. (2015). Nanopore-based fourth-generation DNA sequencing technology. *Genomics, proteomics & bioinformatics, 13*(1), 4-16. doi:http://dx.doi.org/10.1016/j.gpb.2015.01.009

Garibyan, & Avashia. (2013). Polymerase chain reaction. *J Invest Dermatol, 133*(3), 1-4. doi:http://dx.doi.org/10.1038/jid.2013.1

Gauthier, Vincent, Charette, & Derome. (2018). A brief history of bioinformatics. *Briefings in Bioinformatics, 20*(6), 1981-1996. doi:http://dx.doi.org/10.1093/bib/bby063

Gentekaki, Curtis, Stairs, Klimeš, Eliáš, Salas-Leiva, et al. (2017). Extreme genome diversity in the hyper-prevalent parasitic eukaryote Blastocystis. *PLoS Biol, 15*(9), e2003769. doi:http://dx.doi.org/10.1371/journal.pbio.2003769

Gong, Gong, Wang, Yu, & Dong. (2016). Involvement of Reduced Microbial Diversity in Inflammatory Bowel Disease. *Gastroenterology research and practice, 2016*, 6951091-6951091. doi:http://dx.doi.org/10.1155/2016/6951091

Guardiola, Uriz, Taberlet, Coissac, Wangensteen, & Turon. (2015). Deep-Sea, Deep-Sequencing: Metabarcoding Extracellular DNA from Sediments of Marine Canyons. *PloS one, 10*(10), e0139633. doi:http://dx.doi.org/10.1371/journal.pone.0139633

Hedin, van der Gast, Rogers, Cuthbertson, McCartney, Stagg, et al. (2016). Siblings of patients with Crohn's disease exhibit a biologically relevant dysbiosis in mucosal microbial metacommunities. *Gut, 65*(6), 944-953. doi:http://dx.doi.org/10.1136/gutjnl-2014-308896

Hoffman. (2019). Chapter 25 - Analysis of Variance. I. One-Way. In J. I. E. Hoffman (Ed.), *Basic Biostatistics for Medical and Biomedical Practitioners (Second Edition)* (https://doi.org/10.1016/B978-0-12-817084-7.00025-5pp. 391-417): Academic Press.

Ip, Loose, Tyson, de Cesare, Brown, Jain, et al. (2015). MinION Analysis and Reference Consortium: Phase 1 data release and analysis. *F1000Res, 4*, 1075. doi:http://dx.doi.org/10.12688/f1000research.7201.1

Jain, Koren, Miga, Quick, Rand, Sasani, et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology, 36*(4), 338-345. doi:http://dx.doi.org/10.1038/nbt.4060

Jain, Olsen, Paten, & Akeson. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology, 17*(1), 239-239. doi:http://dx.doi.org/10.1186/s13059-016-1103-0
Johnson, Spakowicz, Hong, Petersen, Demkowicz, Chen, et al. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat Commun, 10*(1), 5029. doi:http://dx.doi.org/10.1038/s41467-019-13036-1
Kaplan. (2015). The global burden of IBD: from 2015 to 2025. *Nat Rev Gastroenterol Hepatol, 12*(12), 720-727. doi:http://dx.doi.org/10.1038/nrgastro.2015.150
Kaur, Chen, Luther, & Kao. (2011). Intestinal dysbiosis in inflammatory bowel disease. *Gut Microbes, 2*(4), 211-216. doi:http://dx.doi.org/10.4161/gmic.2.4.17863
Khan, Ullah, Zha, Bai, Khan, Zhao, et al. (2019). Alteration of Gut Microbiota in Inflammatory Bowel Disease (IBD): Cause or Consequence? IBD Treatment Targeting the Gut Microbiome. *Pathogens (Basel, Switzerland), 8*(3), 126. doi:http://dx.doi.org/10.3390/pathogens8030126
Khan, Ullman, Ford, Abreu, Abadir, Marshall, et al. (2011). Antibiotic therapy in inflammatory bowel disease: a systematic review and meta-analysis. *Am J Gastroenterol, 106*(4), 661-673. doi:http://dx.doi.org/10.1038/ajg.2011.72
Khor, Gardet, & Xavier. (2011). Genetics and pathogenesis of inflammatory bowel disease. *Nature, 474*(7351), 307-317. doi:http://dx.doi.org/10.1038/nature10209
Kibegwa, Bett, Gachuiri, Stomeo, & Mujibi. (2020). A Comparison of Two DNA Metagenomic Bioinformatic Pipelines While Evaluating the Microbial Diversity in Feces of Tanzanian Small Holder Dairy Cattle. *BioMed Research International, 2020*, 2348560. doi:http://dx.doi.org/10.1155/2020/2348560
Kim, Shin, Guevarra, Lee, Kim, Seol, et al. (2017). Deciphering Diversity Indices for a Better Understanding of Microbial Communities. *J Microbiol Biotechnol, 27*(12), 2089-2093. doi:http://dx.doi.org/10.4014/jmb.1709.09027
Koboziev, Reinoso Webb, Furr, & Grisham. (2014). Role of the enteric microbiota in intestinal homeostasis and inflammation. *Free radical biology & medicine, 68*, 122-133. doi:http://dx.doi.org/10.1016/j.freeradbiomed.2013.11.008
Krogsgaard, Engsbro, Stensvold, Nielsen, & Bytzer. (2015). The Prevalence of Intestinal Parasites Is Not Greater Among Individuals With Irritable Bowel Syndrome: A Population-based Case-control Study. *Clinical Gastroenterology and Hepatology, 13*(3), 507-513.e502. doi:http://dx.doi.org/https://doi.org/10.1016/j.cgh.2014.07.065
Lavelle, & Sokol. (2020). Gut microbiota-derived metabolites as key actors in inflammatory bowel disease. *Nature Reviews Gastroenterology & Hepatology, 17*(4), 223-237. doi:http://dx.doi.org/10.1038/s41575-019-0258-z
Laver, Harrison, O'Neill, Moore, Farbos, Paszkiewicz, et al. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification, 3*, 1-8. doi:http://dx.doi.org/https://doi.org/10.1016/j.bdq.2015.02.001
Leggett, Alcon-Giner, Heavens, Caim, Brook, Kujawska, et al. (2020). Rapid MinION profiling of preterm microbiota and antimicrobial-resistant pathogens. *Nat Microbiol, 5*(3), 430-442. doi:http://dx.doi.org/10.1038/s41564-019-0626-z
Li, Jin, Zhang, Pan, Wu, Liu, et al. (2021). Recovery of human gut microbiota genomes with third-generation sequencing. *Cell Death & Disease, 12*(6), 569. doi:http://dx.doi.org/10.1038/s41419-021-03829-y
Lirhus, Høivik, Moum, Anisdahl, & Melberg. (2021). Incidence and Prevalence of Inflammatory Bowel Disease in Norway and the Impact of Different Case Definitions: A Nationwide Registry Study. *Clin Epidemiol, 13*, 287-294. doi:http://dx.doi.org/10.2147/clep.S303797

