University of South-Eastern Norway

Faculty of Technology, Natural Sciences and Maritime Sciences

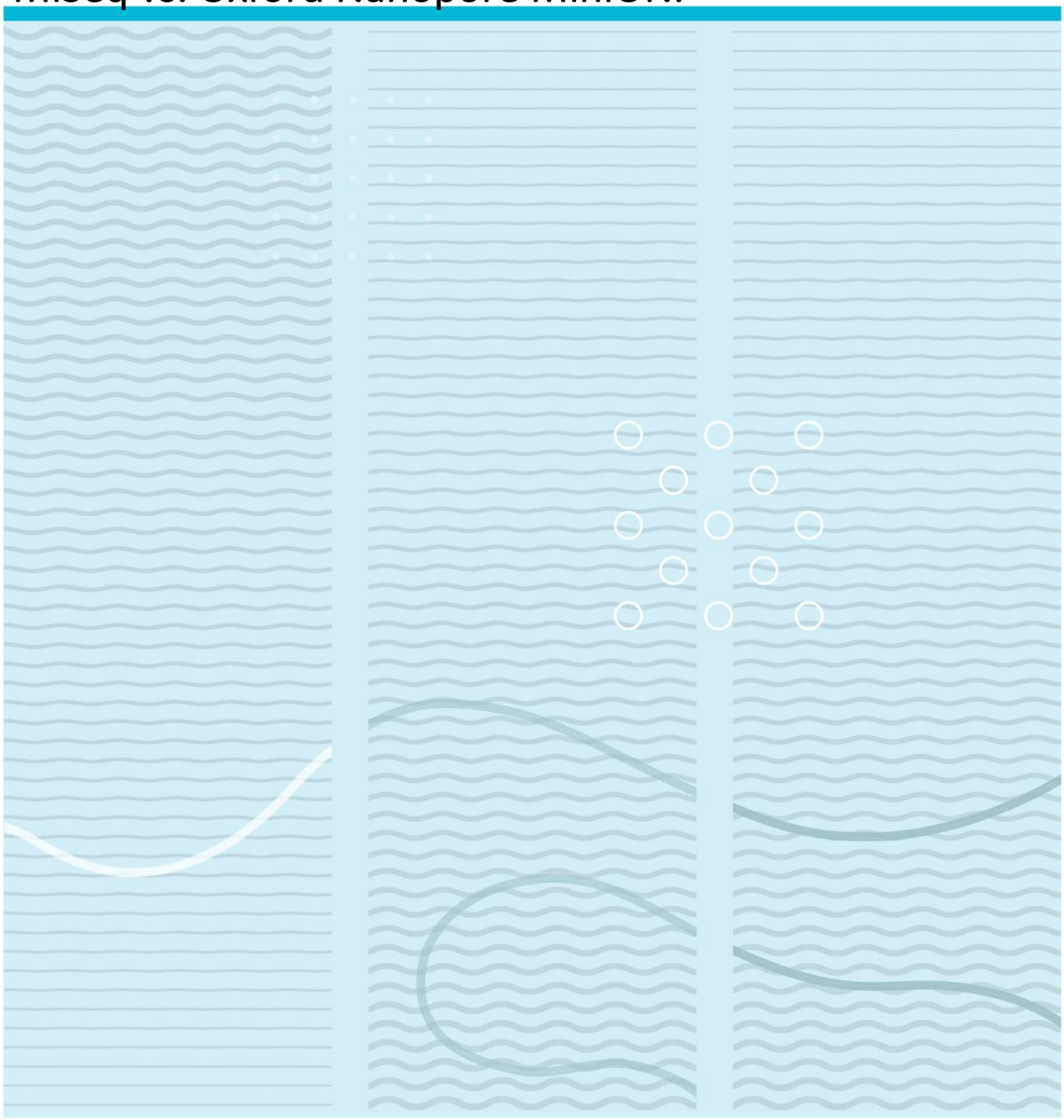
Master's Thesis

Study programme: Master of Environmental Science

Spring 2021

Jorunn Hellekås

Next-generation sequencing of soil samples. Illumina MiSeq vs. Oxford Nanopore MinION.



University of South-Eastern Norway
Faculty of Faculty of Technology, Natural Sciences and Maritime Sciences
Department of Natural Sciences and Environmental Health
PO Box 235
NO-3603 Kongsberg, Norway

http://www.usn.no

This thesis is worth 60 study points.

Summary

Most of the world's soil resources are of poor condition and with an expanding population it is important to improve soil management. Activities of microorganisms living in soil are essential for soil and plant health, and study of soil microorganisms can give important information of microbial diversity and function, which can give status of soil condition.

Next-generation sequencing give us the opportunity to study soil microorganisms at the most basic level. Illumina MiSeq is the most used sequencing platform. It provides a lot of data output but is limited by short read length. Oxford Nanopore MinION is the newest third generation sequencing technology which offers long-read sequences, but a main disadvantage is a relatively low read accuracy.

We used Illumina MiSeq and Oxford MinION to compare the taxonomic resolution of bacteria by sequencing of soil samples. Additionally, the soil samples were subject to different treatments: pesticide, fertilizer, pesticide*fertilizer and untreated, to see if it influenced the microbial community.

The two technologies got very similar results at both alpha- and beta diversity analyses. Application of fertilizer gave a significant effect on the microbial community at both technologies. Different phyla were dominating within the two sequencing platforms. Proteobacteria dominated with the MinION analysis, while Actinobacteria dominated with the Illumina MiSeq analysis. However, we conclude that there was not a difference in the taxonomic resolution between Illumina MiSeq and Oxford MinION.

Contents

Foreword

I am sincerely glad Mona introduced me to the topic soil health when I started my master's degree. It led me to an interestingly year writing my master's thesis.

I would like to thank my supervisor Mona for creating the subject of this thesis, and for advice and guidance through the writing process. I would also like to thank my supervisor Jørn Henrik for help and guidance with both writing and statistical analyses. My sincere thanks to Sofie for all her work at the lab, and for helping me when I asked for something. Huge thanks to Bionér and Standard Bio for letting me take part in their soil project.

Bø i Telemark, 16.05.2021

Jorunn Hellekås

1. Introduction

The soil contains a great diversity of organisms and is essential for all life on earth. Soil provides important ecosystem services like water regulation, erosion control, and climate regulation, and is important to human health. Large quantities of food are derived from the soil because it provides the right condition for plant growth (Brevik et al., 2018). The complex soil organic matter is a storage of nutrients, where physical-chemical processes control nutrient sorption and availability for the plant roots. This soil function, to support and sustain plant growth, is accomplished by a cooperation between the plant itself, the soil mineral matrix and microbes (Brevik et al., 2020). A healthy soil produces healthy crops that in turn feed people and animals. The Food and Agriculture Organisation for the United Nations (FAO) has raised a concern regarding soil health. Their report from 2015 says that one-third of earth's land is acutely degraded, and with an expanding population it is important to improve the soil health and create a sustainable soil management (FAO & ITPS, 2015).

Microbes in soil are central to soil and plant health. They act as decomposers of dead plants and animals, creating organic matter. The degradation of dead material release nutrients that enhance soil fertility and finally improves plant productivity (Mohanram & Kumar, 2019). In the rhizosphere, soil bacteria surround plant roots and interact with them. Through processes like nitrogen fixation and solubilization of phosphorus compounds, vital molecules are converted to a form accessible to the plant. Pathogens are prevented to infect plant roots as soil bacteria fight for nutrients and space, and produce antibiotics (Khatoon et al., 2020).

The soil microbial community can be influenced by soil variables, e.g., pH, carbon-to-nitrogen ratio, and available phosphorus, which will vary with the use of fertilizers and pesticides. Study of microorganisms in soil can give important information of composition and function of microbial populations, which can give status of soil condition (Hermans et al., 2017). So far only a few soil microorganisms have been studied in the lab due to challenges in replicating their natural environment (Solden et al., 2016). However, culture-independent techniques, using molecular analyses have been developed. Next-generation sequencing gives the opportunity of differentiating terrestrial life forms at the

most fundamental level. This technology offers massively parallel sequencing of DNA/RNA, which can reveal the hereditary and biochemical properties of microorganisms (Heather & Chain, 2016). To identify microbial communities in different environments, metabarcoding is widely used, where barcode means a short variable gene region. The process of metabarcoding includes DNA extraction from a sample and amplification by polymerase chain reaction (PCR), using specific primers which amplify the target gene region. The amplified products are sequenced and then characterized using bioinformatic tools (Orgiazzi et al., 2015).

The 16S ribosomal RNA gene of prokaryotes is largely used as the target gene when studying microbial taxonomy. This gene is part of the small ribosomal subunit present in all prokaryotic cells and have useful characteristics for bacteria identification, like its small size (~1500 bp) and its nine variable regions (V1-V9) flanked by conserved stretches (Figure 1) (Santos et al., 2020).

Figure 1: The 16S rRNA gene (1500 bp) consist of 9 variable regions (V1-V9) and conserved regions (blue) (figure by Jorunn Hellekås).

The variable regions can be used to distinguish between different bacterial species because of different evolutionary development. The conserved regions make it possible to use universal designed primers that will be able to bind to a wide variety of DNA templates for identification of many microorganisms (Santos et al., 2020). When sequencing the 16S rRNA gene, similar sequence variants are often clustered into operational taxonomic units (OTUs), where each cluster represent a taxonomic unit of a bacteria species or genus. The OTUs are generated to chosen threshold lines (e.g. 97% or 99%) for comparison with reference databases (Johnson et al., 2019). Different sequencing methods exists, and the concept of sequencing by synthesis is commonly used. Here, the template strand is used to synthesise a new DNA strand, and the nucleotides are monitored when they are incorporated, creating the new strand. This sequencing method is offered by companies like Thermo Fisher (e.g., Ion Torrent) and Illumina. A newer sequencing method is single-molecule sequencing which are used by

Oxford Nanopore Technologies. This method monitors single nucleotides as the template strand passes an ionic current through a nanopore (McCombie et al., 2019).

Illumina is a widely used second generation sequencing technology, and the company offers different sequencing platforms like the MiniSeq, MiSeq and NextSeq, where the sequencing length varies between the platforms. For analyses with Illumina, the DNA samples are amplified by PCR, and the PCR-products are cleansed. Then, library preparation starts with fragmentation of amplified DNA, followed by ligation of adaptor sequences to each end of the DNA fragments. The prepared DNA and necessary reagents are then loaded into the sequencer and further to a solid-surface flow cell coated with primers complementary to the adaptor sequences. The fragmented DNA binds to the primers at the cell surface and is amplified by a DNA polymerase, producing clusters of millions of copies of the initial DNA fragment (Figure 2A). The fragments are then sequenced using sequencing-by-synthesis (SBS) chemistry. Fluorescently labelled nucleotides (A, C, G and T) are monitored as they are incorporated one base at a time, creating a new DNA strand (Figure 2B and 2C) (Quainoo et al., 2017).

