University of Stavanger

FACULTY OF SCIENCE AND TECHNOLOGY

MASTER THESIS

Curriculum: Biological Chemistry	Spring semester,2019 Confidential
Author: Isha Sehgal	 ..... (author signature)
Tutor: Daniela Maria Pampanin	
Master thesis title:	Can gut microbial community changes in Atlantic cod be used as biomarker of water pollution?
Keywords: Ecotoxicology, Atlantic cod, Gut microbiome, QIIME2, Amplicon sequence variants (ASVs), Vibrionales (Aliivibrio), Mycoplasmatales.	Number of pages: 55 + appendices/other: 62 Stavanger, 15.06.2019 date/year

UNIVERSITY OF STAVANGER

Can gut microbial community changes in Atlantic cod be used as biomarker of water pollution?

by

Isha Sehgal

A thesis submitted in partial fulfillment for the

degree of Master of Science in Biological Chemistry

in the

Department of Chemistry, Bioscience and Environmental Technology

Faculty of Science and Technology

June 2019

"If we knew what it was we were doing, it would not be called research, would it?"

Albert Einstein

Declaration of Authorship

I, ISHA SEHGAL, declare that this thesis titled, 'Can gut microbial community changes in Atlantic cod be used as biomarker of water pollution?' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: 

Date: 15/06/2019

Abstract

The gut microbiome of juvenile Atlantic cod (*Gadus morhua*) was studied to evaluate changes related to the exposure of dispersed crude oil and find potential biomarker of exposures. Fish was exposed to dispersed crude oil (0.05 ppm) for 1,3,7 and 28 days, as a reference a group of fish was maintained in the laboratory without oil exposure as a control group (n=4). Gut samples were dissected and stored at -80°C for analysis. Various DNA extraction methods were tested, to find the optimal conditions. The Illumina MiSeq platform was used for sequencing the V4-V5 region of the 16s rRNA gene.

The sequencing results were examined with the software- QIIME2 and the R package. An average of 359,912±183,643 quality filtered reads were generated and 3438 amplicon sequence variants (ASVs) were identified after data processing of 32 samples. A non-metric multidimensional scaling (NMDS) analysis based on Bray-Curtis distance matrix was done to understand the distribution of samples. Pielou's evenness and Shannon indices were used for diversity analysis. The taxonomy study revealed that the most abundant orders in the juvenile Atlantic cod gut were Vibrionales, Mycoplasmatales, Actinobacteridae, Alteromonadales, Rhodobacterales and Pseudomonadales. DESeq2 analysis depicted upregulation of 9 and downregulation of 65 ASVs.

From the obtained data, it was possible to conclude that the dispersed crude oil exposure promotes changes in the gut microbiota and affects their diversity. At the order level of classification, Vibrionales (Aliivibrio) and Mycoplasmatales seem to be indicative of exposure to the oil contamination. The present thesis work is building on the current knowledge base in this area of study and data generated is regarded value addition for future research.

Keywords: Gut microbiome, Atlantic cod, V4-V5 region, 16s rRNA, QIIME2, Amplicon sequence variants (ASVs), Vibrionales (Aliivibrio), Mycoplasmatales.

Acknowledgements

I express deep sense of gratitude to my supervisor Assoc. Prof. Daniela Maria Pampanin for her support, patience and encouragement all through my research work and thesis compilation. Her guidance has been a source of motivation throughout the project. And I am thankful to her for providing me an opportunity to be part of sampling at International Research Institute of Stavanger (now NORCE) Environment laboratory in Mekjarvik, Stavanger.

I would like to thank to Dr. Andrea Bagi for her guidance with the data treatment. Her generosity to make time for discussions and sharing data was an immense help for me during analysis.

Many thanks to Juline Walter for coaching me during the lab work and training me in gut sampling.

Finally, I would like to thank my husband for his support all through the studies and research work.

Contents

List of Tables

List of Figures

1 Introduction

1.1 Background to oil pollution problem

During the production of oil and gas from offshore oilfields, large amount of discharge water also known as produced water (PW) is released into the sea after treatment. Chemically, the produced water is complex and consists mainly of the polycyclic aromatic hydrocarbons (PAH) along with some alkylated phenols that are harmful for the sea organisms (Røe Utvik et al., 1999). Particularly the PAHs are of major concern due to their cancer and mutagenic potential and these can be present in dissolved as well as in sediment state in the PW. Thus, it is extremely important to monitor and control the discharge of PW in sea.

There are environmental regulations for oil and gas production offshore facilities for treating the PW prior discharge to sea. However, the treatment systems do not assure 100% efficiency and renders the discharge with dissolved low and high molecular weight hydrocarbons that are difficult to remove (Latimer & Zheng, 2003; Pampanin & Sydnes, 2013). For example, in Norwegian oil and gas sector, the Norwegian Environment Agency regulates discharge of PW in sea from offshore installations. However, significant volume of PW within the regulatory threshold of maximum allowable concentration of oil gets discharged to sea on daily basis.

Various methods of environment monitoring have been established and chemical monitoring had been the common choice for these studies in the earlier times. Unfortunately, with monitoring of chemicals in abiotic environment; it is only possible to determine the concentration of that pollutant, but the most important question remains unanswered i.e. what is the effect of pollutant on the organisms that get exposed?

To address the problem of biological effect, various environment monitoring programs have been trying to understand the threshold levels, accumulation and effects of organic pollutants inside the marine organisms' system. In this regard, the most developed and sought-after method to determine the detrimental effects of oil discharge exposure on organisms is to study biomarkers (Sanni et al., 2017).

1.1.1 Importance of biomarkers in environmental monitoring

The organisms respond when they are exposed to chemicals and the impact or the effect of these could be toxic. Biomarkers measure the effect of chemical exposure from molecular level to organisms' level, trying to extrapolate the information to predict the effect at population level. The use of biomarkers helps to predict if the organisms, living in a specific habitat are healthy or undergoing physiological changes that could develop into life threatening diseases (Peakall & Walker, 1994). The concept of biomarker allows to monitor the responses of a healthy organism to increasing levels of pollutant. It is expected that the health of the organism will eventually decline that may not be visible initially as a fully developed disease but there would be different kinds of physiological and biochemical changes that show the impact of pollutant on the organisms' health. It is possible for the organism to return to its normal healthy state by using its own repair mechanisms if the pollutant is expelled and withdrawn from its body- So biomarkers act as a forewarning of the implications of the pollution levels; that could be deadly for the ecosystem (Depledge & Fossi, 1994).

Fossi and Leonzio (1993) formulated a protocol for using biomarkers in analysing the risk for the ecosystem. This protocol was divided into 3 stages: 1st stage involved identification of ecosystems that were at risk. The 2nd stage was identifying the critical or endangered species and the target population. The 3rd stage analyzes the effect of the pollutants on the ecosystem and it was suggested to include laboratory as well as field studies to establish strong scientific proof for the impact of chemicals on the environment.

1.1.2 Atlantic Cod as bioindicator species

It is very common to use sea organisms like molluscs and fish in biomonitoring studies (Viarengo et al., 2007). To date, there have been several studies on different types of fish to understand the impact of oil pollution and among the various species studied, there is one that is very common and that is, Atlantic cod (Gadus morhua). Atlantic cod is extensively distributed and has a high marketable value that makes it a popular sentinel organism (Nahrgang et al., 2013; Pampanin et al., 2016). Cod thrives in various regions of the European coast as shown in the figure 1.

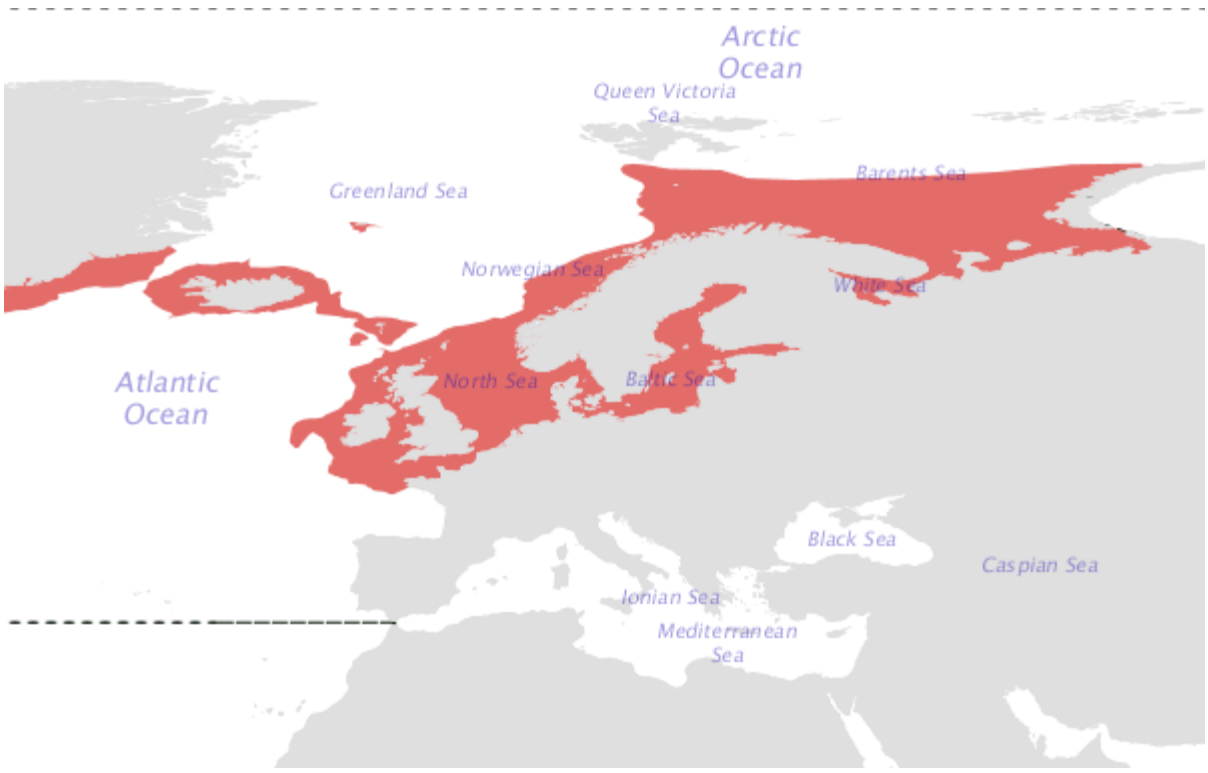

Figure 1. Distribution map of Atlantic cod (Image Source: Food and Agriculture Organization of the United Nations)

The effect of oil exposure on cod has been studied well and there have been reports of the presence of PAH in fish bile, changes in the Ethoxyresorufin O-deethylase (EROD), catalase (CAT) and glutathione S-transferases (GST) activities in the liver. It is also reported that exposure to PAH leads to formation of DNA adducts (Sundt et al., 2012; Pampanin et al., 2016). Reports of changes in gut microbiota due to oil pollution have also been published (Bagi et al., 2018). However, there is a huge scope in this area and further studies have to be conducted to understand clearly the normal microbiota in the gut of fish and also the transformation or changes in the gut microbial community due to the pollution impact.