Lu, Breitwieser, Thielen, & Salzberg. (2017). Bracken: Estimating species abundance in metagenomics data. *PeerJ Computer Science, 3*, e104. doi:http://dx.doi.org/10.7717/peerj-cs.104

Lu, Giordano, & Ning. (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, proteomics & bioinformatics, 14*(5), 265-279. doi:http://dx.doi.org/10.1016/j.gpb.2016.05.004

Lu, & Salzberg. (2020). Ultrafast and accurate 16S rRNA microbial community analysis using Kraken 2. *Microbiome, 8*(1), 124. doi:http://dx.doi.org/10.1186/s40168-020-00900-2

Ma, Stachler, & Bibby. (2017). Evaluation of Oxford Nanopore MinION Sequencing for 16S rRNA Microbiome Characterization. *bioRxiv,* 10.1101/099960, 099960. doi:http://dx.doi.org/10.1101/099960

Madoui, Engelen, Cruaud, Belser, Bertrand, Alberti, et al. (2015). Genome assembly using Nanopore-guided long and error-free DNA reads. *BMC genomics, 16*(1), 327-327. doi:http://dx.doi.org/10.1186/s12864-015-1519-z

Maloney, Molokin, & Santin. (2019). Next generation amplicon sequencing improves detection of Blastocystis mixed subtype infections. *Infection, Genetics and Evolution, 73*, 119-125. doi:http://dx.doi.org/https://doi.org/10.1016/j.meegid.2019.04.013

Maloney, Molokin, & Santin. (2020). Use of Oxford Nanopore MinION to generate full-length sequences of the Blastocystis small subunit (SSU) rRNA gene. *Parasites & Vectors, 13*(1), 595. doi:http://dx.doi.org/10.1186/s13071-020-04484-6

Manichanh, Borruel, Casellas, & Guarner. (2012). The gut microbiota in IBD. *Nat Rev Gastroenterol Hepatol, 9*(10), 599-608. doi:http://dx.doi.org/10.1038/nrgastro.2012.152

Matsuo, Komiya, Yasumizu, Yasuoka, Mizushima, Takagi, et al. (2021). Full-length 16S rRNA gene amplicon analysis of human gut microbiota using MinION™ nanopore sequencing confers species-level resolution. *BMC Microbiology, 21*(1), 35. doi:http://dx.doi.org/10.1186/s12866-021-02094-5

McMurdie, & Holmes. (2013). phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PloS one, 8*(4), e61217. doi:http://dx.doi.org/10.1371/journal.pone.0061217

Nell, Suerbaum, & Josenhans. (2010). The impact of the microbiota on the pathogenesis of IBD: lessons from mouse infection models. *Nat Rev Microbiol, 8*(8), 564-577. doi:http://dx.doi.org/10.1038/nrmicro2403

Owczarek, Rodacki, Domagała-Rodacka, Cibor, & Mach. (2016). Diet and nutritional factors in inflammatory bowel diseases. *World journal of gastroenterology, 22*(3), 895-905. doi:http://dx.doi.org/10.3748/wjg.v22.i3.895

Park, & Kim. (2016). Trends in Next-Generation Sequencing and a New Era for Whole Genome Sequencing. *International neurourology journal, 20*(Suppl 2), S76-S83. doi:http://dx.doi.org/10.5213/inj.1632742.371

Peters, Spatharis, Dario, Dwyer, Roca, Kintner, et al. (2018). Environmental DNA: A New Low-Cost Monitoring Tool for Pathogens in Salmonid Aquaculture. *Front Microbiol, 9*, 3009. doi:http://dx.doi.org/10.3389/fmicb.2018.03009

Petersen, Martin, Moschetti, Kershaw, Tsongalis, & Kraft. (2019). Third-Generation Sequencing in the Clinical Laboratory: Exploring the Advantages and Challenges of Nanopore Sequencing. *Journal of Clinical Microbiology, 58*(1), e01315-01319. doi:http://dx.doi.org/doi:10.1128/JCM.01315-19