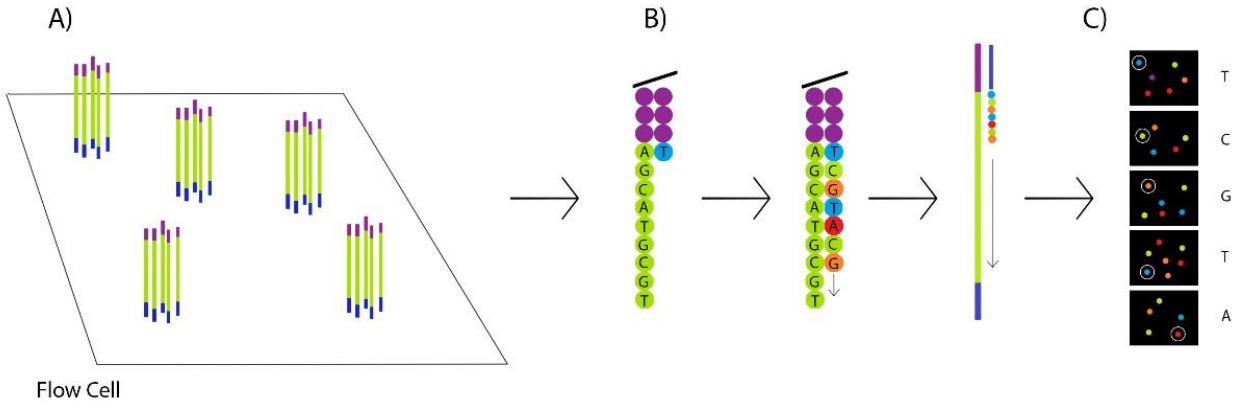


Figure 2: Illumina sequencing. A: The flow cell has clusters of millions of copies of the initial DNA. B: Nucleotides are incorporated base per base. C: The incorporated nucleotides are monitored by a computer (figure by Jorunn Hellekås).

Illumina MiSeq sequencing (hereinafter referred to as MiSeq) produces short reads, approximately 300 bp. When targeting the 16S rRNA gene, only sub-regions of the gene are sequenced, such as V4 or V6, or two variable regions, such as V3-V4. Taxonomy is assigned based on these short variable regions (Johnson et al., 2019). A lot of data output

(20 GB) is produced as the system can run up to 96 samples at the same time, with a read accuracy of 99.9%. Also, a benefit of MiSeq sequencing is that many bioinformatic tools and programs are designed to process sequencing data from variable region V3 and V4 (Santos et al., 2020).

Oxford Nanopore Technologies (ONT) made the small, portable MinION sequencer platform commercially available in 2014 (hereinafter referred to as MinION) (Rang et al., 2018). The sequencing principle involves nanopore technology and differ from Illumina technologies as a single molecule of DNA is detected without chemical labelling the sample. The MinION device consist of a flowcell with sensors and nanopores, where the flowcell is inserted into the device before each sequencing run. Barcodes and adaptors are ligated to the DNA samples during library preparation. The adaptors will interact with proteins attached to the nanopores so the DNA strand can enter the pores (Figure 3A). The nanopores has an ionic current which are disrupted when a nucleotide traverses it and is displayed as a "squiggle" plot (Figure 3B). Each base type is linked to a characteristic alteration in the current (squiggles), allowing sequencing of the DNA strand as it moves through the nanopore (Plesivkova et al., 2019).

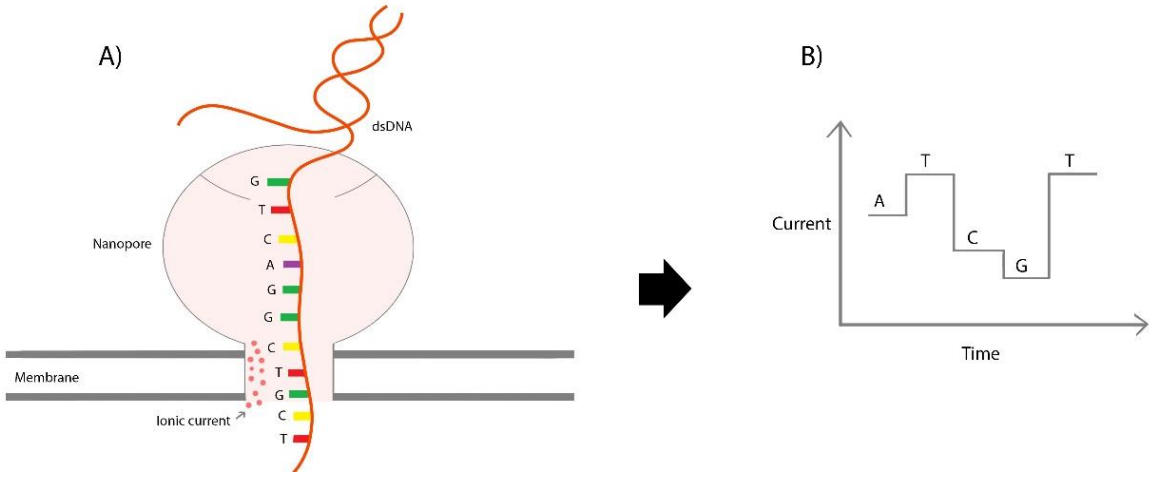


Figure 3: MinION sequencing. A: The DNA strand is moving through the nanopore disturbing the ionic current in a base-specific manner. B: The passing nucleotides are monitored as a squiggle plot (figure by Jorunn Hellekås).

The MinION can sequence thousands of basepairs in one run and hence the whole 16S rRNA gene can be sequenced. The detection of nucleotides passing the nanopores gives results in real time. This gives the opportunity to start data analysis early in the process, and also stop the sequencing when sufficient results are obtained (Santos et al., 2020). A

known problem of sequencing with MinION is high error rates of up to 15% per base (Quainoo et al., 2017; Rang et al., 2018). However, ONT constantly evolves their technology, and reports that new chemistry of the flow cell is coming soon with a read accuracy of up to 99% (Oxford Nanopore Technologies, 2020).

Identification of bacteria at species level can be crucial in several situations. For instance, outbreaks at hospitals require quick information to infer transmission dynamics and identify species with potential antibiotic resistance genes (McGann et al., 2016). Foodborne pathogens can cause disease, and it is therefore important to find the source of transmission in order to prevent further damage (Quick et al., 2015). Moreover, soil borne pathogens may threaten plant health and hence crop production (De Corato, 2020). However, it can be challenging to receive sequencing results and classify at species level. Johnson et al. (2019) performed an in-silico experiment to test if different sub-regions of the 16S rRNA gene could provide accurate, taxonomic classification at the species-level. They achieved poor results and concluded that partial sequencing of the 16S rRNA gene cannot discriminate between species. Also, some studies claims that the 16S rRNA gene alone is not sufficient for species identification, since the sequence similarity varies substantially between genera and species (Cusco et al., 2018; Srinivasan et al., 2015).

The aim of this study was to compare the diversity estimates and taxonomic resolution determined by sequencing of soil samples with Illumina MiSeq and ONT MinION. Additionally, the soil was subject to different treatments, to see if it influenced the microbial community. The whole 16S rRNA gene was sequenced with the MinION, while the variable regions V3-V4 were sequenced with MiSeq.

2. Materials and methods

2.1. Soil sampling and treatments

Mixed soil from location 59.3868012, 9.2255803 in Midt-Telemark was distributed into 12 pallets (400L x 12) in mid-June 2020. Horse pasture mix (SPIRE) was sowed in all 12 pallets. Each pallet got 5 grams of seed sown evenly according to the manufacturer's description. The soil was watered every other day until grass sprouts. Then the pallets were subject to various treatments: three pallets were treated with pesticide (Roundup Glyphosate mixed with water (1:100)), three pallets were treated with mineral fertilizer (60-gram Yara "25-2-6"), three pallets were treated with both pesticide and mineral fertilizer, and three pallets remained untreated. The treatments of fertilizer and pesticide were done according to the manufacturer's description.

In the autumn of 2020, soil for DNA analysis (10-25 gram) was collected in plastic tubes from the middle of the 12 pallets in two rounds. We left out one sample because further analysing required a maximum of 23 samples plus negative control. Their sample ID are listed in Table 1. The samples were stored in the freezer (-80 C^{o}) until DNA extraction.

Table 1: Table of the 23 collected soil samples with sample ID and type of treatment.

Sample ID	**Treatment**	**Sample ID**	**Treatment**
MUTP1_0909	Untreated	MUTP1_0211	Untreated
MUTP2_0909	Untreated	MUTP2_0211	Untreated
MUTP3_0909	Untreated	MUTP3_0211	Untreated
MFP1_0909	Fertilizer	MFP1_0211	Fertilizer
MFP2_0909	Fertilizer	MFP2_0211	Fertilizer
MFP3_0909	Fertilizer	MFP3_0211	Fertilizer
MPP1_0909	Pesticide	MPP1_0211	Pesticide
MPP2_0909	Pesticide	MPP2_0211	Pesticide
MPP3_0909	Pesticide	MPP3_1510	Pesticide
MPFP1_0909	Pesticide + Fertilizer	MPFP1_0211	Pesticide + Fertilizer
MPFP2_0909	Pesticide + Fertilizer	MPFP2_0211	Pesticide + Fertilizer
MPFP3_0909	Pesticide + Fertilizer		

2.2. Microbial DNA extraction

DNA was extracted from 0.25 g of soil using the DNeasy PowerMax Soil kit (QIAGEN, Germany) following manufacturer's instructions. DNA yield (ng/µl) and purity ($A_{260/280}$) were quantified spectrophotometrically with a NanoDrop Lite Spectrophotometer (Thermo Fisher Scientific, USA). DNA yield was also quantified fluorometrically with a Qubit 3.0 fluorometer (Invitrogen, USA). Both NanoDrop and Qubit were used according to manufacturer's protocols. Extracted DNA was stored in the freezer (-20 C^o) until further analysing.

2.3. Illumina MiSeq sequencing

The extracted DNA samples were placed in a 96-well plate, where each sample had a minimum amount of 200 ng DNA, and sent to the Norwegian Sequencing Centre (NCS) (Oslo, Norway). Here, the DNA samples were subject for 16S DNA library preparation and sequenced on an Illumina MiSeq instrument (Illumina, USA) with a 300-bp paired end reads (MiSeq v3 chemistry) with 10% PhiX spike-in. NSC followed the protocol of Fadrosh et al. (2014) to prepare amplicon libraries of the V3-V4 region (approx. 469 bp) of the prokaryotic 16S rRNA gene. A positive control sample (ZymoBIOMICS Microbial Community DNA standard II, Zymo Research) and a negative control sample (H_2O) was included. The 16S amplification primer sequences were as follows:

- 319F forward primer: 5' ACTCCTACGGGAGGCAGCAG 3'
- 806R reverse primer: 5' GGACTACNVGGGTWTCTAAT 3'

The sequencing results were received as FASTQ files.