The effect of oil exposure on plasma proteins has been also studied in cod. Enerstvedt et al. (2017) studied proteins in the blood plasma of cod fish after injecting it with naphthalene, chrysene and their dihydrodiols. The study found that the PAH exposure produced an immune response and showed upregulation of certain proteins that could act as protein biomarkers of exposure.

1.1.3 Gut microbiota

The gut microbiota serves multiple functions, they aid in digestion, metabolism of xenobiotic compounds, and immunity of the host organism (Ghanbari et al., 2015; Bagi et al., 2018). The study and characterization of gut bacteria saw a new light in the advent of next generation sequencing technique with immense potential to untap the hidden facts related to bacterial composition, their role in metabolism of xenobiotics and the functions of various genes that were not known earlier because of unculturable bacterial species (Ghanbari et al., 2015). The study of gut microbes in the gastrointestinal cavity of marine organisms offers a huge opportunity since very limited studies have been conducted in this discipline (Bagi et al., 2018; Johny et al., 2018).

1.1.4 Next Generation Sequencing

Frederick Sanger developed the first DNA sequencing technology and later on newer and improved versions of sequencing technique came into existence. These new sequencing methods are called as next generation sequencing technologies and they are categorized further into second, third and fourth generation based on various parameters (Kumar & Kocour, 2017).

Next generation sequencing (NGS) is a popular and a cost-effective DNA sequencing technique, that is widely used in genomic research. It is also known as parallel or deep sequencing in which millions of tiny DNA pieces or fragments are sequenced in parallel at a very high speed. With the help of bioinformatics and its various tools, the small fragments are joined together to map the whole genome of the organism (Behjati & Tarpey, 2013). The NGS platform includes different steps in its protocol before finally giving the sequencing results. One of the most important steps is the library preparation that involves the attachment of short adapter sequences to the fragmented DNA before they are amplified. It is worthy to note that only a few nanograms of DNA is enough to prepare the library (Mardis, 2008). This technique bypasses the bacterial clone preparation because it does not need the cellular environment to prepare the library (Dijk et al., 2014). These high throughput technologies like any other technology undergo evolution and better versions become available with time. However, all of these techniques have the common workflow and can be divided into: (a) processing of sample that involves extraction of DNA, its quantification using nano-drop or qubit and preparation of library; (b) use of bioinformatics to generate and analyze the sequences (Gargis et al., 2016). A general workflow of NGS is given in figure 2.

As depicted in the figure 2, the dry laboratory work (bioinformatics part) encompasses primary, secondary and tertiary analyzes. In the primary analyzes, quality scores are given to base calls, generated from signals of the instrument. These scores depict the likelihood of having the

correct base in place. During secondary treatment of data, the information is processed further, and the sequence reads are compared to the available dataset, for example a reference sequence. In cases, where the reference sequence is not available then de novo assembly technique is used to create the whole novel sequence. In the tertiary part, interpretation and reporting of results takes place that includes the annotation of genome and taxonomic profiling (Gargis et al., 2016).

The fragments of DNA produced during sequencing can vary in length; short length reads, or long-length reads (Gargis et al., 2016; Ambardar et al., 2016). Various NGS technologies differ in their length of reads, sequencing mechanism and output, that also affects the cost and run time (Loman et al., 2012).

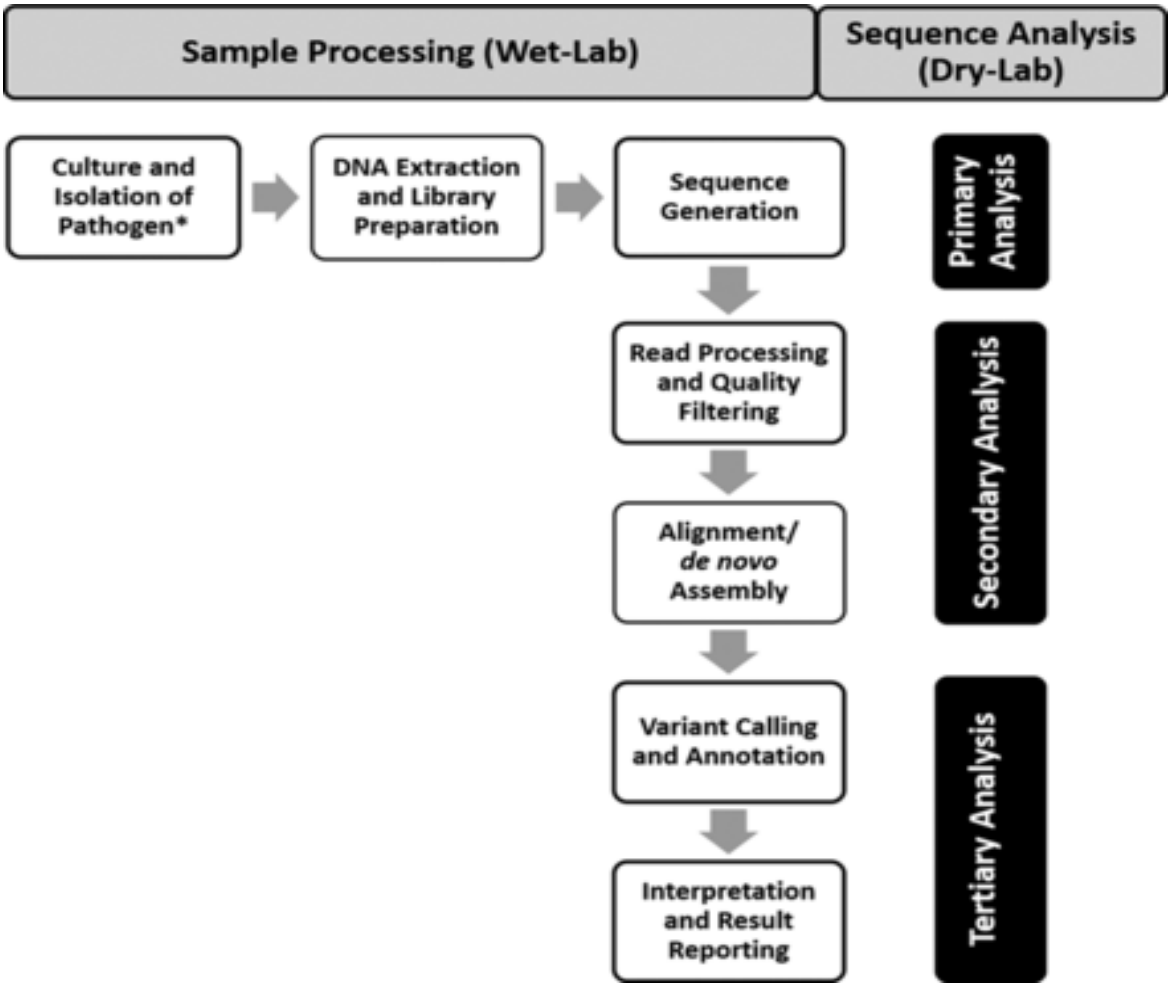

Figure 2. A general flow of next generation sequencing (Source: Gargis et al., 2016)

1.1.5 Illumina Sequencing

The study of microbial DNA from environmental samples has been gaining momentum with sequencing technologies becoming more advanced and cost efficient in recent times. Illumina, a popular name in genomics technologies launched their first DNA sequencing machine in the year 2006; built on the principle of sequencing by synthesis involved the use of a glass slide called flow cell. The flow cell consisting of eight lanes is lined on the interior surface with oligonucleotides, attached covalently to these cells. These oligos have sequences complimentary to the adapters that are attached to the library fragments (fragmented DNA of interest). When these library fragments are loaded on the flow cell, the oligos on the flow cells hybridizes with the library fragments that is followed by bridge amplification steps (the DNA fragments bend over forming bridges) mediated by an isothermal polymerase. The amplification takes place, generating millions of clusters of library fragments. The next step is the sequencing, and it starts with the supply of polymerase and fluorescently labelled nucleotides, which are chemically altered at their 3´ end to allow incorporation of only a single base for every cycle. The image of the cluster helps to identify the base that gets integrated. The following step deblocks the 3´ end by removing the fluorescent group that allows the addition of next base to the cluster. At the end, the sequence is recorded and undergoes quality checks to filter low quality reads (Mardis, 2008; Shokralla et al., 2012).

1.1.6 The illumina workflow

The steps involved in the illumina sequencing are depicted in the figure 3.