Petersen, Stensvold, Mirsepasi, Engberg, Friis-Møller, Porsbo, et al. (2013). Active ulcerative colitis associated with low prevalence of Blastocystis and Dientamoeba fragilis

infection. *Scand J Gastroenterol, 48*(5), 638-639. doi:http://dx.doi.org/10.3109/00365521.2013.780094
Phillips, & Douglas. (2018). The Global Market for Next-Generation Sequencing Tests Continues Its Torrid Pace. *The Journal of precision medicine, 4*, https://www.thejournalofprecisionmedicine.com/wp-content/uploads/2018/2011/Phillips-Online.pdf.
Pigneur, & Sokol. (2016). Fecal microbiota transplantation in inflammatory bowel disease: the quest for the holy grail. *Mucosal Immunol, 9*(6), 1360-1365. doi:http://dx.doi.org/10.1038/mi.2016.67
Plesivkova, Richards, & Harbison. (2018). A review of the potential of the MinION™ single-molecule sequencing system for forensic applications. *Wiley Interdisciplinary Reviews: Forensic Science, 1*. doi:http://dx.doi.org/10.1002/wfs2.1323
Quainoo, Coolen, van Hijum, Huynen, Melchers, van Schaik, et al. (2017). Whole-Genome Sequencing of Bacterial Pathogens: the Future of Nosocomial Outbreak Analysis. *Clin Microbiol Rev, 30*(4), 1015-1063. doi:http://dx.doi.org/10.1128/cmr.00016-17
Quast, Pruesse, Yilmaz, Gerken, Schweer, Yarza, et al. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res, 41*(Database issue), D590-596. doi:http://dx.doi.org/10.1093/nar/gks1219
Rang, Kloosterman, & de Ridder. (2018). From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol, 19*(1), 90. doi:http://dx.doi.org/10.1186/s13059-018-1462-9
Rodríguez, Murphy, Stanton, Ross, Kober, Juge, et al. (2015). The composition of the gut microbiota throughout life, with an emphasis on early life. *Microbial ecology in health and disease, 26*, 26050-26050. doi:http://dx.doi.org/10.3402/mehd.v26.26050
Rook. (2010). 99th Dahlem conference on infection, inflammation and chronic inflammatory disorders: darwinian medicine and the 'hygiene' or 'old friends' hypothesis. *Clin Exp Immunol, 160*(1), 70-79. doi:http://dx.doi.org/10.1111/j.1365-2249.2010.04133.x
Santos, van Aerle, Barrientos, & Martinez-Urtaza. (2020). Computational methods for 16S metabarcoding studies using Nanopore sequencing data. *Computational and Structural Biotechnology Journal, 18*, 296-305. doi:http://dx.doi.org/https://doi.org/10.1016/j.csbj.2020.01.005
Scanlan. (2012). Blastocystis: past pitfalls and future perspectives. *Trends in Parasitology, 28*(8), 327-334. doi:http://dx.doi.org/https://doi.org/10.1016/j.pt.2012.05.001
Schirmer, Garner, Vlamakis, & Xavier. (2019). Microbial genes and pathways in inflammatory bowel disease. *Nature Reviews Microbiology, 17*(8), 497-511. doi:http://dx.doi.org/10.1038/s41579-019-0213-6
Schroeder, & Jenkins. (2018). How robust are popular beta diversity indices to sampling error? *Ecosphere, 9*. doi:http://dx.doi.org/10.1002/ecs2.2100
Slatko, Gardner, & Ausubel. (2018). Overview of Next-Generation Sequencing Technologies. *Current protocols in molecular biology, 122*(1), e59-e59. doi:http://dx.doi.org/10.1002/cpmb.59
Stensvold, & Clark. (2016). Current status of Blastocystis: A personal view. *Parasitology International, 65*(6, Part B), 763-771. doi:http://dx.doi.org/https://doi.org/10.1016/j.parint.2016.05.015
Stensvold, & van der Giezen. (2018). Associations between Gut Microbiota and Common Luminal Intestinal Parasites. *Trends in Parasitology, 34*(5), 369-377. doi:http://dx.doi.org/https://doi.org/10.1016/j.pt.2018.02.004
Stojanov, Berlec, & Štrukelj. (2020). The Influence of Probiotics on the Firmicutes/Bacteroidetes Ratio in the Treatment of Obesity and Inflammatory Bowel

disease. *Microorganisms, 8*(11). doi:http://dx.doi.org/10.3390/microorganisms8111715
Sun, Huang, Zhang, Zhu, Haiminen, Carrieri, et al. (2021). Challenges in benchmarking metagenomic profilers. *Nature Methods, 18*(6), 618-626. doi:http://dx.doi.org/10.1038/s41592-021-01141-3
Tamboli, Neut, Desreumaux, & Colombel. (2004). Dysbiosis in inflammatory bowel disease. *Gut, 53*(1), 1-4. doi:http://dx.doi.org/10.1136/gut.53.1.1
Thukral. (2017). A review on measurement of Alpha diversity in biology. *Agricultural Research Journal, 54*(1), 1-10. doi:http://dx.doi.org/10.5958/2395-146x.2017.00001.1
Tomasello, Tralongo, Damiani, Sinagra, Di Trapani, Zeenny, et al. (2014). Dismicrobism in inflammatory bowel disease and colorectal cancer: changes in response of colocytes. *World journal of gastroenterology, 20*(48), 18121-18130. doi:http://dx.doi.org/10.3748/wjg.v20.i48.18121
Twyman, & Primrose. (2006). *Principles of Gene Manipulation and Genomics (Seventh Edition)*.1405135441
Vasiljevic, Lim, Humble, Seah, Kratzer, Morf, et al. (2021). Developmental validation of Oxford Nanopore Technology MinION sequence data and the NGSpeciesID bioinformatic pipeline for forensic genetic species identification. *Forensic Science International: Genetics, 53*, 102493. doi:http://dx.doi.org/https://doi.org/10.1016/j.fsigen.2021.102493
Vich Vila, Imhann, Collij, Jankipersadsing, Gurry, Mujagic, et al. (2018). Gut microbiota composition and functional changes in inflammatory bowel disease and irritable bowel syndrome. *Sci Transl Med, 10*(472). doi:http://dx.doi.org/10.1126/scitranslmed.aap8914
Wei, Hung, Kao, Lin, Lee, Chang, et al. (2020). Classification of Changes in the Fecal Microbiota Associated with Colonic Adenomatous Polyps Using a Long-Read Sequencing Platform. *Genes, 11*(11). doi:http://dx.doi.org/10.3390/genes11111374
Wick. Porechop *GitHub*. Retrieved from https://github.com/rrwick/Porechop
Willis. (2019). Rarefaction, Alpha Diversity, and Statistics. *Frontiers in Microbiology, 10*. doi:http://dx.doi.org/10.3389/fmicb.2019.02407
Wood, Lu, & Langmead. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology, 20*(1), 257. doi:http://dx.doi.org/10.1186/s13059-019-1891-0
Wood, & Salzberg. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology, 15*(3), R46. doi:http://dx.doi.org/10.1186/gb-2014-15-3-r46
Wu, Xiong, & Yu. (2015). Taxonomic resolutions based on 18S rRNA genes: a case study of subclass copepoda. *PloS one, 10*(6), e0131498-e0131498. doi:http://dx.doi.org/10.1371/journal.pone.0131498
Xavier, & Podolsky. (2007). Unravelling the pathogenesis of inflammatory bowel disease. *Nature, 448*(7152), 427-434. doi:http://dx.doi.org/10.1038/nature06005
Yeh, McNichol, Needham, Fichot, Berdjeb, & Fuhrman. (2021). Comprehensive single-PCR 16S and 18S rRNA community analysis validated with mock communities, and estimation of sequencing bias against 18S. *Environmental Microbiology, 23*. doi:http://dx.doi.org/10.1111/1462-2920.15553
Yu, & Rodriguez. (2017). Clinical presentation of Crohn's, ulcerative colitis, and indeterminate colitis: Symptoms, extraintestinal manifestations, and disease phenotypes. *Seminars in Pediatric Surgery, 26*(6), 349-355. doi:http://dx.doi.org/https://doi.org/10.1053/j.sempedsurg.2017.10.003