2.4. Library preparation and ONT MinION sequencing

For library preparation, we followed the Oxford Nanopore Technologies protocol "16S Barcoding Kit 1-24 SQK-16S024" (16S_g086_v1_revJ_14Aug2019) according to the manufacturer's description. The extracted DNA samples were first amplified using long-range PCR (Thermal cycler model PTC-200, MJ research, USA) in 50 µl volume containing 10 ng of DNA and 25 µl LongAmp Hot Start Taq 2x Master Mix (Biolabs, New England) and 10 µl of premade barcoded primers (16S Barcoding Kit 1-24 (SQK-16S024), Oxford

Nanopore) which amplified the 16S rRNA gene spanning the variable regions V1-V9. The PCR conditions were as follows 1 minute of initial denaturation at 95C°, 30 cycles of 20 seconds denaturation at 95C°, 30 seconds of annealing at 55 C° and elongation for 2 minutes at 65 C°, followed by a final extension of 5 minutes at 65 C°. We used 30 amplification cycles instead of 25 cycles, which the protocol recommends, due to low output of PCR products. Each of the 23 amplified samples were diluted to a DNA concentration of around 10 ng/µl. The samples were then pooled using 2 µl from each. Subsequently, the pooled samples and necessary reagents were loaded on the MinION flow cell FLO-MIN106D (R9) (Oxford Nanopore Technologies, UK) and the sequencing was started.

2.5. Bioinformatic and statistical analyses

Illumina MiSeq 16S sequencing data processing

Raw sequencing reads were demultiplexed and primers and barcodes were trimmed off using an inhouse software at the Norwegian Sequencing Centre (https://github.com/nsc-norway/triple_index-demultiplexing/tree/master/src). Resulting FASTQ files were further pre-processed, quality filtered and analysed using the QIIME2 pipeline. Reads were truncated to 280 bases and paired sequence reads were joined, quality-filtered and chimeras were removed using DADA2 in the QIIME2 pipeline. Taxonomic classification of ASVs generated from DADA2 were done using the Naive Bayes classifier algorithm trained on data from SILVA v. 138 using the feature classifier in QIIME2.

ONT MinION 16S sequencing data processing

Raw FAST5 reads were basecalled, sorted by barcodes and converted to FASTQ files using the MinKNOW software (v. 20.06.4). Reads with a phred score <7 and a length lower than 1200 and higher than 2000 bp were excluded from further analyses using the software EPI2ME. EPI2ME was further used to assign the nanopore reads to taxonomy using the 16S module. Here, filtered reads were blasted against the NCBI 16S bacterial database and reads were assigned to taxonomic units based on sequence similarity.

Statistical analyses

We used RStudio (RStudio, 2021) (v. 1.4.1103) for all statistical analyses. The samples were rarefied in order to account for differences in number of sequences obtained from the different samples.

R package Phyloseq (https://github.com/joey711/phyloseq) was used to compare different estimates of the alpha diversity, i.e., the diversity within samples, and to plot the relative abundance of the different groups. We calculated the observed species richness, the abundance-based coverage estimator (ACE) and the Shannon index. Observed species richness reflect the number of species in the samples while ACE reflect the species relative abundance (number of individuals in each species). The Shannon index accounts for both species richness and their relative abundance.

R package Vegan (https://github.com/vegandevs/vegan/) was used to calculate the beta diversity, i.e., the diversity between samples, in a Principal Coordinates Analysis (PCoA) plot of Bray-Curtis dissimilarities. The significance and the effect of treatments on the bacterial community (beta diversity) were calculated using PERMANOVA using the vegan R-package.

3. Results

The results of MiSeq sequencing ranged from about 20 000 reads (sample MFP2_0909) to just above 100 000 reads (sample MPFP2_0909) (Figure 4, left). MinION sequencing ranged from just below 50 000 reads (sample MUTP2_0909) to above 150 000 reads (sample MUTP1_0211), where over half of the samples had over 100 000 sequencing reads (Figure 4, right). Sample MUTP2_0211 got poorly sequencing results in the MinION sequencing and was removed for further analyses, giving us a total of 22 samples.

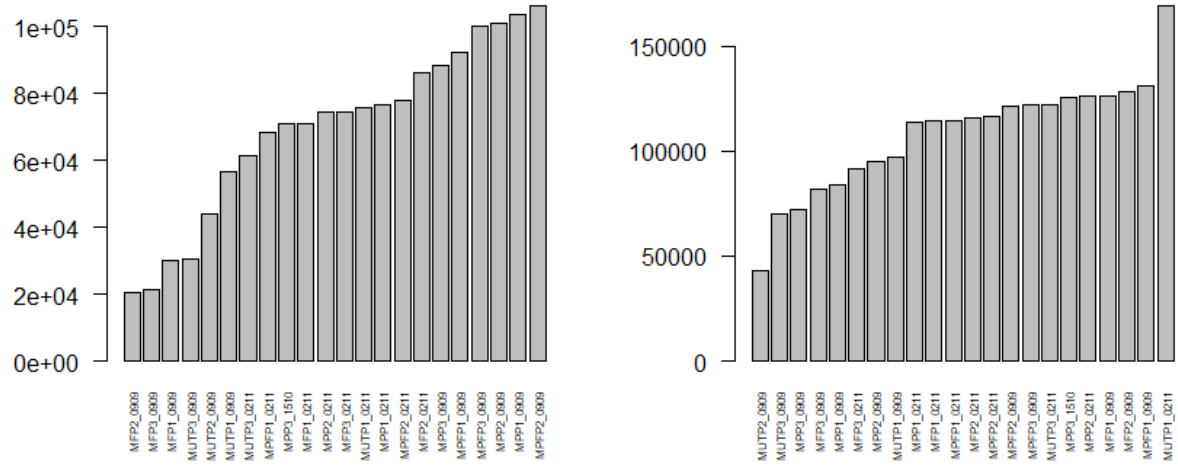


Figure 4: Number of sequencing reads generated by MiSeq (left) and MinION (right). The 22 samples are represented on the horizontal line and number of reads on the vertical line.

The variation in number of reads produced from each sample may affect further calculations of diversity estimates. To avoid this we rarefied the data, we also removed reads that did not assign to any domain such as bacteria and archaea and removed the taxa that was only found in very low frequency.

3.1. Alpha diversity

We compared different estimates of the alpha diversity: observed species richness, the Abundance-based Coverage Estimator (ACE) and the Shannon index, of the two technologies.

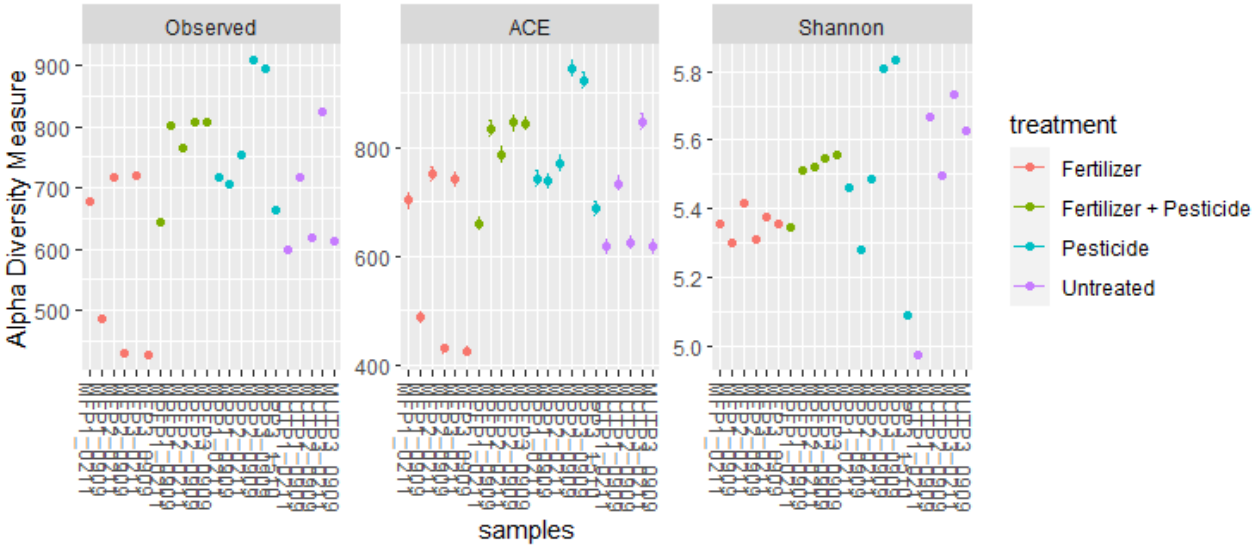


*Figure 5: Alpha diversity analyses of the samples sequenced on MiSeq. The calculated analyses are Observed species richness, ACE, and Shannon index. The data points represent the 22 samples, and the different colours represent the four treatments: red = fertilizer, green = fertilizer*pesticide, blue = pesticide, and purple = untreated.*

MiSeq sequencing results showed that three of the samples treated with fertilizer had the lowest number of observed species (under 500) (Figure 5). The highest number of observed species (around 900) belonged to two samples treated with pesticide. The calculation of ACE was quite similar to the result of observed species and ranged from just above 400 to around 900. The Shannon index showed that the samples treated with both fertilizer and fertilizer*pesticide was more grouped than the rest of the samples which were spread out.

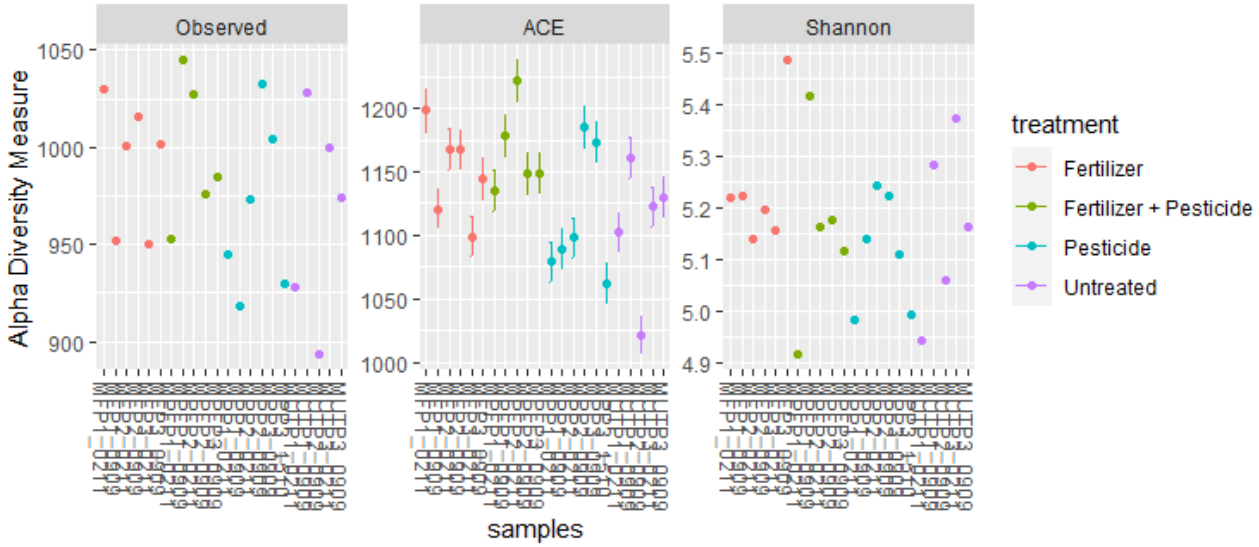


*Figure 6: Alpha diversity analyses of the samples sequenced on MinION. The calculated analyses are Observed species richness, ACE, and Shannon index. The data points represent the 22 samples, and the different colours represent the four treatments: red = fertilizer, green = fertilizer*pesticide, blue = pesticide, and purple = untreated.*

Analyse results of observed species when sequencing on the MinION, showed that one sample was below 900 and the rest of the samples were between 900 and 1050 (Figure 6). The abundance in ACE ranged from just above 1000 to around 1250. The Shannon index showed a grouping of the samples treated with fertilizer, except one outlier. The rest of the samples were spread out.