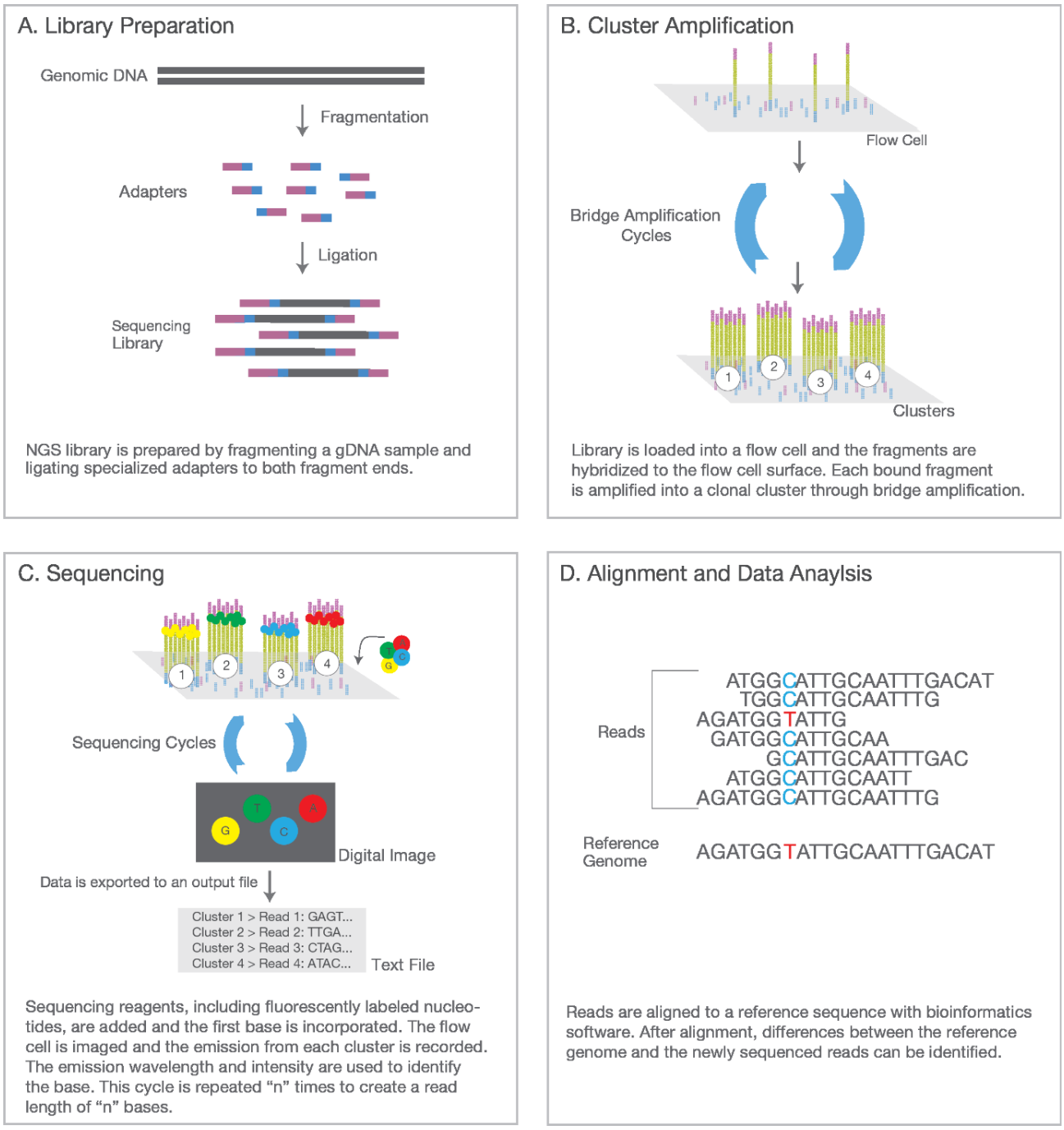

Figure 3. The steps in the illumina workflow (Image Source: Illumina)

1.1.7 Taxonomic Profiling

Results from sequencing are used to group the microorganisms into various taxa along with their relative abundance and that is known as taxonomic profiling. This classification can be done based on sequencing of marker genes or the whole genome. These two approaches are used to study microbiota in environmental samples and different terminologies like metataxonomics and metagenomics are commonly used. The term metataxonomics is used for studies involving sequencing of marker genes like 16s rRNA (a highly conserved region in prokaryotes), 18s rRNA (in eukaryotes) or internal transcribed spacer (ITS) in fungi. On the other hand, metagenomics is the sequencing of all the genomes (containing all the genes) of microbes in an environmental sample (Breitwieser et al., 2017). However, the rRNA sequencing had also been called as metagenomics in the previous studies. According to Marchesi and Ravel (2015), these terms imply different meanings and are ought to be used correctly to avoid confusion and misuse among the scientific community.

In both approaches, the sequence reads go through the binning process. The reads are then grouped into operational taxonomic units (OTUs) or amplicon sequence variants (ASVs). The clustering is based on the following:

I. similarity threshold value of sequence reads that are compared with a reference database
II. nucleotide composition such as GC content
III. combination of I and II

1.1.8 ASVs versus OTUs

Amplicon sequence variants (ASVs) are becoming more popular than Operational Taxonomic Units (OTUs) for the illumina sequencing analysis. The ASV method is more sensitive and specific than OTUs. While the OTU method depends on specifying a threshold value for the clustering of sequences, the ASVs deduces the sample sequence before any kind of error from amplification or sequencing gets introduced into the sequences. With ASV method, it is possible to distinguish sequences that differ in as small as one nucleotide.

Since OTU clustering works either on relative abundances or on comparing the processed reads against a reference database, it is likely that any variation that is not presented by the reference database would not be reported and get lost. On the other hand, ASVs provide all kind of variations from the sample, since it combines both *de novo* OTUs and closed reference OTUs. The ASV data is also more reliable, reproducible and a valid comparison between samples could be made (figure 4).

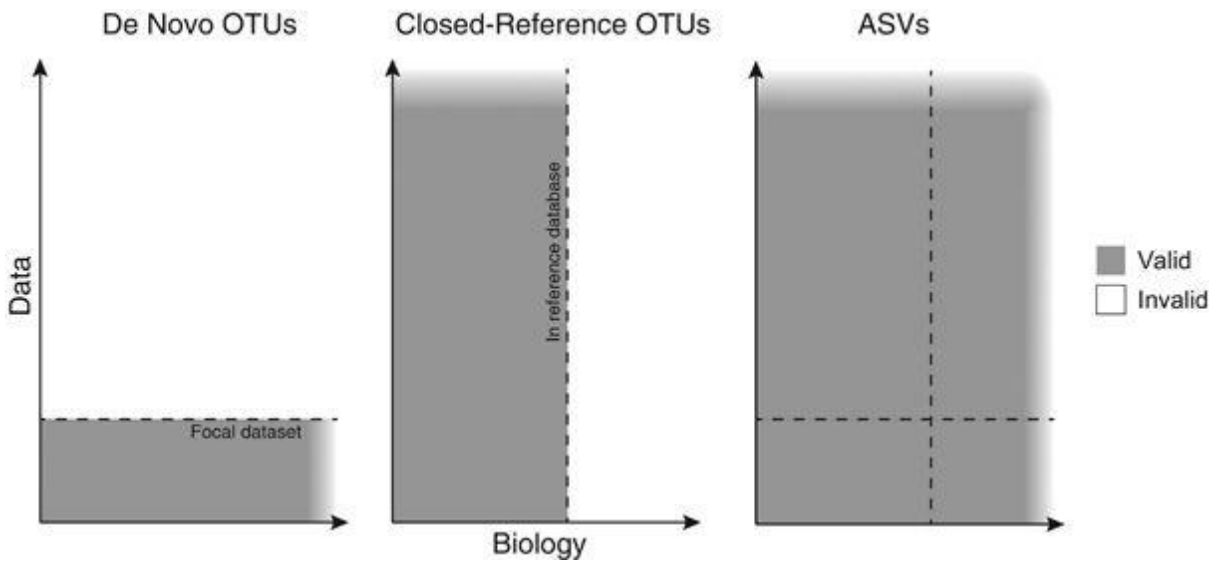

Figure 4. Comparison of data validity for de novo OTUs, closed reference OTUs and ASV method (Callahan et al., 2017)

1.2 Aim of the study

The present work aims to achieve the following goals:

1. to study the microbial community changes in the gut of juvenile cod fish (*G. morhua*);
2. to compare results with other research work and;
3. to establish a biomarker of water pollution based on the results of gut microbiome.

2 Materials and Methods

2.1 Sample Collection

Juvenile Atlantic cods were sourced from Centre for Marine Aquaculture (Nofima), Tromsø and brought to the International Research Institute of Stavanger (now NORCE) Environment laboratory in Mekjarvik (Stavanger). Large glass fibre tanks of 1000 litre capacity were used to store the fish (figure 5). Salinity, temperature, flow of water, light and darkness levels were controlled as well as monitored throughout the study. About 2 weeks were given to allow the fishes to acclimate.

Figure 5. Storage of juvenile cods at NORCE Environment laboratory in Stavanger.

2.2 Exposure

Fish were exposed to crude oil to mimic produced water in sea. The oil was administered at a concentration of 0.05 ppm and the samples were divided as- 64 in control group and 64 in the oil exposed group, a total of 128 samples. The exposure set up involved big glass tanks supplied with one inlet for sea water and the other for regulated dispersion of oil droplets into the tank. The set up was based on continuous flow system (CFS) as described in Sanni et al., 1998 (figure 6). There was no mortality during the entire study period.