Zallot, Quilliot, Chevaux, Peyrin-Biroulet, Guéant-Rodriguez, Freling, et al. (2013). Dietary beliefs and behavior among inflammatory bowel disease patients. *Inflamm Bowel Dis, 19*(1), 66-72. doi:http://dx.doi.org/10.1002/ibd.22965

Zhang, & Li. (2014). Inflammatory bowel disease: pathogenesis. *World journal of gastroenterology, 20*(1), 91-99. doi:http://dx.doi.org/10.3748/wjg.v20.i1.91

Zhang, Wang, Wang, Wang, & Li. (2020). Estimate of the sequenced proportion of the global prokaryotic genome. *Microbiome, 8*(1), 134. doi:http://dx.doi.org/10.1186/s40168-020-00903-z

Appendix A – Supplementary material

A. Top 10 species abundance for all SUS samples by MinION sequencing.

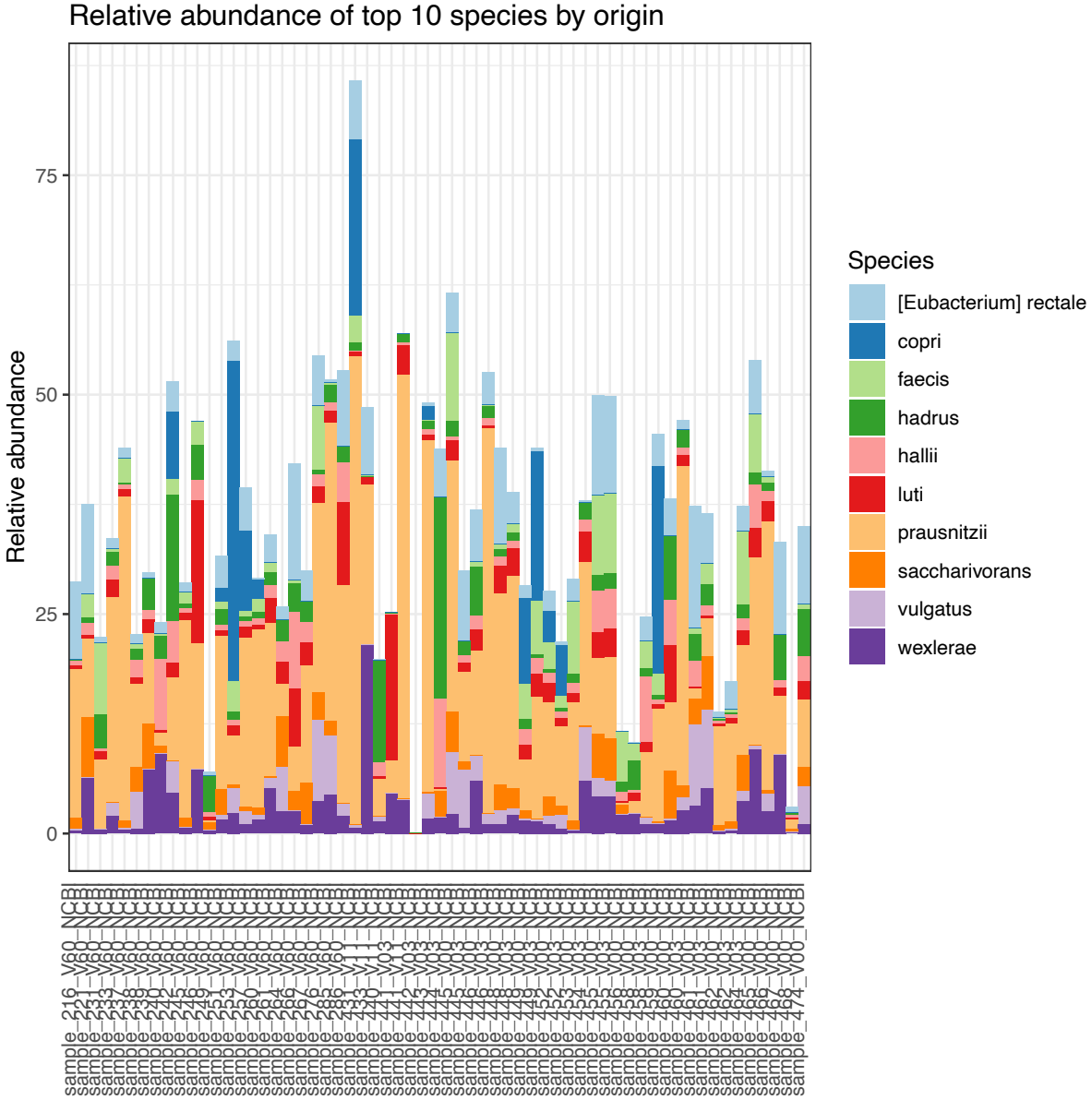


Figure 25. Top 10 species by origin for all SUS samples, sequenced with MinION. The different genera are shown in the different colors displayed next to the plots. The y-axis is relative abundance, while the x-axis is the samples.