3.2. Correlation test

We ran a correlation test to check if the alpha diversity analyses from MinION and MiSeq were correlated. The calculations showed us that the correlation was not significant ($p > 0.05$) (Table 2).

Table 2: Calculated correlation of Observed species, ACE and Shannon index of the samples sequenced on MinION and MiSeq. The table show the t-value, p-value, 95% confidence interval and the correlation coefficient (R).

	T	P	95% CI	R
Observed:	1.3035	0.2072	-0.1607 0.6274	0.2798
ACE:	1.1452	0.2657	-0.1938 0.6063	0.2481
Shannon:	1.8331	0.0817	-0.0504 0.6905	0.3793

We made a correlation plot which visually suggested that some positive correlation was present (Figure 7).

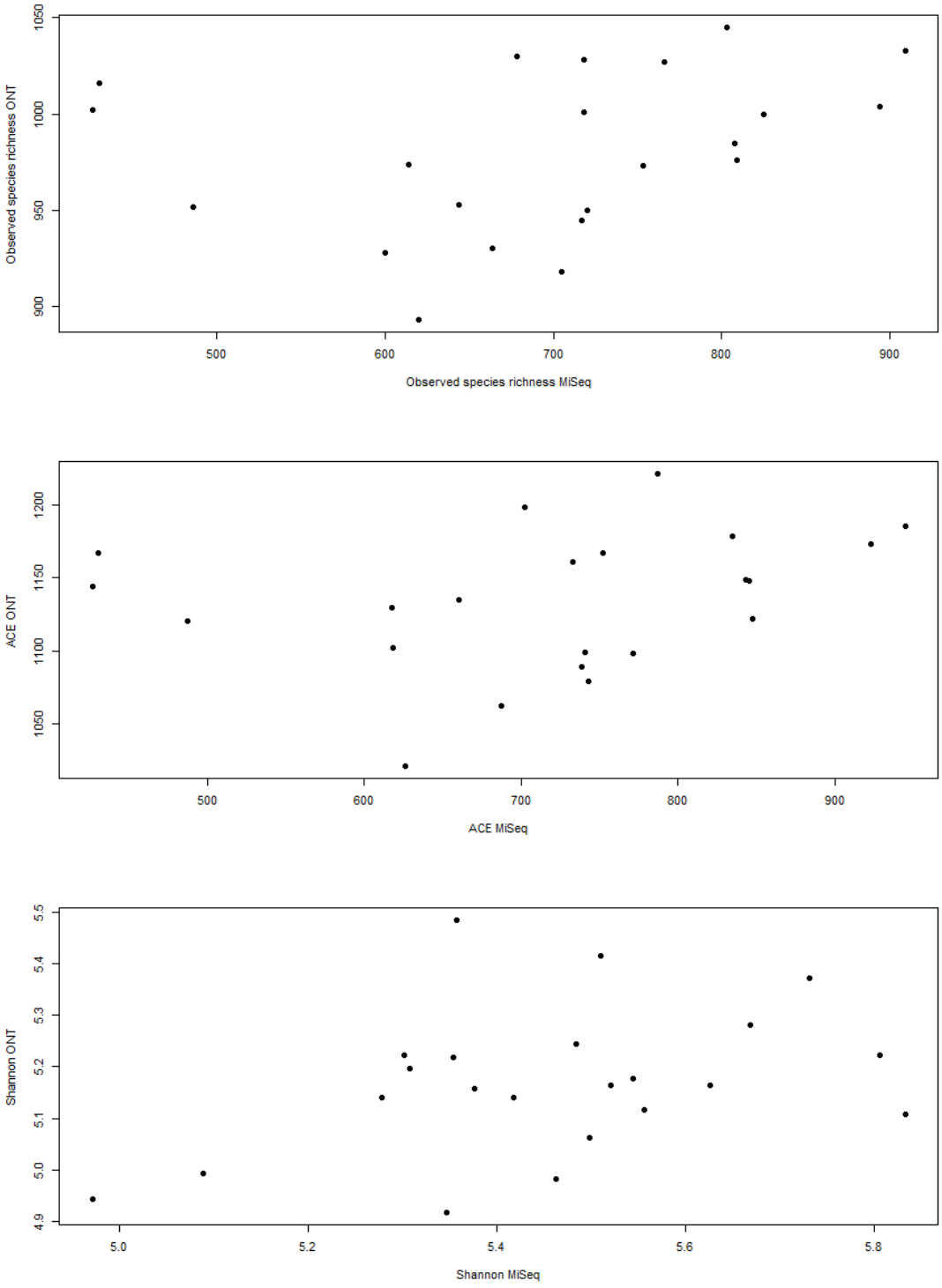


Figure 7: Correlation plot of observed (top), ACE (middle), and Shannon index (bottom) of the samples sequenced on MinION and MiSeq. MiSeq is represented on the horizontal line and MinION is represented on the vertical line. The data points represent the 22 samples.

From the correlation plot it was evident that two samples were outliers (two data points up left in the plots of observed species richness and ACE, sample MFP3_0909 and MFP2_0909) which may be the reason there was no significant correlation. We removed these two samples and calculated the correlation again.

Table 3: Calculated correlation of Observed species, ACE and Shannon index of the samples sequenced on MinION and MiSeq after removing two outliers. The table show the t-value, p-value, 95% confidence interval and the correlation coefficient (R).

	T	P	95% CI	R
Observed:	3.0329	0.0072	0.1872 0.8145	0.5816
ACE:	2.2136	0.0400	0.0252 0.7513	0.4626
Shannon:	2.5797	0.0189	0.1000 0.7822	0.5195

The correlation was now significant with all p-values < 0.05 and higher correlation coefficients compared to previous calculations (Table 3).

3.3. Beta diversity

We further calculated the beta diversity to see how divergent the samples and the two technologies were from each other.

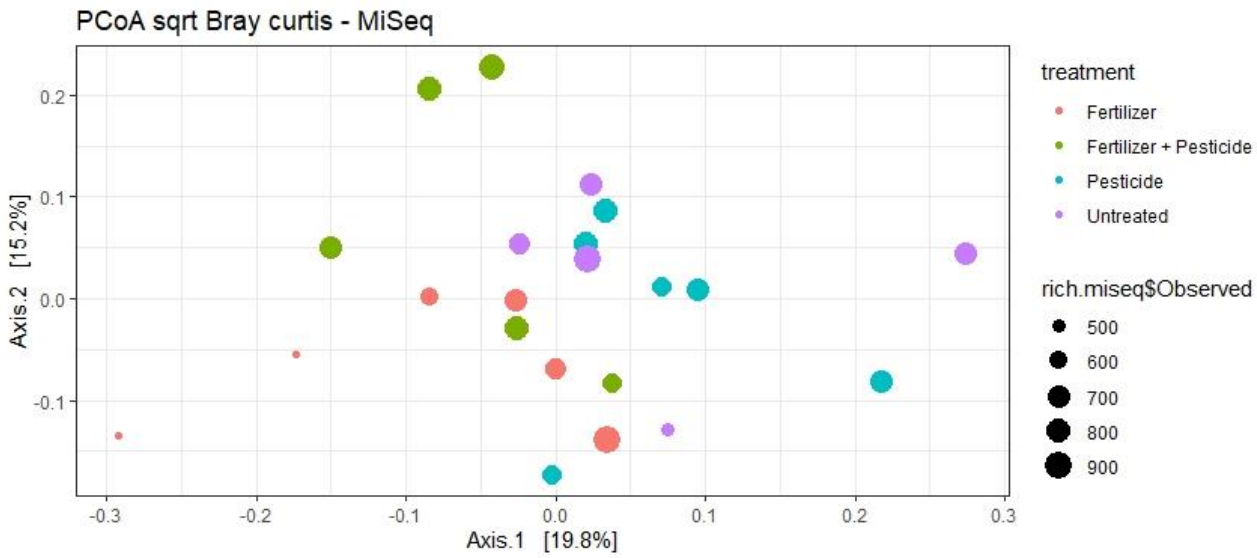


*Figure 8: PCoA plot of Bray-Curtis dissimilarity between microbial communities of the samples sequenced on MiSeq. The data points represent the 22 samples, and the different colours represent the four treatments: red = fertilizer, green = fertilizer*pesticide, blue = pesticide, and purple = untreated.*

We expected some tendency of clustering regarding soil samples with similar treatment. This was not seen visually in the PCoA plot of the MiSeq analysis. Most of the samples were centred and the different treatments overlapped each other (Figure 8).

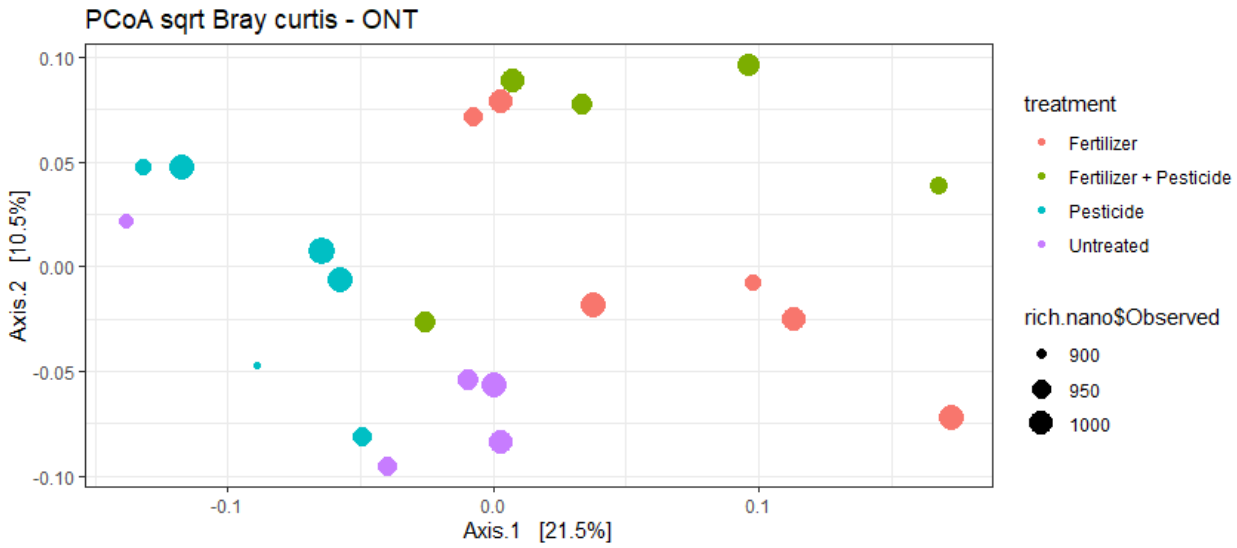


Figure 9: PCoA plot of Bray-Curtis dissimilarity between microbial communities of the samples sequenced on MinION. The data points represent the 22 samples, and the different colours represent the four treatments: red = fertilizer, green = fertilizer + pesticide, blue = pesticide, and purple = untreated.