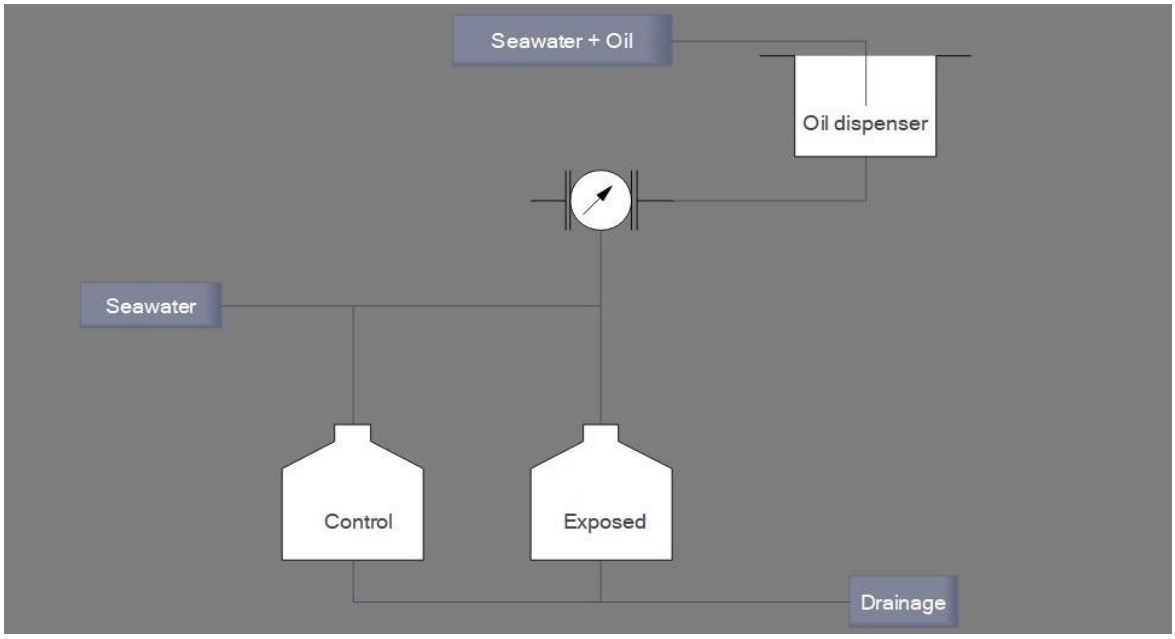

Figure 6. Exposure set up (adapted from Bagi et al., 2018)

2.3 Sampling

The juvenile cods were anaesthetised and then sacrificed by a sudden blow on their head. Fish were dissected for sample collection. The workbench was cleaned each time by wiping with ethanol. Sterilized scissors, scalpels and blades were used for sampling the gut.

Samples were collected in cryotubes, quickly frozen into a box of liquid nitrogen and stored at -80°C for further analysis.

The sampling was done after day 1, 3, 7, and 28 days. The general health condition of each fish was evaluated by calculating the condition index (CI) and hepatosomatic index (HSI), that are based on the length of the fish, whole body weight and liver weight.

The CI (Lambert & Dutil, 1997), also known as Fulton factor, is calculated as:

$$K = W/L\text{^}3 \times 100$$

where W, L is the weight and length of fish.

The HSI, a measure of energy reserves of the body was calculated by the formula given below (Lambert & Dutil, 1997):

$$HSI = \frac{LW}{W} \times 100$$

Here, LW and W is the liver weight and somatic weight of the fish.

2.4 Method Optimization for DNA Extraction

Various methods of DNA extraction are available therefore, the first step was to establish a suitable method for cod gut samples. The frozen gut content stored in the freezer was thawed on ice to proceed with DNA extraction.

2.4.1 Phase I

In the initial phase of experimentation, the whole gastrointestinal tract of adult cod fish was used to process the gut content (figure 7).

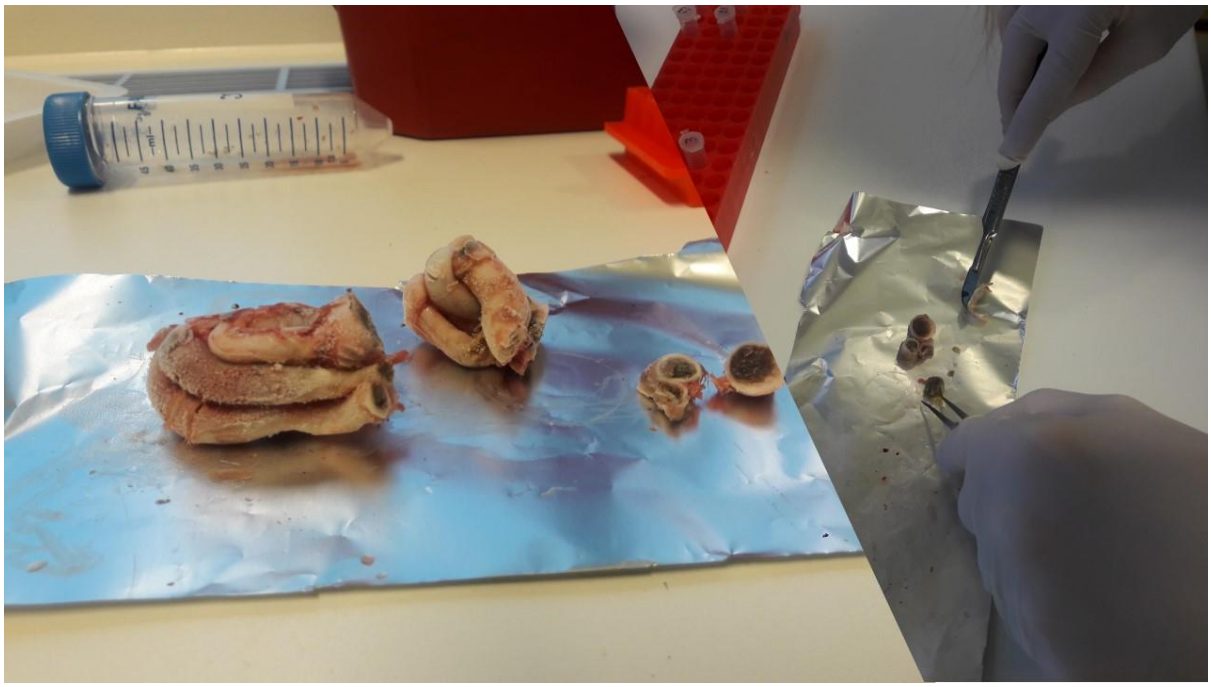

Figure 7. Processing of gastrointestinal tract for extraction of DNA.

Various methods for DNA extraction were tried in combination with different purification techniques for the gut samples (figure 8).

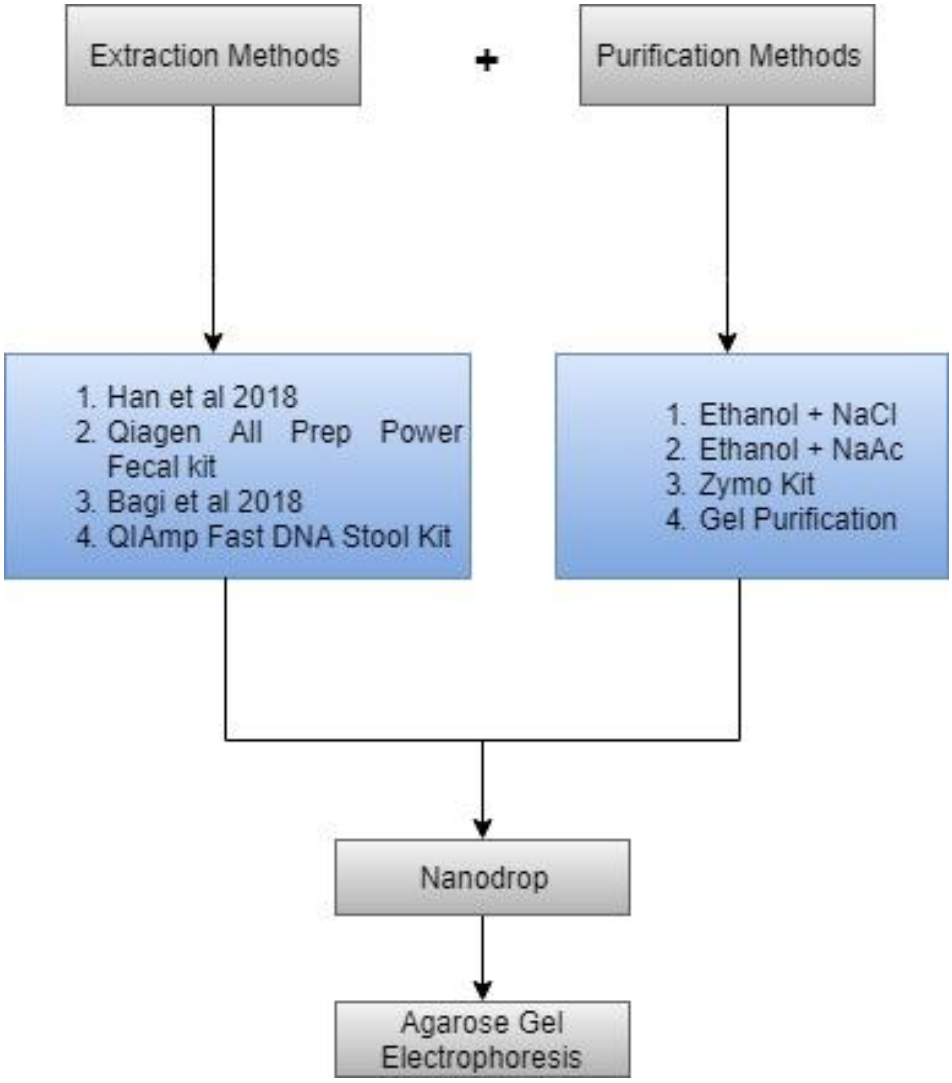

Figure 8. DNA extraction method optimization chart.