B. Top 10 genus across the two different groups of *Blastocystis* patients (*Blastocystis* positive/negative).

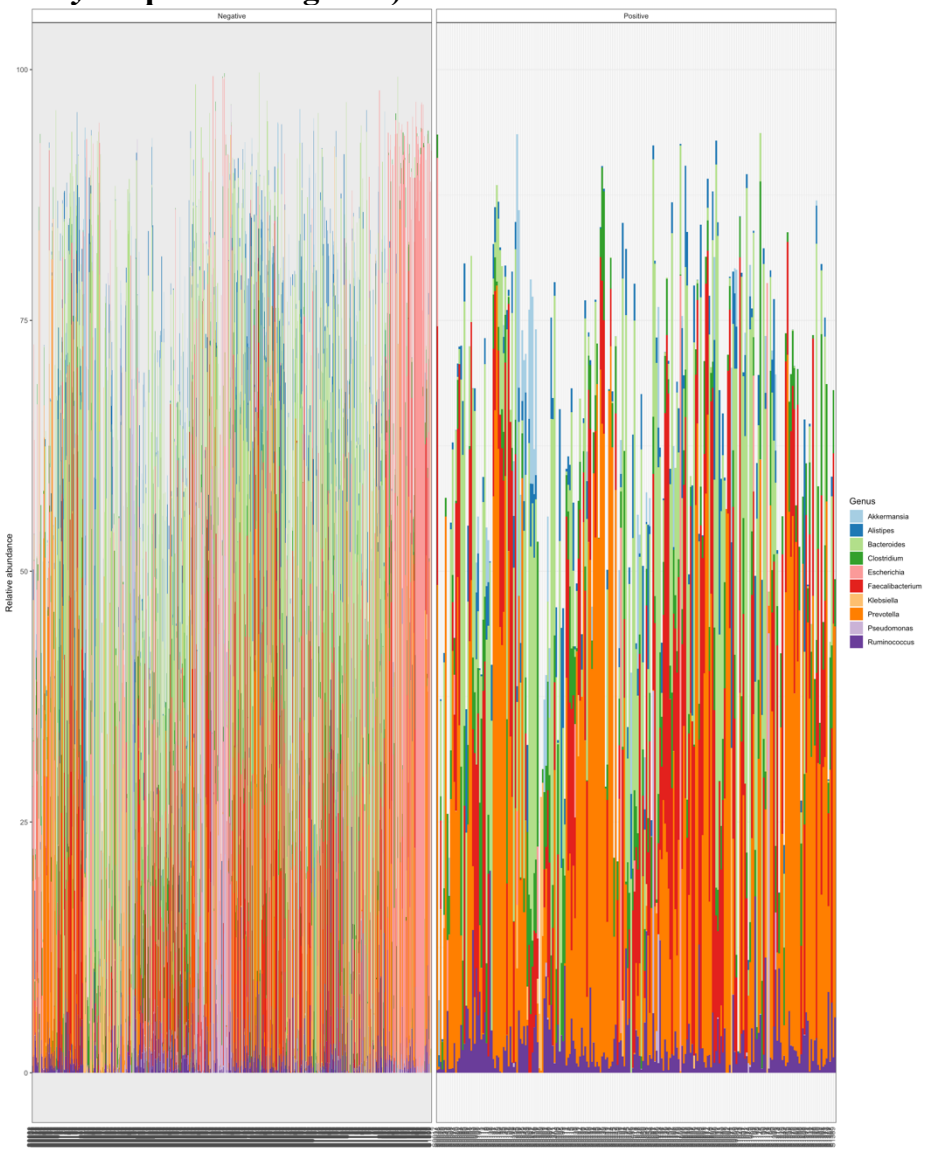


Figure 26. Top 10 genera by origin, sequenced with Illumina MiSeq. The two boxes represent the two groups; *Blastocystis* negative (left) and *Blastocystis* positive (right). The different genera are shown in the different colors displayed next to the plots. The y-axis is relative abundance, while the x-axis is the samples.