The samples analysed with MinION distinguished better the different treatments (Figure 9). The samples within each treatment showed a tendency of grouping and there was not

much overlapping between the treatments. We did a PERMANOVA test to check if the different treatments of the soil gave a significant effect on the bacterial community of the samples sequenced on MiSeq and MinION.

Table 4: Summary of PERMANOVA test of soil treatments on the microbial community of the samples analysed with MiSeq. The table show Df: degrees of freedom, SS: sum of squares, MS: mean of squares, F-statistics, R^2 and P-value.

Treatment	Df	SS	MS	F	R^2	P-value
Fertilizer	1	0.183	0.183	2.770	0.119	0.001
Pesticide	1	0.083	0.083	1.263	0.054	0.17
Fertilizer*Pesticide	1	0.084	0.084	1.266	0.054	0.155
Residuals	18	1.189	0.066	0.773		
Total	21	1.538				

The PERMANOVA test of the samples analysed with MiSeq showed that treatment with fertilizer gave a significant effect on the microbial community ($p = 0.001$) while treatment with pesticide and the interaction between the two treatments were not significant ($p > 0.05$) (Table 4). This was similar to the samples analysed with MinION. The PERMANOVA test showed that treatment with fertilizer gave a significant effect on the microbial community ($p = 0.001$) while the other treatments did not ($p > 0.05$) (Table 5).

Table 5: Summary of PERMANOVA test of soil treatments on the microbial community of the samples analysed with MinION. The table show Df: degrees of freedom, SS: sum of squares, MS: mean of squares, F-statistics, R^2 and P-value.

Treatment	Df	SS	MS	F	R^2	P-value
Fertilizer	1	0.085	0.085	2.585	0.112	0.001
Pesticide	1	0.046	0.046	1.380	0.060	0.089
Fertilizer*Pesticide	1	0.037	0.037	1.122	0.049	0.27
Residuals	18	0.594	0.033	0.780		
Total	21	0.762				

3.4. Dominating groups of bacteria

We analysed which groups of bacteria that were dominating in the soil samples. The abundance of the different phyla varied between the two technologies (Figure 10).

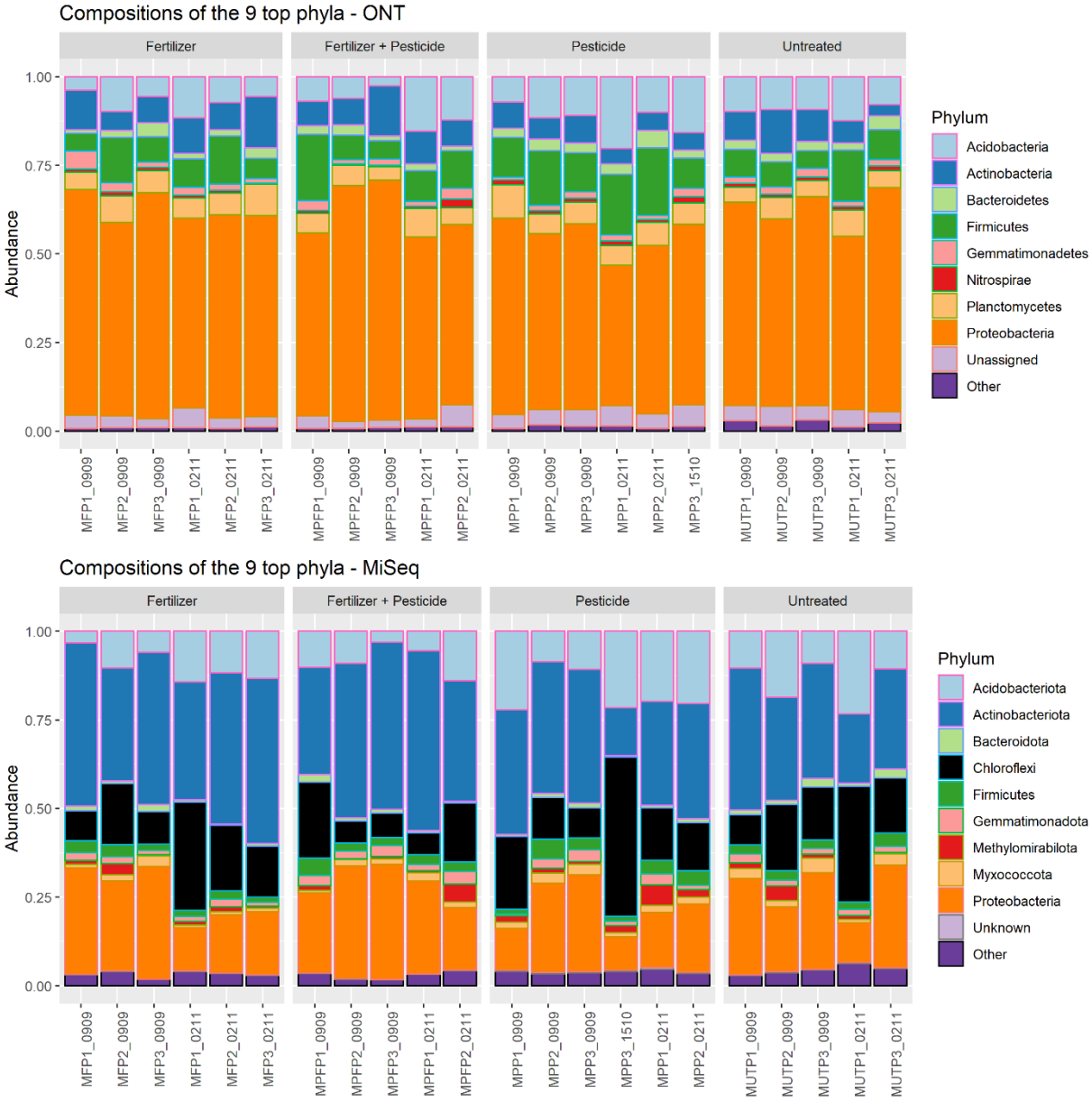


Figure 10: The figures show 9 phyla with highest abundance when analysed with MinION (top) and MiSeq (bottom). The four boxes represent the different treatments, and the columns represent the 22 samples which are listed on the horizontal line. The different colours represent the phyla written to the right of the figures.

The analyses with MinION showed that Proteobacteria had the greatest abundance (Figure 10, top). The three phyla with lowest abundance were Bacteroidetes, Gemmatimonadetes and Nitrospirae. In general, there was no clear pattern regarding soil treatment and phyla. This also applied to the samples analysed with MiSeq (Figure 10,

bottom). Here, Bacteroidota, Methylomirabilota and Myxococcota were the three phyla with lowest abundance. The phylum Chloroflexi made up a relatively large proportion of the phyla detected with MiSeq but was not represented by MinION. Further on, Actinobacteriota had the greatest abundance with MiSeq while the same phylum had quite low abundance with MinION. We investigated Actinobacteria to see how bacteria families within this phylum distributed between the two technologies.

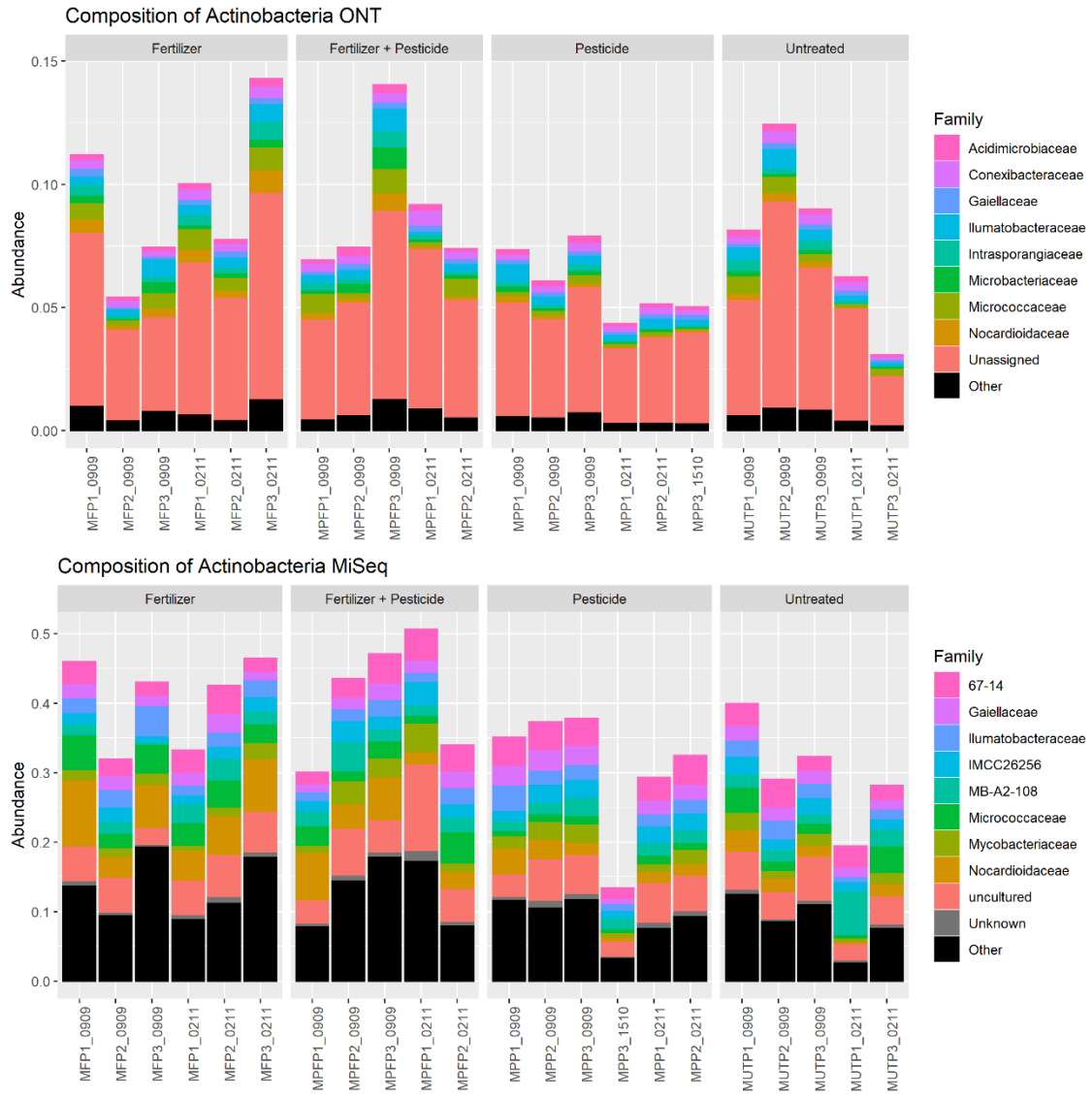


Figure 11: The figures show 9 top families within the phylum Actinobacteria when analysed with MinION (top) and MiSeq (bottom). The four boxes represent the different treatments, and the columns represent the 22 samples which are listed on the horizontal line. The different colours represent the families written to the right of the figures.