2.4.1.1 Han et al. 2018

Some modifications were made to this protocol like a pre-wash step with distilled water was introduced in the start of the protocol. Hexadecyl trimethyl ammonium bromide (CTAB), instead of Tris-EDTA (TE) buffer, was used along with lysis buffer and lysozyme for chemically disrupting the cells. Tissue lyser (3 min at 30 Hz) was used for physical disruption of the cells. Enzymes like proteinase K and RNase were used to get rid of impurities. Equal volumes of phenol: chloroform: isoamyl alcohol (PCI) (25:24:1) followed by treatment with equal volume of chloroform: isoamyl alcohol (CI) was used for precipitating the DNA. This

was followed by adding Sodium acetate (1/10) volume, subsequent washing with 95% ice-cold ethanol, overnight incubation at -20°C, centrifugation at 4°C and washing with 70% ethanol two times. The DNA pellet was dried and resuspended in 10 mM Tris-HCl (pH 8). The amount of eluted DNA was measured with Nanodrop spectrophotometer.

2.4.1.2 Qiagen All Prep Power Fecal Kit

The protocol was followed as per the manufacturer´s instructions. Around 0.2 g of gut sample was weighed and then homogenised after addition of lysis buffer, lysozyme and glass beads. Dithiothreitol (DTT) was also used to aid disruption of the cell membrane. The kit was equipped with a special inhibitor removal technology that works on removing contaminants that cause interferences during the downstream processing. The DNA MinElute Spin columns allowed effective binding of DNA and it was cleaned further by washing twice with wash buffers (AW1, AW2). In the end, the DNA was eluted with the elution buffer and quantified using Nanodrop (figure 9).

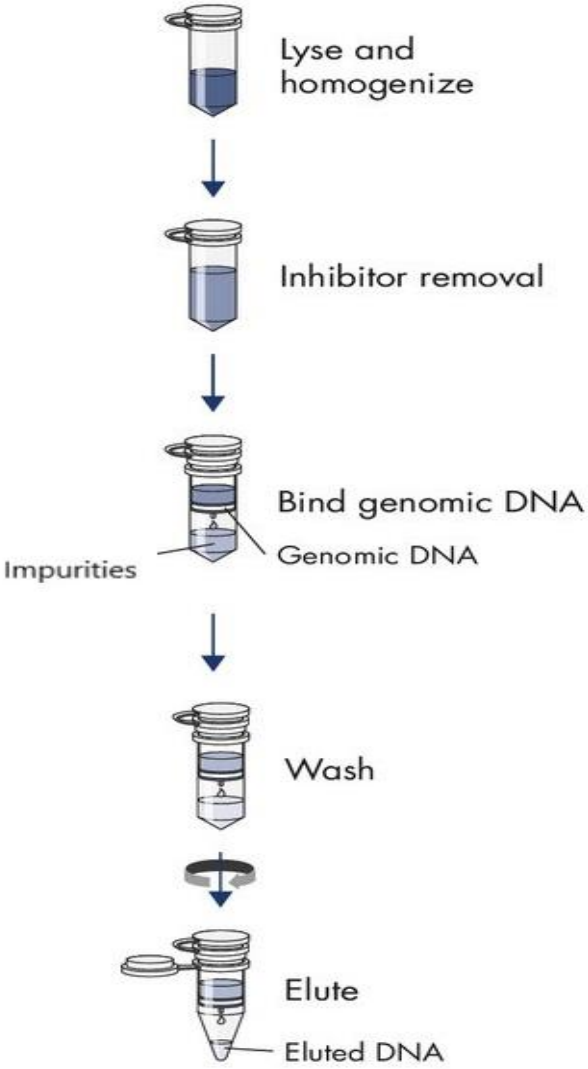

Figure 9. The work flowchart of Qiagen All Prep Power Fecal Kit (Adapted from QIAGEN)

2.4.1.3 Bagi et al. 2018

Few changes were made to this method that are briefly described in the following steps:

1. 0.15 g sample was taken and 1 mL of ATL lysis buffer was added.
2. The contents were vortexed and tissue lyser was used two times followed by incubation on ice for 5 minutes.
3. The sample was vortexed again and then centrifuged at 10,000 rpm for 1 minute at room temperature.
4. The supernatant was transferred into a fresh 1.5 mL microcentrifuge tube and kept for overnight incubation at 55°C.
5. 20 µL of RNAse A (100 mg/mL) was added for 30 minutes digestion of RNA at 37°C.
6. To approximately 600 µL of lysate, 1 mL of phenol : chloroform : isoamyl alcohol (PCI, 25:24:1) was added and the samples were mixed vigorously.
7. It was followed by 15 minutes incubation on ice and centrifugation at 5000 rpm for 15 minutes at 4°C.
8. The supernatant was transferred into a fresh tube and step 7 was repeated again with 1 mL of PCI.
9. To the supernatant obtained from step 8, 1 mL of chloroform : isoamyl alcohol (CI, 24:1) was added.
10. The next step was to add 2 volumes of ice-cold ethanol (100%) and sodium acetate (1/10, 0.3 mM, pH 5.3).
11. The contents were mixed and incubated at -20°C for overnight.
12. The samples were centrifuged at 21000 rpm for 30 minutes at 4°C.
13. Washed twice with 500 µL of 70% ethanol and dried for an hour.
14. The DNA pellet was re-suspended in 60 µL of milliq water.
15. The quantification was done on nanodrop spectrophotometer.

2.4.1.4 QIAmp Fast DNA Stool Mini Kit

The kit was used according to the manufacturer´s instructions. All centrifugations were done at 20,000 rcf at 22°C. The gut DNA sample (stool) was weighed between 180-220 mg in a 2 ml microcentrifuge tube and placed on ice. After that, 1 ml of InhibitEX buffer was added and the sample was vortexed for a minute to mix it properly. The buffer brought about lysis of the bacterial cells. This mixture was heated at 70°C for 5 minutes and vortexed for 15 seconds. The sample was then centrifuged for 1 minute to obtain pellet of the sample. A new 1.5 ml microcentrifuge tube was taken and 15 µl of proteinase K was pipetted into it. Around 200 µl of the supernatant from the centrifuged tube was pipetted into this tube containing proteinase K and about 200 µl of buffer AL was also added to the same tube. The sample was vortexed for 15 seconds and then incubated in a thermomixer at 70°C for 10 minutes. After the incubation, 200 µl of 100% ethanol was added and mixed by vortexing for few seconds. Following this, the QIAmp Spin columns were taken and placed into a 2 ml collection tube. Around 600 µl of the lysate from the previous step was applied onto the column and centrifuged for a minute. The spin column was placed into a new collection tube and the tube containing the filtrate was discarded. This step was repeated until all the lysate had been applied on the column. The next step was to add 500 µl of buffer AW1 to the column and again centrifuge for 1 minute. The tube containing the filtrate was discarded and the spin column was transferred into a new collection tube. Then 500 µL of buffer AW2 was added and the centrifugation was done for 3 minutes. The tube that contained the filtrate was discarded and the column was now transferred into a new 1.5 ml microcentrifuge tube. The last buffer to be added was the elution buffer and around 80 µl of buffer ATE was pipetted onto the column, incubated for 5 minutes at room temperature and the centrifuged for 1 minute for elution of DNA. A pictorial view of the protocol is shown in figure 10.

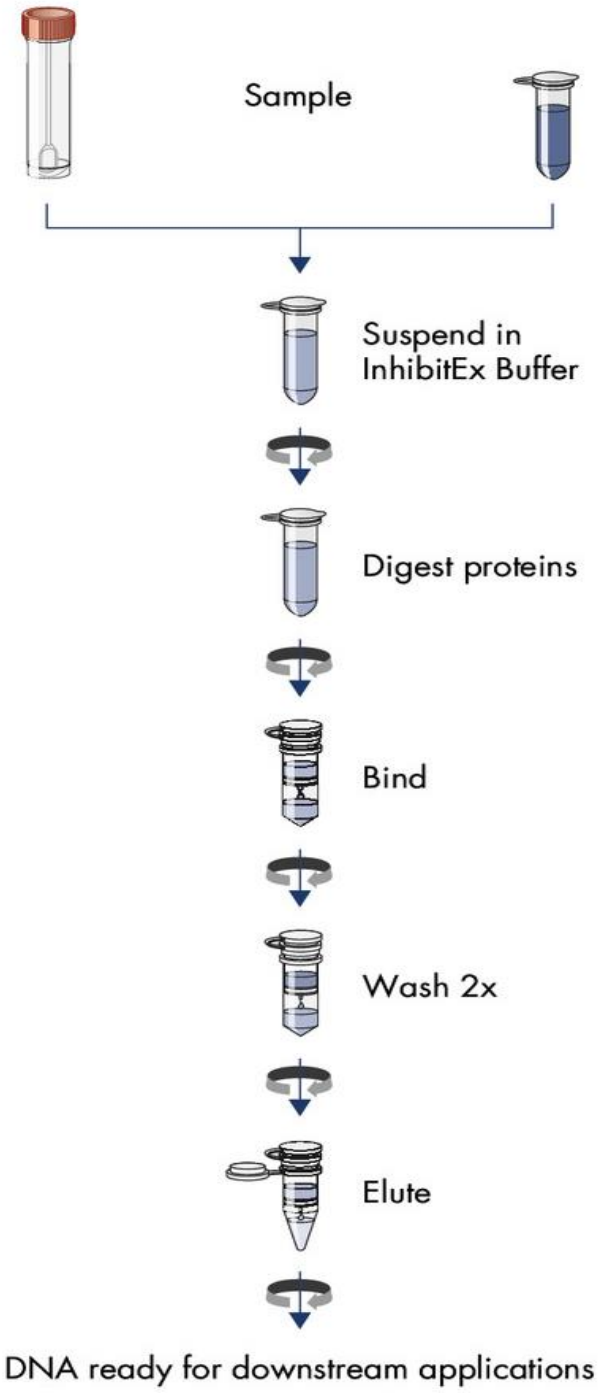

Figure 10. The DNA extraction protocol for QIAmp Fast DNA Stool Mini Kit (Image Source: QIAGEN)

2.4.1.4.1 Purification techniques

Certain purification methods were also used after DNA extraction process to improve the purity levels further.