C. Top 10 phyla across the two different groups of *Blastocystis* patients (*Blastocystis* positive/negative).

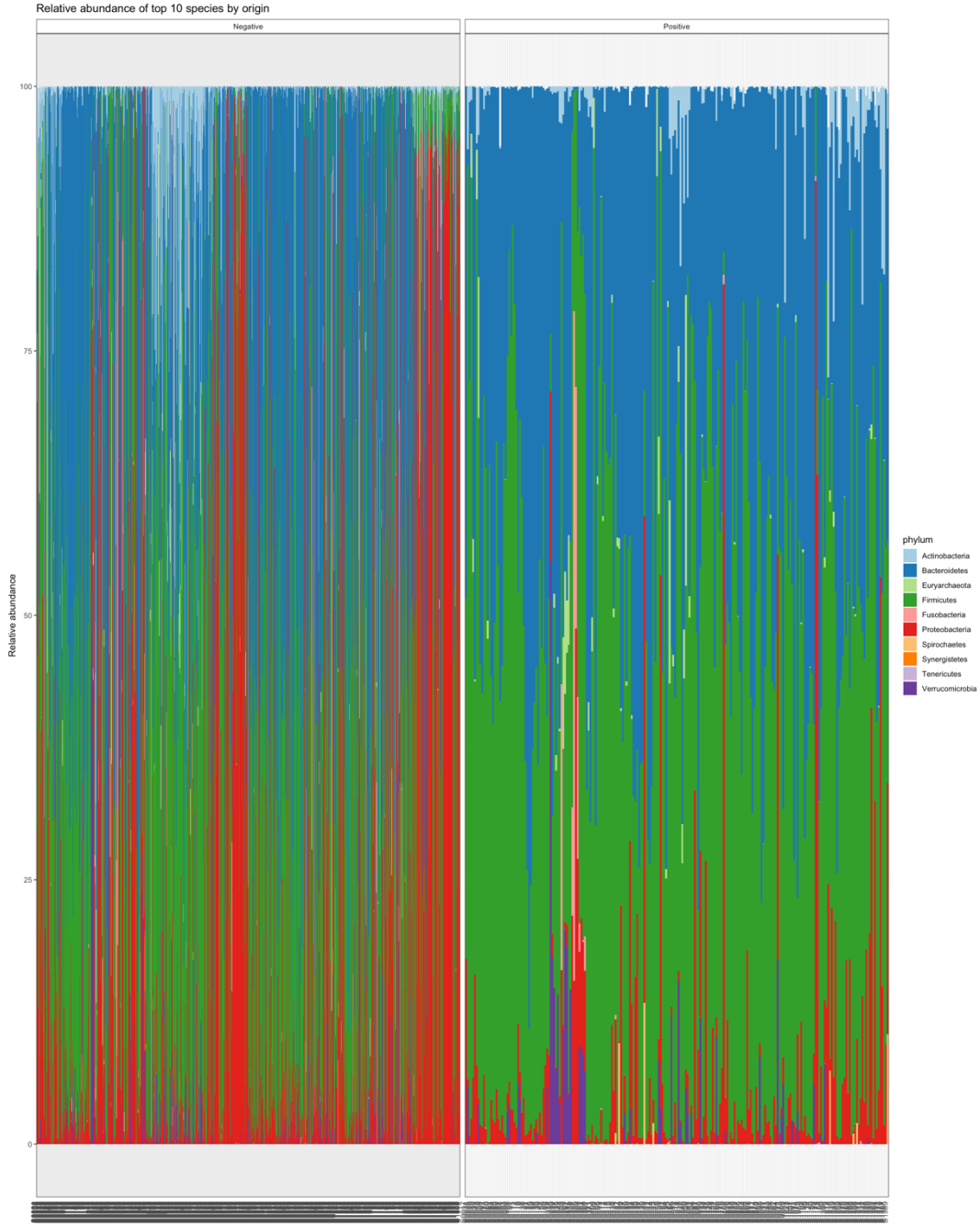


Figure 27. Top 10 phyla by origin between *Blastocystis* positive and negative samples, sequenced by Illumina MiSeq. The two boxes represent the two groups; *Blastocystis* negative (left) and *Blastocystis* positive (right). The different phyla are shown in the different colors next to the plots. The y-axis is relative abundance, while the x-axis is the samples.

D. Top 10 abundant phyla for *Blastocystis* positive/negative by Illumina sequencing.

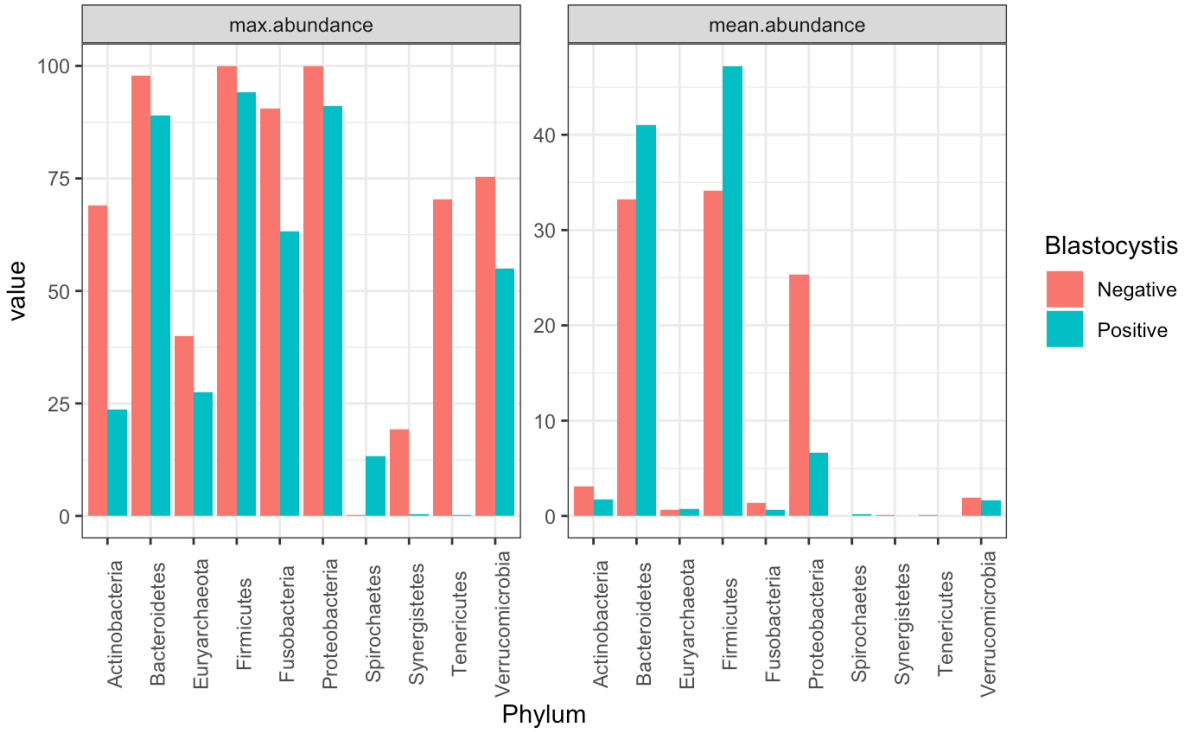


Figure 28. Top 10 most abundant phyla for the *Blastocystis* negative and positive groups by Illumina sequencing. The two colors represent the two groups of interest. Red = *Blastocystis* negative, Blue = *Blastocystis* positive. Plot a) shows max. abundance, b) shows mean abundance on phylum level. The y-axis is the amount of each phylum; the x-axis shows the phylum names.