The relative abundance of reads assigned to Actinobacteria was higher when analysed with MiSeq than with MinION (Figure 11). The families *Micrococcaceae* and *Nocardioidaceae* were the only ones detected at top 9 families by both MinION and

MiSeq. The black area (other) represents other bacteria families with lower abundance than the 9 top families. MinION had a great proportion of “unassigned” reads (Figure 11, top). This means that the reference database used did not have the information required to classify the reads at the family level.

We further investigated Proteobacteria, which had a great abundance when analysed with MinION.

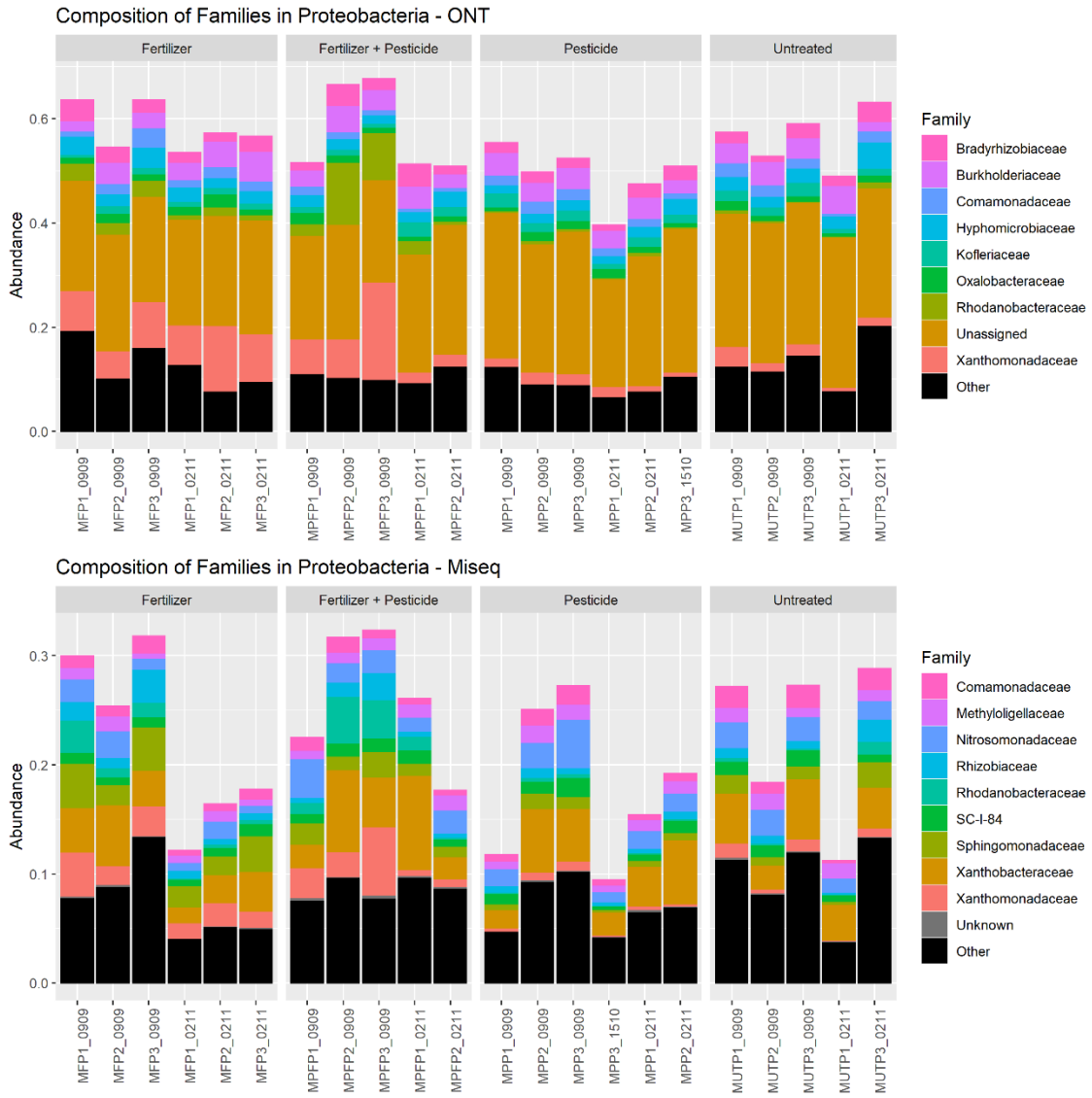


Figure 12: The figures show 9 top families within the phylum Proteobacteria when analysed with MinION (top) and MiSeq (bottom). The four boxes represent the different treatments, and the columns represent the 22 samples which are listed on the horizontal line. The different colours represent the families written to the right of the figures.

The relative abundance of reads assigned to Proteobacteria was higher when analysed with MinION than with MiSeq (Figure 12). However, the relative abundance of unassigned reads was higher when analysed with MinION than with MiSeq. The family

Xanthomonadaceae had a higher abundance in the samples treated with fertilizer and fertilizer*pesticide than the other treatments at both technologies. We decided to look into *Xanthomonadaceae* at species level in order to compare the species determination with the two technologies.

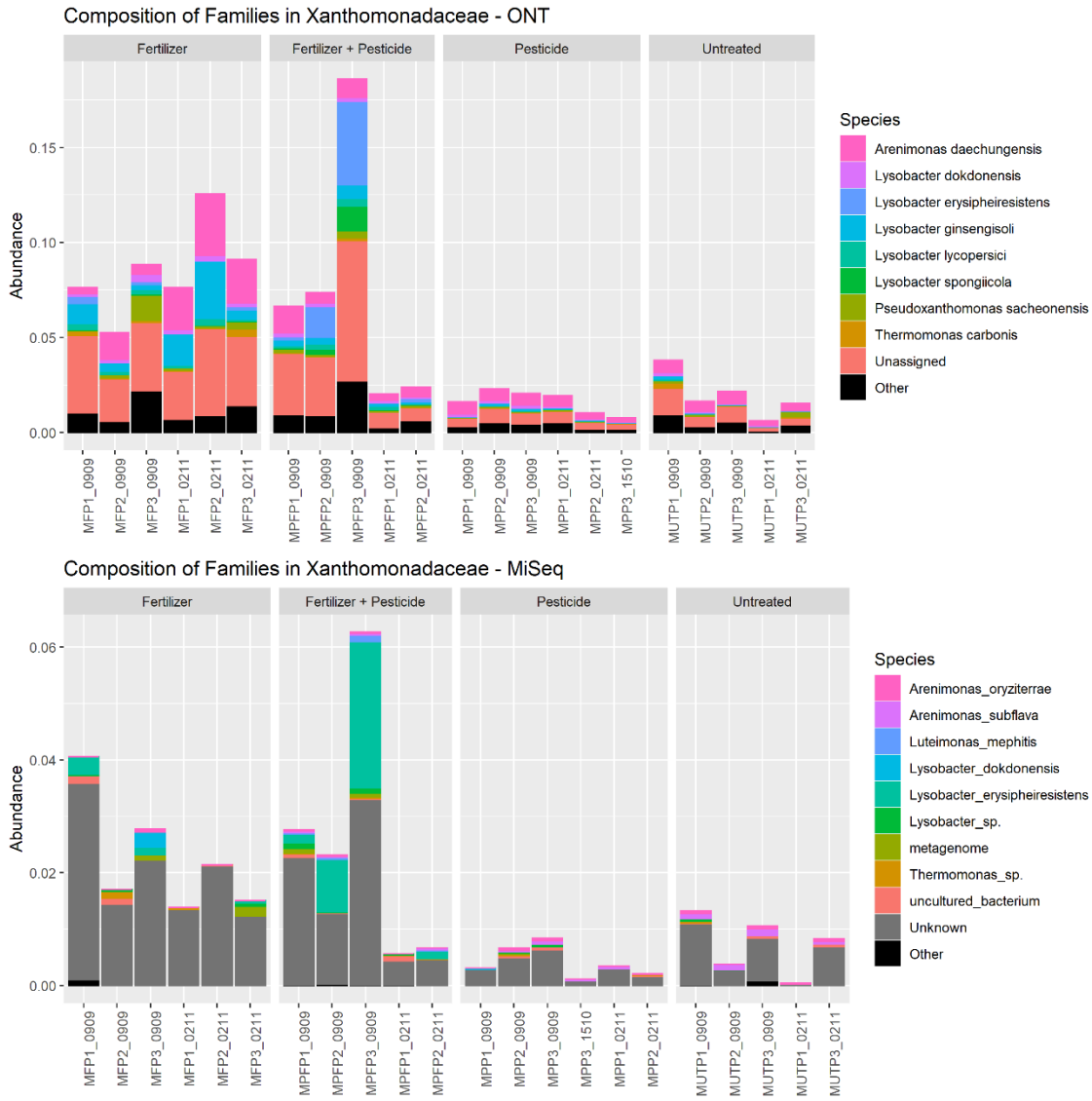


Figure 13: The figures show 9 top species within the family Xanthomonadaceae when analysed with MinION (top) and MiSeq (bottom). The four boxes represent the different treatments, and the columns represent the 22 samples which are listed on the horizontal line. The different colours represent the species written to the right of the figures.