2.4.1.4.1.1 Ethanol Purification

1. The eluted DNA from the extraction process was treated with 180 µL of 100% ice-cold ethanol, 6 µL of sodium chloride (NaCl, 5M) and 2 µL of magnesium chloride ($MgCl_2$, 1M).
2. The contents were mixed and the sample was incubated for overnight at -20°C.
3. Centrifugation was done at 13,000 g for 30 minutes at 4°C.
4. The supernatant was decanted and the pellet was washed with 1 mL of 70% ethanol (ice-cold).
5. The pellet was centrifuged again at 13,000 g for 10 minutes.
6. The supernatant was decanted.
7. The pellet was kept for drying for about an hour and then resuspended in TE buffer (pH 8).
8. The amount of DNA was measured with Nanodrop.

Another variation of this same purification method was also used but instead of using sodium chloride, 3M sodium acetate (NaAc, pH 5.2) was used and rest of the protocol was followed in the same manner.

2.4.1.4.1.2 Zymo Purification

The Genomic DNA Clean and Concentrator kit from Zymo was used according to the instructions from the manufacturer.

1. DNA binding buffer (2-7) times the volume of DNA samples was taken in a 1.5 mL microcentrifuge tube (if DNA sample was 60 µL then 120 µL of binding buffer was added to it).
2. The sample and the DNA binding buffer were mixed by vortexing.
3. The mixture was transferred to the Zymo-Spin Column placed in a collection tube.
4. Centrifugation was performed at 15,000 g for 30 seconds and the flow-through was discarded.
5. The sample was washed with 200 µL of DNA wash buffer and centrifuged again for 30 seconds.
6. The washing was repeated one more time and then about 80 µL of DNA Elution buffer was added to the column and then incubated for 5 minutes at room temperature.

7. The column was transferred to a new 1.5 mL microcentrifuge tube and centrifuged for 30 seconds for eluting the DNA.
8. The DNA was quantified using Nanodrop.

2.4.1.4.1.3 DNA Gel Extraction

The Monarch DNA gel extraction kit was used from New England BioLabs to excise DNA directly from the gel. The kit was used as per the instruction manual from the manufacturer (figure 11).

1. The DNA was excised from the agarose gel with the help of a sterilized scalpel and transferred into a microcentrifuge tube (1.5 mL).
2. The gel slice was weighed and Monarch gel dissolving buffer (4 times or 4 volumes to the DNA gel slice) was added to it.
3. It was incubated at 50°C with periodic vortexing to dissolve the gel.
4. Then the above sample was applied onto a column placed in a collection tube and centrifuged for a minute at 16,000 g.
5. The flow-through was discarded and the column was transferred into a new collection tube.
6. 200 µL of DNA wash buffer was added and the sample was spinned again for a minute.
7. The flow-through was discarded and the washing was repeated once again.
8. To elute the DNA, 20 µL of Elution buffer was added and incubated for a minute at room temperature.
9. The sample was centrifuged for 1 minute and the eluted DNA was collected.

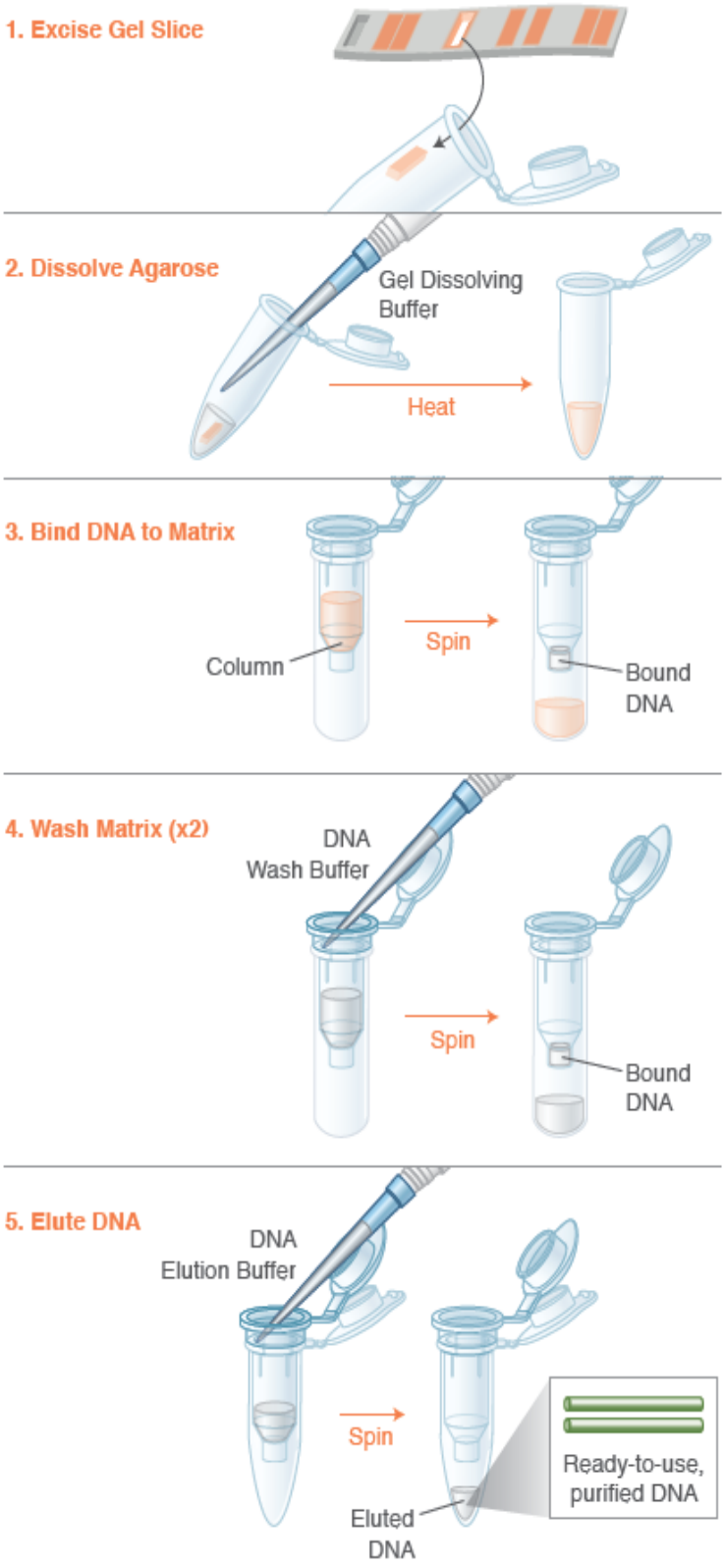

Figure 11. The workflow for Monarch Gel Extraction Kit (Image Source: New England BioLabs)

2.4.2 Phase II

After the first phase of trial, the results from all methods of DNA extraction were compared and it was noted that the Qiagen kit (Fast DNA Stool Mini Kit) gave consistent results for the purity ratios with very low standard deviation (figure 12). It was also noted from the purification techniques that there was loss of DNA even though the purity of the samples enhanced but there was no significant improvement. It was also seen that from all the purification methods tried, the Monarch gel extraction technique led to significant losses in the amount of DNA probably due to the tricky way of cutting the band of DNA from the gel. Therefore, it was decided to go ahead with the extraction of DNA using the Qiagen Fast DNA Stool Mini Kit without applying any purification technique to get the desired results.

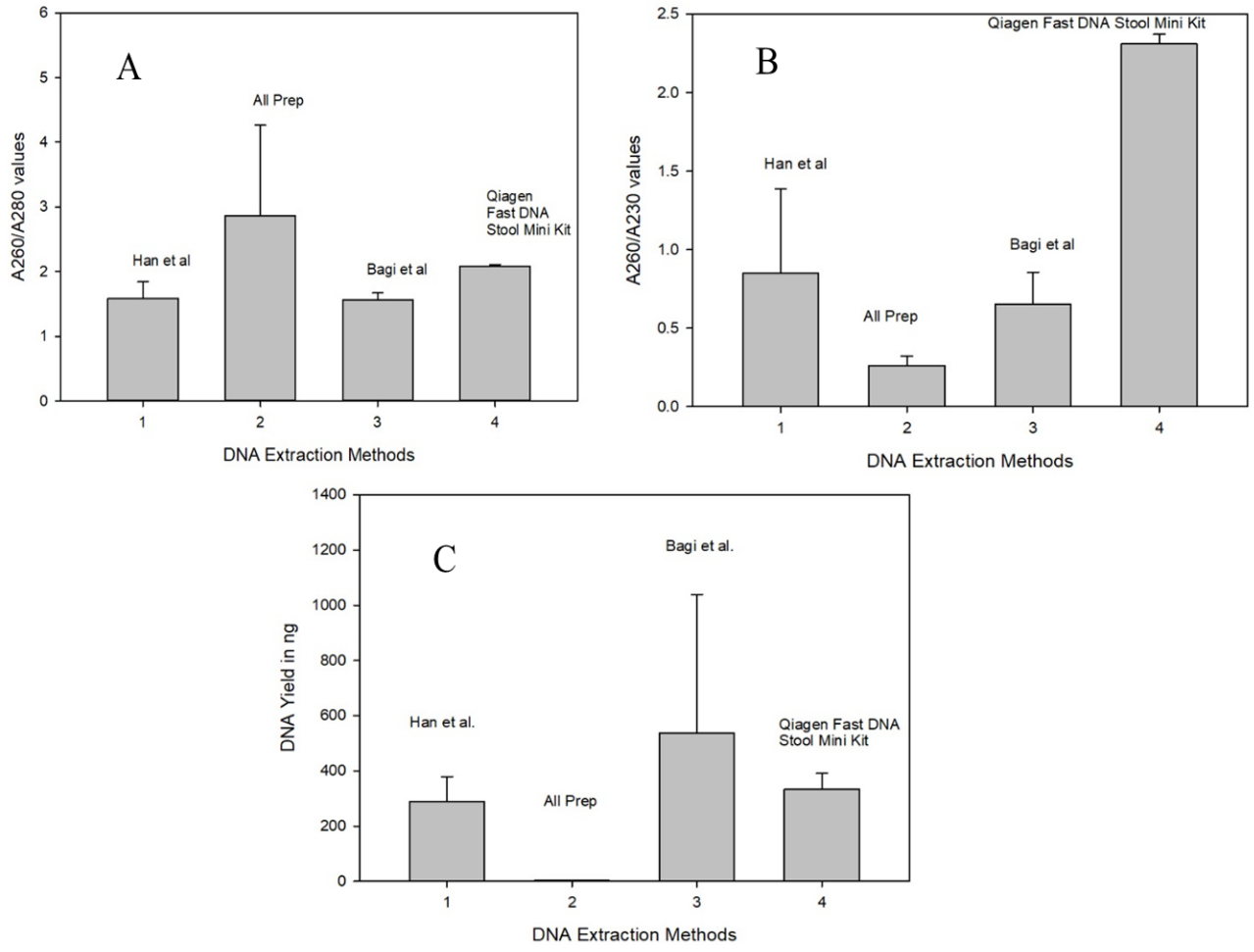

Figure 12. The methods were compared by calculating standard deviation on A260/A280 ratio shown in plot A, the A260/A230 ratio shown in plot B, and the amount of DNA in nanograms (ng) shown in plot C