E. Most abundant phyla for *Blastocystis* positive/negative by MinION sequencing.

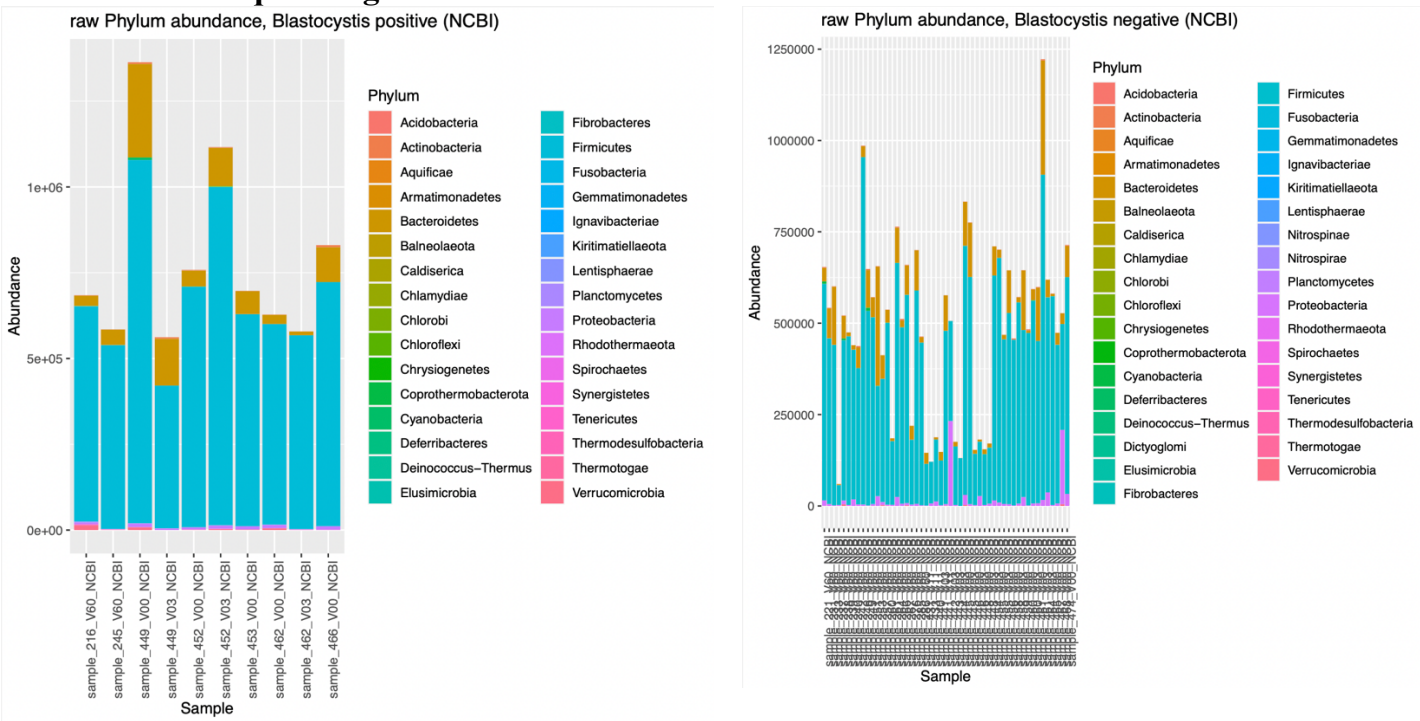


Figure 29. Raw phylum abundance for *Blastocystis* positive and negative, sequenced with MinION from ONT. The different colors represent the different phyla. Plot a (left) is *Blastocystis* positive samples, while plot b (right) is *Blastocystis* negative. Y-axis is the abundance; the -axis is the sample names.