The samples analysed with MinION had a higher relative abundance of reads assigned to *Xanthomonadaceae* compared to the samples analysed with MiSeq (Figure 13). At both technologies, the relative abundance of reads assigned were higher in the samples treated with fertilizer and fertilizer*pesticide than in the untreated and pesticide treated

samples. The analyses with MiSeq had a greater proportion of unknown reads than MinION. Another example of species' level was within the family *Bacillaceae.*

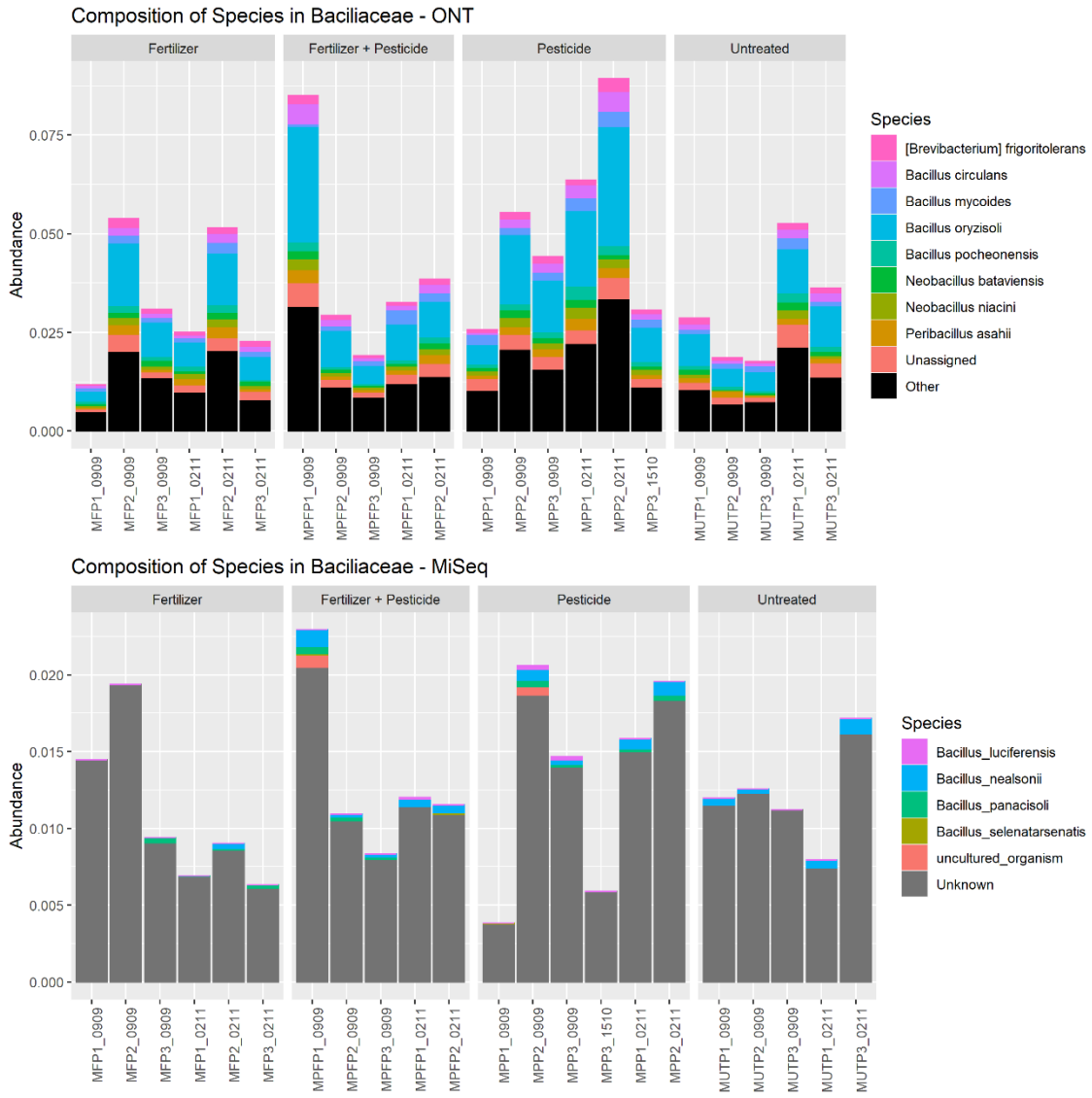


Figure 14: The figures show 9 top species within the family Bacillaceae when analysed with MinION (top) and MiSeq (bottom). The four boxes represent the different treatments, and the columns represent the 22 samples which are listed on the horizontal line. The different colours represent the species written to the right of the figures.

The analyses of *Bacillaceae* resulted in almost only unknown reads when analysed with MiSeq (Figure 14, bottom) and the few species represented belonged to the genus *Bacillus*. The analyses of *Bacillaceae* with MinION revealed several species (Figure 14, top) and the relative abundance of reads was higher compared to MiSeq.

4. Discussion

Soil microorganisms has several important roles, e.g. being decomposers and living in symbiose with plants, providing them nutrients and protecting them from pathogens (Khatoon et al., 2020). To understand and possibly improve the soil health, it is important to study soil microorganisms and learn more about their role in the ecosystem. Accurate data is crucial, and choice of method can be important. Next-generation sequencing can give us insight at the most basic level, and have made it possible to investigate microbial populations in their natural environment (Boers et al., 2019). Illumina MiSeq is one of the most popular sequencing platforms. It provides cheap and accurate data but is limited by the short sequencing length. An alternative to Illumina MiSeq is the Oxford MinION which deliver long, real-time reads (Quainoo et al., 2017).

The results of the alpha diversity analyses showed us that there was little variation between the two technologies. The highest number of observed species was 900 and 1050 when analysed with MiSeq and MinION, respectively. The difference can be explained by both longer and higher number of reads that were analysed with MinION. The correlation analyses between the two technologies were significant after removing two outliers. This is comparable with a study where 50 fecal samples were analysed with MinION and MiSeq (Wei et al., 2020). The authors found a significant correlation ($p = 0.0076$) of the taxonomic profiles at the genus level. However, their species-level assignment showed a substantial difference between MinION and MiSeq ($p = > 0.1$).

The taxonomic identification of bacteria at phyla level showed some differences between the two technologies. In particularly, Actinobacteria had greater abundance when analysed with MiSeq (approx. 40%) than with MinION (<15%) while Proteobacteria had greater abundance with MinION (30-60%) than with MiSeq (<30%). Chloroflexi was not one of the 9 top phyla detected with MinION but got quite large abundance with MiSeq (approx. 10-40%). The different abundancy between the phyla may be due to the reference databases used. We used SILVA and NCBI for taxonomic assignment for MiSeq and MinION analyses, respectively. NCBI 16S database contains around 18 000 16S ribosomal RNA reference sequences (O'Leary et al., 2016), while SILVA database contains 190 000 sequences. This skewed distribution of sequences may lead to different

dominating groups of bacteria and hence lead to different taxonomic assignment (Nygaard et al., 2020; Park & Won, 2018).

The choice of primer pairs may also affect the different taxonomic outcome, as they represent different regions of the 16S rRNA gene. Wasimuddin et al. (2020) tested four commonly used primer pairs in their study. They found that the primer pair 515F-806R revealed highest microbial diversity, and that the primer pair 799F-1193R was biased towards the genus *Bacillus* and hence detected fewer reads of all other genera. Also, Heikema et al. (2020) tested the ONT 16S rRNA gene primers on five pure cultures and found that both forward and reverse primer were not compatible with the 16S rRNA gene of the genus *Corynebacterium.*

At species level, MiSeq got large proportions of unknown reads. This may be due to that the SILVA database do not classify further than the genus level (Balvočiūtė & Huson, 2017). MinION revealed a slightly higher abundancy at species level than MiSeq (Figure 14). However, it is difficult to determine whether these results are reliable because of MinION's high error rates (Rang et al., 2018). In addition to the high error rates, MinION is also limited by the shortage of bioinformatic tools to correctly process long-read amplicon sequences. Most of the available software for taxonomic assignment are designed to process data from short-read sequences e.g., reads from Illumina MiSeq. This makes it challenging to perform valid taxonomic analyses of long-read sequences. Additionally, many reference databases consist mainly of fragments of the 16S rRNA gene and lack amounts of full-length sequences gene. This also may limit the taxonomic identification's reliability when analysing the full sequence of the gene (Santos et al., 2020).

Modified methods to improve the accuracy of MinION sequencing are tested by several scientists. One concept is rolling circle amplification (RCA). Single-stranded DNA is amplified repeatedly in circles, creating a long DNA-strand which contain the target sequence multiple times. The replicated sequences generate more reliable consensus reads. Increasing the number of RCA fragments increased the median accuracy value to 97% (Li et al., 2016) and 97.6% when sequencing on MinION (Baloğlu et al., 2020). However, consensus sequence building with RCA is a time-consuming step. Another

modified method is tested by Karst et al. (2021). They have developed a new amplicon sequencing approach where the DNA template is tagged with unique molecular identifiers (UMIs) during PCR. UMIs are like sample barcodes consisting of random bases and are part of the primers. Each target DNA molecule is tagged with a unique UMI, and the UMI is carried forward to a second round of PCR. After sequencing, the reads are grouped by what UMI they have retained and that allow us to generate high quality consensus sequences. When Karst et al. (2021) tested this method with UMIs and MinION, they sequenced full-length ribosomal RNA operons (~4400bp) in a mock microbial community and achieved results with a mean residual error rate under 0.005%. The length of the target sequence may also affect the accuracy of MinION sequencing. Cusco et al. (2018) used MinION to sequence the full-length 16S rRNA (1500 bp) and the 16S-ITS-23S region from the *rRNA* operon (4300 bp) from an isolate of *Staphylococcus pseudintermedius.* The amplicons from the 16S-ITS-23S region and the 16S rRNA gene were assigned correctly to bacterial species in 98% and 68%, respectively.

We treated the soil with both pesticide (glyphosate) and mineral fertilizers. The PCoA plot of Bray-Curtis dissimilarities (beta diversity) did not reveal a clear pattern with the samples analysed with MiSeq but showed a slight tendency of clustering based on the different treatments when analysed with MinION. However, the calculated results from the PERMANOVA tests showed very similar results between the two technologies. The samples treated with fertilizer gave a significant effect on the microbial community ($p = 0.001$) when analysed with both MinION and MiSeq. We also saw that the family *Xanthomonadaceae* got a greater abundance in the samples treated with fertilizer and fertilizer*pesticide than the other treatments at both technologies. This is contradictory results to a meta-analysis by Bebber and Richards (2020). They examined the impacts of organic and mineral fertilizers on soil microbial diversity, in comparison with unfertilized controls. The collected data from 31 studies revealed that there was no significant difference in taxonomic diversity for prokaryotes between soil treated with mineral fertilizer and control soil.

The PERMANOVA tests showed us that application of glyphosate gave no significant effect on the microbial community ($p > 0.05$). This is comparable with previous studies. Guijarro et al. (2018) investigated soil bacterial communities in two agricultural soil with

5 and almost 10 years of glyphosate application history, and a soil without previous exposure to glyphosate. In their study, they sprayed glyphosate over parts of the three soils and analysed collected soil samples before and after applied glyphosate. The results showed that the relative abundance of specific taxa could not be associated to glyphosate dissipation. Instead, they found that differences in bacterial community structure were strongly correlated with soil organic matter, cation exchange capacity, pH, and clay content. Although Guijarro et al. (2018) did not receive significant results regarding glyphosate, application of the pesticide may have an indirect effect on the soil microbial community. Schlatter et al. (2017) also found that repeated glyphosate applications had minimal impact on soil bacterial communities and diversity. However, a small increase of copiotrophic bacteria, which thrives in carbon rich soil, occurred. Schlatter et al. (2017) concluded that dying roots after application of glyphosate were colonized by fast-growing copiotrophic bacteria like Proteobacteria and Bacteroidetes.

5. Conclusion

We compared two next-generation sequencing technologies, MiSeq and MinION, by sequencing 22 soil samples for bacterial identification based on 16S rRNA gene targeting. The results of the alpha diversity were similar between the two technologies. Also, the beta diversity analyses based on effect of the treatments showed us same results for both technologies, where application of fertilizer gave a significant effect on the microbial community. When analysing bacteria groups, different phyla were dominating within the two sequencing platforms. However, we conclude that there was not a difference in the taxonomic resolution between Illumina MiSeq and Oxford MinION.