2.4.2.1 Final Run of extraction with optimized method

Depending on the results of DNA extraction, the final set of experiment was performed with the QIAmp Fast DNA Stool Mini Kit. The frozen gut content stored at -80°C was thawed on ice for further processing with the kit.

Based on the average HSI and Fulton C values, 4 samples from each set-control and those exposed to 0.05 ppm petroleum oil were selected (table 1).

Table 1 The data shows the general health condition of the fish at the time of sampling.

Sampling day	**Day 1**		**Day 3**		**Day 7**		**Day 28**	
Sample set	Average HSI	Average Fulton C	Average HSI	Average Fulton C	Average HSI	Average Fulton C	Average HSI	Average Fulton C
Control	7.77	0.8	8.37	0.87	7.75	0.86	9.17	0.93
Exposed	7.85	0.81	8.14	0.8	8.65	0.85	9.27	0.91

A comparison would be made between the results from samples exposed for 1,3,7, and 28 days and the control samples. These results along with the previous studies would be discussed in detail and finally concluded at the end of this report.

2.5 Sequencing

After DNA extraction, the samples were sent to IIGB HT Sequencing facility in the University of California Riverside (UCR), California, USA. The library preparation and sequencing was done at UCR. TruSeq Nano DNA sample preparation kit was used to prepare the amplicon libraries for 32 samples using the indexed and barcoded V4-V5-primers: 799Fmod3 (5'-CMGGATTAGATACCCKGG-3') and 1115R (5'-AGGGTTGCGCTCGTTG-3'). Bioanalyzer was used for quality control of the libraries. The quantification was done with Qubit and post quantifying the DNA, the samples were loaded into the sequencer. The illumina MiSeq benchtop sequencer (2×300 bp paired end) was used to sequence the V4-V5 region of the 16s rRNA gene for analysing the microbial diversity in the gut of *G.morhua*.

2.6 Data Processing

The raw data files containing the sequences in FASTA format were downloaded from UCR´s web portal using a software called wget that retrieves large data files from web servers (available at GNU Wget webpage). A mapping file which was compatible with the QIIME2 software, containing information about the samples was created from the original metadata file. The large raw data was processed with the QIIME2 Virtual Box (version 2019.4) and feature table files were created to be used as input for the statistical analysis (Caporaso et. al., 2010; Bolyen et. al., 2018). The commands used for processing data in QIIME2 along with the criterion set at each step is summarized in a text file that is attached as a supplementary material (Appendix A).

After importing the raw sequences in QIIME2, the quality scores were assigned and examined. The subsequent steps involved quality filtering, trimming, merging and removal of chimera. These steps were achieved with a plugin called dada2 (version 1.8) (Callahan et. al., 2016). The process of trimming removed the low quality and the primer sequences (first 35 nucleotides) from all the reads using the option --p-trim-left 35 for forward and reverse reads in combination with a truncating step setting the options to --p-trunc-len-f 285 for forward reads and --p-trunc-len-r 210 for reverse reads. This led to removal of low quality (average Q score < 25) nucleotides from the end of each read. At this step, a feature table containing the amplicon sequence variants (ASVs), their read counts in each sample (ASV table) and a fasta file containing the representative sequences was exported for custom statistical analysis and for classification of the sequences.

2.7 Microbial community analysis

2.7.1 Alpha-diversity analysis

Further analysis was done to find out the alpha-diversity of samples using the ASV table based on rarefaction. The diversity and core-metrics-phylogenetic function plugin in QIIME2 was used for this purpose (plugin available at QIIME2 docs). The settings were decided based on the smallest library size in the dataset (--p-sampling-depth 99181). The deciding factor for picking the sampling depth was the sample with the lowest read count. The rarefaction curves were generated, checked and then exported. The alpha-group significance function (Kruskal & Wallis, 1952) was used to check if there was any significant difference between the samples (measuring diversities within samples). This function was based on the Kruskal-Wallis test and the results for observed OTUs, Shannon diversity and Pielou’s evenness were exported and then plotted using the R package ggplot2 (Wickham, 2009).

2.7.2 Exploratory analysis of read counts

Next was the filtering of the ASV table, considering only ASVs with > 10 reads in at least one sample (1820 out of 3438). This filtered ASV table along with its corresponding metadata was imported into R. The non-metric multidimensional scaling (NMDS) approach, using the metaMDS function in the vegan package was used (Oksanen et. al., 2019) and the results were plotted with the ggplot2 package. Also, the environmental variables (metadata about the fish specimens like weight, length etc.) were tried to fit onto ordination with the envfit function. The vectors of continuous variables are fit using this function. When plotted, the arrows point in the direction of increasing gradient, while the length of the arrow is proportional to the significance of the correlation between the environmental variable and the ordination. Next, the ordisurf function was used to visualize "smooth surfaces" for continuous variables onto the ordination.

2.7.3 Taxonomic analysis

Normalization of the read counts in the ASV table was achieved by converting them into relative abundance values. This was done by dividing the read count of each ASV by the total number of reads in the corresponding sample i.e., library size. The representative sequences were uploaded on the RDP Classifier website and using the default settings of the RDP Naive Bayesian rRNA Classifier Version 2.11 (September 2015), the taxonomy was obtained. The ASV table that contained the relative abundances was then merged with the taxonomical assignments. In the first selection, only ASVs with > 0.01 % abundance in at least one sample were selected while removing the rest from the table. In the next selection, bacterial sequences were selected by retaining only those ASVs that classified into the Domain of Bacteria with > 80 % confidence. This approach was also tried for plotting the phylum and order level compositions and only those ASVs were kept, that had a confidence value > 80 % for the corresponding taxonomic level (phylum or order) respectively. Phyloseq was used to prepare the relative abundance table, taxonomy table and the sample metadata files and after that imported into the R package for plotting (McMurdie & Holmes, 2013).

2.7.4 Differential abundance

The DESeq2 function in R was used with default parameters to find out ASVs that were either enriched or depleted across the 32 samples when oil exposed ones were compared to controls (Love, Huber & Anders, 2014). The read counts of ASVs that were present with > 10 reads in at least one sample was used as the input table. Results were generated according to the calculated adjusted *p*-value (padj) and filtered, to include only ASVs with padj < 0.05.

3 Results

3.1 Sequencing data

The sequencing results produced by dada2 plugin are depicted in table 2. According to this data, about 444,283 ± 242,925 raw reads per sample were generated. After truncation, 368,281±187,282 reads and post removal of sequencing errors, around 367,725±187,120 reads were produced. The quality filtered reads were merged (forward+reverse reads), giving 365,679±186,208 reads that underwent chimera removal and finally provided about 359,912±183,643 reads from each sample.

3438 ASVs, resulted from the processing of data with the shortest and the longest sequence of 250 nt, 413 nt respectively.

Table 2. Sequence processing statistics showing averages (Average) and standard deviations (S.D.) for the samples (n = 32) with respect to the number of initial raw sequencing reads (Raw) and the number of sequences retained after each processing step

	Raw	Filtered	Denoised	Merged	Non-chimeric
Average	444,283	368,281	367,725	365,679	359,912
S.D.	242,925	187,282	187,120	186,208	183,643

3.2 Gut microbe data analysis

3.2.1 Rarefaction curve

The observed OTUs and the sequencing depth were plotted, after rarefaction in QIIME2 (figure 13). The curve was analyzed for alpha-diversity in the samples. It showed that the sequencing depth of 99181 reads/sample was enough for coverage of the diversity of sequences present in all samples as the curves were flattened beyond this point.