F. Beta-diversity of Blastocystis positive and negative samples.

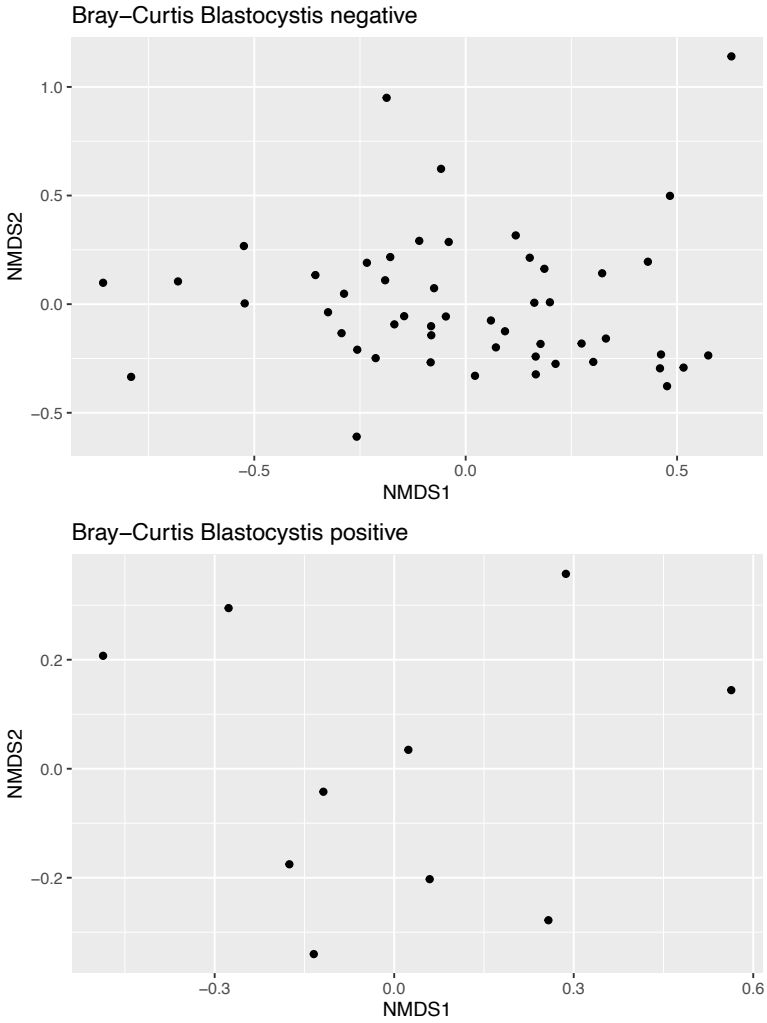


Figure 30. PCoA plot of Bray-Curtis dissimilarity index *Blastocystis* positive and negative samples. sequenced by MinION. The samples are scattered as black points. The y-axis is NMDS2; the x-axis is NMDS1 (Non-metric multidimensional scaling). The top figure is *Blastocystis* negative, and the lower is *Blastocystis* positive.

G. DNA quality for all samples after DNA extraction and quantitation

Table 15. DNA quality is measured after DNA extraction.

Sample ID	ng/µL	260/280	260/230	Sample ID	ng/µL	260/280	260/230
216 V60	38,7	1,8	1,36	246 V60	13,6	1,72	1,02
216 V60	34,7	1,86	1,86	246 V60	35,8	1,82	1,43
221 V60	106,8	1,85	2,05	249 V60	119,1	1,84	1,85
231 V60	32,1	1,78	3,04	251 V60	170	1,83	1,7
233 V60	51,2	1,83	1,94	253 V60	100,7	1,84	1,98
237 V60	18,2	1,71	1,95	257 V60	42,5	1,88	1,78
237 V60	49,2	1,85	1,8	257 V60	42,7	1,85	2,02
238 V60	165,6	1,83	1,56	260 V60	162,2	1,87	2,01
238 V60	183,5	1,86	1,81	261 V60	90	1,85	1,79
239 V60	60,4	1,83	1,89	264 V60	94,9	1,77	1,22
240 V60	84,5	1,81	2,67	264 V60	90,6	1,81	1,58
242 V60	75,1	1,82	2,04	266 V60	118,4	1,86	1,94
245 V60	160,9	1,81	1,47	267 V60	82,2	1,82	3,09
245 V60	178,9	1,85	1,71	276 V60	30	1,86	1,83
246 V60	14,5	1,68	0,79	284 V60	157,2	1,85	2,03
Sample ID	**ng/µL**	**260/280**	**260/230**	**Sample ID**	**ng/µL**	**260/280**	**260/230**
285 V60	71,4	1,84	1,55	440 V3	8,6	1,92	1,05
286 V60	120,3	1,87	1,98	440 V3	12,4	1,36	0,31
430 V11	5,6	1,48	0,79	440 V11	61,6	1,73	1,14
430 V11	8,1	1,66	1,85	441 V3	34,6	1,76	4,49
430 V11	14,1	1,76	0,99	441 V11	108,8	1,87	1,83
430 V11	10,5	1,63	1,14	442 V3	458,3	1,87	2,45
431 V11	17,5	1,67	1,19	442 V3	376,4	1,87	2,4
431 V11	19,9	1,94	2,24	443 V3	38,2	1,84	1,49
433 V11	21,2	2,03	1,27	443 V3	29,9	1,84	1,55
433 V11	21,8	1,91	1,21	443 V3	33,2	1,84	1,76
440 V3	7,4	1,23	0,34	444 V3	71	1,88	1,74
440 V3	7,3	1,36	0,38	445 V0	326,3	1,87	2,31
440 V3	3,4	1,1	-0,63	445 V0	295,5	1,88	2,39
440 V3	1,9	1,54	1,14	445 V3	362,2	1,87	2,38
440 V3	15	1,6	0,49	445 V3	252,5	1,87	2,32
Sample ID	**ng/µL**	**260/280**	**260/230**	**Sample ID**	**ng/µL**	**260/280**	**260/230**
446 V0	39,6	1,83	1,79	456 V0	83	1,84	2,22
446 V3	25,1	1,84	1,4	458 V0	83,9	1,84	1,88
446 V3	26,6	1,85	1,62	458 V0	122,3	1,87	2,12
448 V0	47,7	1,91	1,77	458 V3	201,6	1,83	1,56
448 V3	27,9	1,84	2,46	458 V3	232,3	1,88	2,02
449 V0	229,7	1,83	1,84	459 V0	94,7	1,85	1,66
449 V3	144,5	1,86	2,1	460 V0	234	1,88	2,24

450 V0	158,1	1,85	2,38	460 V3	132,7	1,87	1,64
451 V0	220	1,86	2,41	460 V3	137,3	1,9	2,22
451 V3	7,1	2,04	1,1	461 V0	159,6	1,88	2,23
451 V3	4,4	1,79	1,45	461 V3	265,2	1,86	2,18
451 V3	4,1	1,64	1,53	462 V0	148,9	1,84	2,08
451 V3	6,5	1,98	3,3	462 V3	40,8	1,9	2,39
452 V0	235,7	1,86	2,09	463 V0	1123,1	1,88	2,39
452 V0	126,2	1,85	1,94	464 V3	75,4	1,82	1,45
452 V3	145,9	1,85	1,96	464 V3	62,7	1,88	2,2
453 V0	158,1	1,85	2,38	465 V0	233,7	1,88	2,12
453 V3	46,5	1,72	1,18	466 V0	37,4	1,81	0,92
453 V3	42,2	1,84	1,59	466 V0	34,7	1,86	1,51
454 V3	274,2	1,88	2,24	467 V0	283,8	1,88	2,32
455 V0	141,8	1,86	2,12	468 V0	401,9	1,88	2,39
455 V3	195	1,84	1,95	474 V0	203,2	1,88	2,26