Sequencing with Illumina MiSeq produce a lot of data output with high accuracy but is limited by the short-read length and poor taxonomic classification at species level (Santos et al., 2020). MinION is restricted by its' high error rates. However, it retains advantages like its portability and real-time sequencing (Rang et al., 2018). Long-read amplicon sequencing is upcoming, and further improvements of the nanopore technology will make Oxford MinION a serious competitor to the short-read sequencing platforms in near future.

References

- Baloğlu, B., Chen, Z., Elbrecht, V., Braukmann, T., MacDonald, S., & Steinke, D. (2020). *A workflow for accurate metabarcoding using nanopore MinION sequencing* [Preprint]. Bioinformatics. https://doi.org/10.1101/2020.05.21.108852
- Balvočiūtė, M., & Huson, D. H. (2017). SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare? *BMC Genomics*, *18*(S2), 114. https://doi.org/10.1186/s12864-017-3501-4
- Bebber, D. P., & Richards, V. R. (2020). *A meta-analysis of the effect of organic and mineral fertilizers on soil microbial diversity* [Preprint]. Ecology. https://doi.org/10.1101/2020.10.04.325373
- Boers, S. A., Jansen, R., & Hays, J. P. (2019). Understanding and overcoming the pitfalls and biases of next-generation sequencing (NGS) methods for use in the routine clinical microbiological diagnostic laboratory. *European Journal of Clinical Microbiology & Infectious Diseases*, *38*(6), 1059–1070. https://doi.org/10.1007/s10096-019-03520-3
- Brevik, E. C., Pereg, L., Steffan, J. J., & Burgess, L. C. (2018). Soil ecosystem services and human health. *Current Opinion in Environmental Science & Health*, *5*, 87–92. https://doi.org/10.1016/j.coesh.2018.07.003
- Brevik, E. C., Slaughter, L., Singh, B. R., Steffan, J. J., Collier, D., Barnhart, P., & Pereira, P. (2020). Soil and Human Health: Current Status and Future Needs. *Air, Soil and Water Research*, *13*, 117862212093444. https://doi.org/10.1177/1178622120934441

- Cusco, A., Catozzi, C., Vines, J., Sanchez, A., & Francino, O. (2018). *Microbiota profiling with long amplicons using Nanopore sequencing: Full-length 16S rRNA gene and whole rrn operon* [Preprint]. Microbiology. https://doi.org/10.1101/450734
- De Corato, U. (2020). Disease-suppressive compost enhances natural soil suppressiveness against soil-borne plant pathogens: A critical review. *Rhizosphere*, *13*, 100192. https://doi.org/10.1016/j.rhisph.2020.100192
- Fadrosh, D. W., Ma, B., Gajer, P., Sengamalay, N., Ott, S., Brotman, R. M., & Ravel, J. (2014). An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. *Microbiome*, *2*(1), 6. https://doi.org/10.1186/2049-2618-2-6
- FAO, & ITPS. (2015). *Status of the world's soil resources: Main report.* Food and Agriculture Organization of the United Nations : Intergovernmental Technical Panel on Soils.
- Guijarro, K. H., Aparicio, V., De Gerónimo, E., Castellote, M., Figuerola, E. L., Costa, J. L., & Erijman, L. (2018). Soil microbial communities and glyphosate decay in soils with different herbicide application history. *Science of The Total Environment*, *634*, 974–982. https://doi.org/10.1016/j.scitotenv.2018.03.393
- Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003
- Heikema, A. P., Horst-Kreft, D., Boers, S. A., Jansen, R., Hiltemann, S. D., de Koning, W., Kraaij, R., de Ridder, M. A. J., van Houten, C. B., Bont, L. J., Stubbs, A. P., & Hays, J. P. (2020). Comparison of Illumina versus Nanopore 16S rRNA Gene

Sequencing of the Human Nasal Microbiota. *Genes*, *11*(9), 1105. https://doi.org/10.3390/genes11091105

- Hermans, S. M., Buckley, H. L., Case, B. S., Curran-Cournane, F., Taylor, M., & Lear, G. (2017). Bacteria as Emerging Indicators of Soil Condition. *Applied and Environmental Microbiology*, *83*(1), e02826-16, e02826-16. https://doi.org/10.1128/AEM.02826-16
- Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., Leopold, S. R., Hanson, B. M., Agresta, H. O., Gerstein, M., Sodergren, E., & Weinstock, G. M. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications*, *10*(1), 5029. https://doi.org/10.1038/s41467-019-13036-1
- Karst, S. M., Ziels, R. M., Kirkegaard, R. H., Sørensen, E. A., McDonald, D., Zhu, Q., Knight, R., & Albertsen, M. (2021). High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *Nature Methods*, *18*(2), 165–169. https://doi.org/10.1038/s41592-020-01041-y
- Khatoon, Z., Huang, S., Rafique, M., Fakhar, A., Aqeel Kamran, M., & Santoyo, G. (2020). Unlocking the potential of plant growth-promoting rhizobacteria on soil health and the sustainability of agricultural systems. *Journal of Environmental Management*, 20.
- Li, C., Chng, K. R., Boey, E. J. H., Ng, A. H. Q., Wilm, A., & Nagarajan, N. (2016). INC-Seq: Accurate single molecule reads using nanopore sequencing. *GigaScience*, *5*(1), 34. https://doi.org/10.1186/s13742-016-0140-7

- McCombie, W. R., McPherson, J. D., & Mardis, E. R. (2019). Next-Generation Sequencing Technologies. *Cold Spring Harbor Perspectives in Medicine*, *9*(11), a036798. https://doi.org/10.1101/cshperspect.a036798
- McGann, P., Bunin, J. L., Snesrud, E., Singh, S., Maybank, R., Ong, A. C., Kwak, Y. I., Seronello, S., Clifford, R. J., Hinkle, M., Yamada, S., Barnhill, J., & Lesho, E. (2016). Real time application of whole genome sequencing for outbreak investigation – What is an achievable turnaround time? *Diagnostic Microbiology and Infectious Disease*, *85*(3), 277–282. https://doi.org/10.1016/j.diagmicrobio.2016.04.020
- Mohanram, S., & Kumar, P. (2019). Rhizosphere microbiome: Revisiting the synergy of plant-microbe interactions. *Annals of Microbiology*, *69*(4), 307–320. https://doi.org/10.1007/s13213-019-01448-9
- Nygaard, A. B., Tunsjø, H. S., Meisal, R., & Charnock, C. (2020). A preliminary study on the potential of Nanopore MinION and Illumina MiSeq 16S rRNA gene sequencing to characterize building-dust microbiomes. *Scientific Reports*, *10*(1), 3209. https://doi.org/10.1038/s41598-020-59771-0
- O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733–D745. https://doi.org/10.1093/nar/gkv1189
- Orgiazzi, A., Dunbar, M. B., Panagos, P., de Groot, G. A., & Lemanceau, P. (2015). Soil biodiversity and DNA barcodes: Opportunities and challenges. *Soil Biology and Biochemistry*, *80*, 244–250. https://doi.org/10.1016/j.soilbio.2014.10.014

- Oxford Nanopore Technologies. (2020). *Product comparison*. https://nanoporetech.com/products/comparison
- Park, S.-C., & Won, S. (2018). Evaluation of 16S rRNA Databases for Taxonomic Assignments Using a Mock Community. *Genomics & Informatics*, *16*(4), e24. https://doi.org/10.5808/GI.2018.16.4.e24
- Plesivkova, D., Richards, R., & Harbison, S. (2019). A review of the potential of the MinION™ single-molecule sequencing system for forensic applications. *Wiley Interdisciplinary Reviews: Forensic Science*, *1*(1), e1323. https://doi.org/10.1002/wfs2.1323
- Quainoo, S., Coolen, J. P. M., van Hijum, S. A. F. T., Huynen, M. A., Melchers, W. J. G., van Schaik, W., & Wertheim, H. F. L. (2017). Whole-Genome Sequencing of Bacterial Pathogens: The Future of Nosocomial Outbreak Analysis. *Clinical Microbiology Reviews*, *30*(4), 1015–1063. https://doi.org/10.1128/CMR.00016-17
- Quick, J., Ashton, P., Calus, S., Chatt, C., Gossain, S., Hawker, J., Nair, S., Neal, K., Nye, K., Peters, T., De Pinna, E., Robinson, E., Struthers, K., Webber, M., Catto, A., Dallman, T. J., Hawkey, P., & Loman, N. J. (2015). Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. *Genome Biology*, *16*(1), 114. https://doi.org/10.1186/s13059-015-0677-2
- Rang, F. J., Kloosterman, W. P., & de Ridder, J. (2018). From squiggle to basepair: Computational approaches for improving nanopore sequencing read accuracy. *Genome Biology*, *19*(1), 90. https://doi.org/10.1186/s13059-018-1462-9
- RStudio, T. (2021). *RStudio: Integrated Development Environment for R.* RStudio. https://www.rstudio.com/

- Santos, A., van Aerle, R., Barrientos, L., & Martinez-Urtaza, J. (2020). Computational methods for 16S metabarcoding studies using Nanopore sequencing data. *Computational and Structural Biotechnology Journal*, *18*, 296–305. https://doi.org/10.1016/j.csbj.2020.01.005
- Schlatter, D. C., Yin, C., Hulbert, S., Burke, I., & Paulitz, T. (2017). Impacts of Repeated Glyphosate Use on Wheat-Associated Bacteria Are Small and Depend on Glyphosate Use History. *Applied and Environmental Microbiology*, *83*(22), e01354-17, e01354-17. https://doi.org/10.1128/AEM.01354-17
- Solden, L., Lloyd, K., & Wrighton, K. (2016). The bright side of microbial dark matter: Lessons learned from the uncultivated majority. *Current Opinion in Microbiology*, *31*, 217–226. https://doi.org/10.1016/j.mib.2016.04.020
- Srinivasan, R., Karaoz, U., Volegova, M., MacKichan, J., Kato-Maeda, M., Miller, S., Nadarajan, R., Brodie, E. L., & Lynch, S. V. (2015). Use of 16S rRNA Gene for Identification of a Broad Range of Clinically Relevant Bacterial Pathogens. *PLOS ONE*, 22.
- Wasimuddin, Schlaeppi, K., Ronchi, F., Leib, S. L., Erb, M., & Ramette, A. (2020). Evaluation of primer pairs for microbiome profiling from soils to humans within the One Health framework. *Molecular Ecology Resources*, *20*(6), 1558–1571. https://doi.org/10.1111/1755-0998.13215
- Wei, P.-L., Hung, C.-S., Kao, Y.-W., Lin, Y.-C., Lee, C.-Y., Chang, T.-H., Shia, B.-C., & Lin, J.-C. (2020). Characterization of Fecal Microbiota with Clinical Specimen Using Long-Read and Short-Read Sequencing Platform. *International Journal of Molecular Sciences*, *21*(19), 7110. https://doi.org/10.3390/ijms21197110