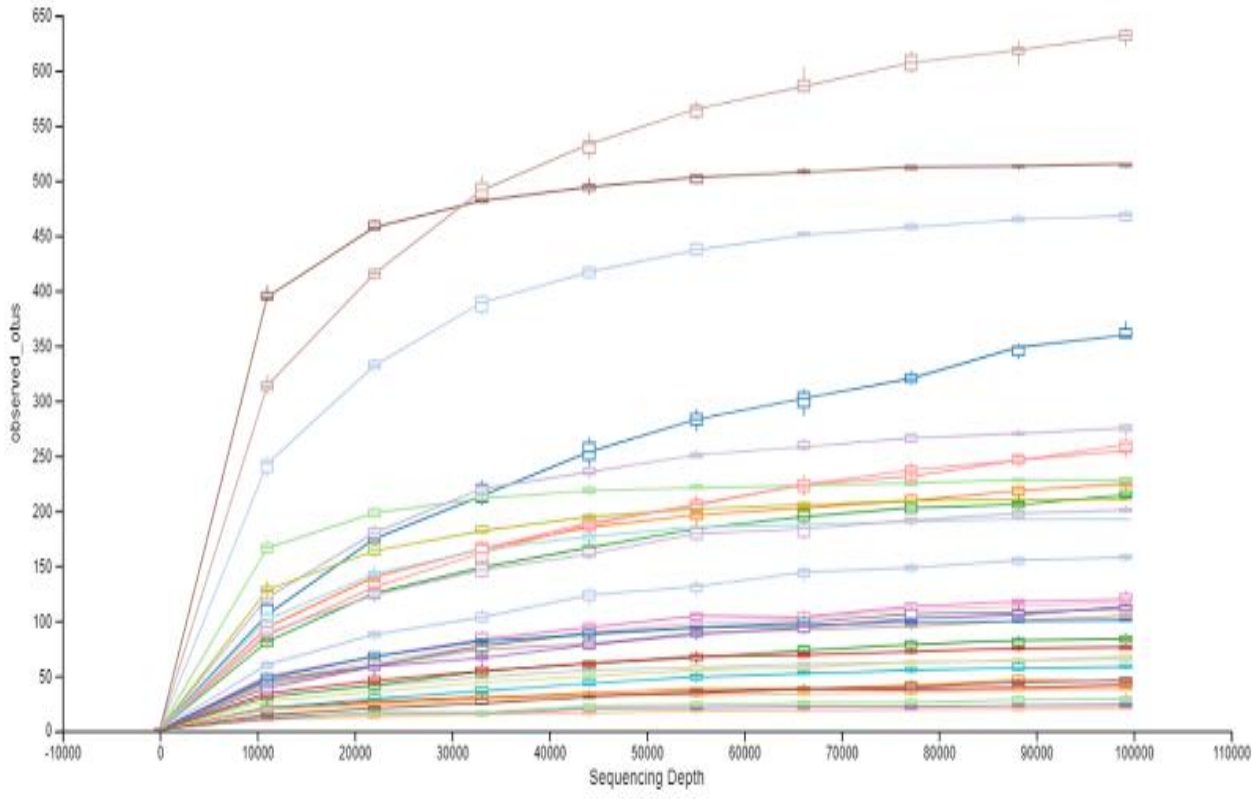

Figure 13. Rarefaction curves of individual samples (n = 32). Number of observed ASVs is plotted against the number of sequences analyzed. Rarefaction depth = 99181 sequences.

3.2.2 NMDS Plot

The NMDS analysis based on Bray-Curtis distance matrix presented a scattered distribution view for the gut microbiota samples. There was no clear grouping of samples for both conditions- control and oil exposed. Even the sampling days did not seem to influence the results (figure 14).

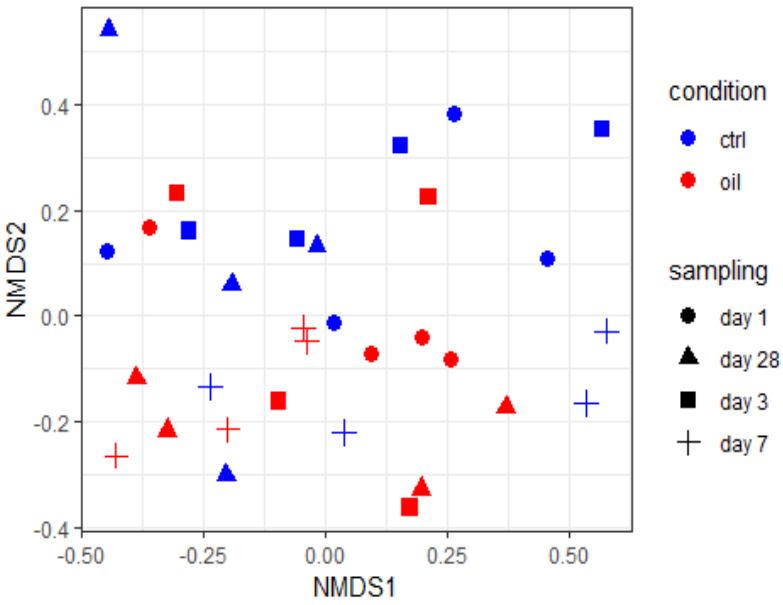

Figure 14. Non-metric multidimensional scaling (NMDS) analysis based on read abundances of ASVs present with > 10 reads in at least one sample using Bray-Curtis distance metrics. Conditions were ctrl = control and oil = oil exposed samples.

The envfit analysis showed that the ASV counts ($p = 0.001$) and Pielou's evenness ($p = 0.036$) had a significant correlation with this ordination and the factors like Fulton's condition index and HSI did not have a significant correlation in this case. Further, the ordisurf function was used to see smooth surfaces based on the ASV read counts of each sample (figure 15).

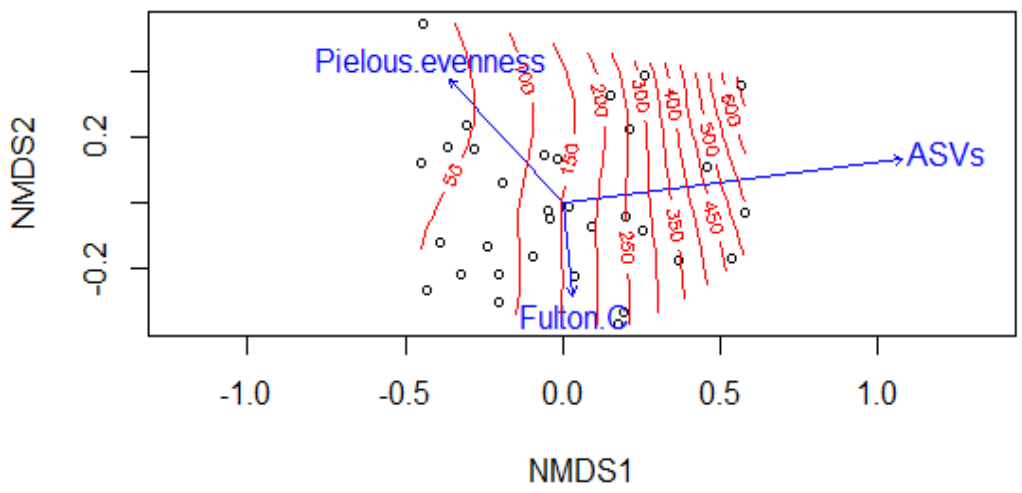

Figure 15. Non-metric multidimensional scaling (NMDS) and envfit analysis based on read abundances of ASVs present with > 10 reads in at least one sample using Bray-Curtis distance metrics. Blue arrows represent the environmental variables that were selected for for correlation analysis. Red lines show smooth surfaces based on ASV read counts in corresponding samples fitted using ordisurf. Note: only Pielou's evenness (Pielous.evenness) and ASV counts (ASVs) has significant correlations ($p < 0.01$).

Factors like weight, length, liver weight, Fulton's condition index and HSI were used for performing another analysis. Based on the ASV counts, the NMDS plot revealed that a sample from day 7 control group was positioned far away from the rest of the samples in the plot. The rest of the samples formed a cluster and were placed close to each other (figure 16).

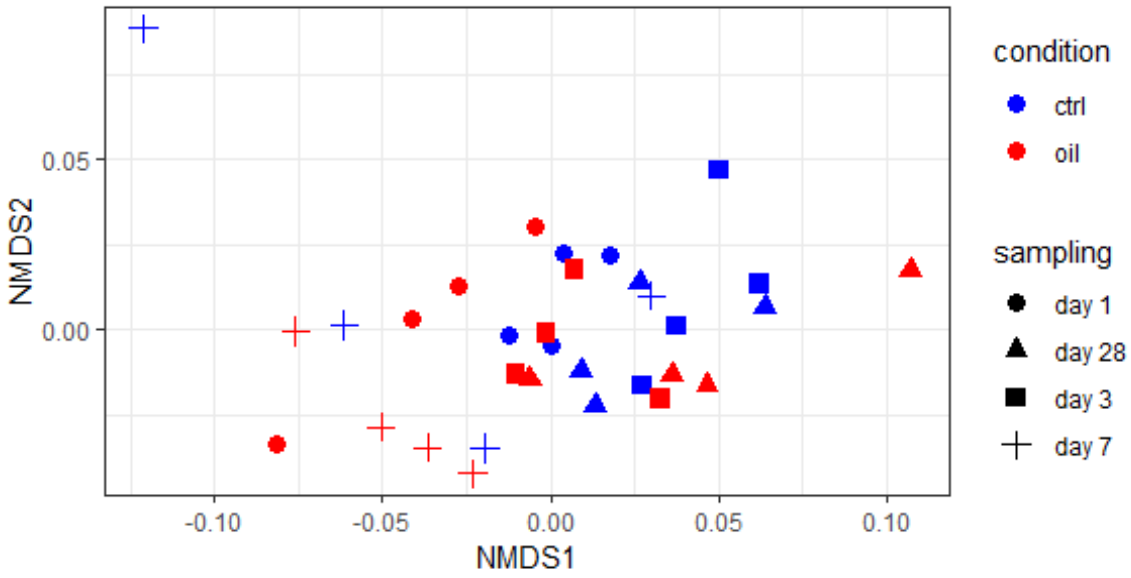

Figure 16. Non-metric multidimensional scaling (NMDS) analysis based on the fish parameters that were collected during the sampling events, i.e., weight, length, liver weight, Fulton's condition index and HSI using Bray-Curtis distance metrics. Conditions were ctrl = control samples and oil = oil exposed samples.

3.2.3 Taxonomy

Taxonomic information and ASV relative abundance were combined and filtered based on confidence values for the classifier. Phylum and order level composition was visualized for ASVs with a taxa confidence of > 80 %. The phylum level plot of Figure 17 shows ASVs with > 0.01 % relative abundance in at least one sample (667 ASVs).

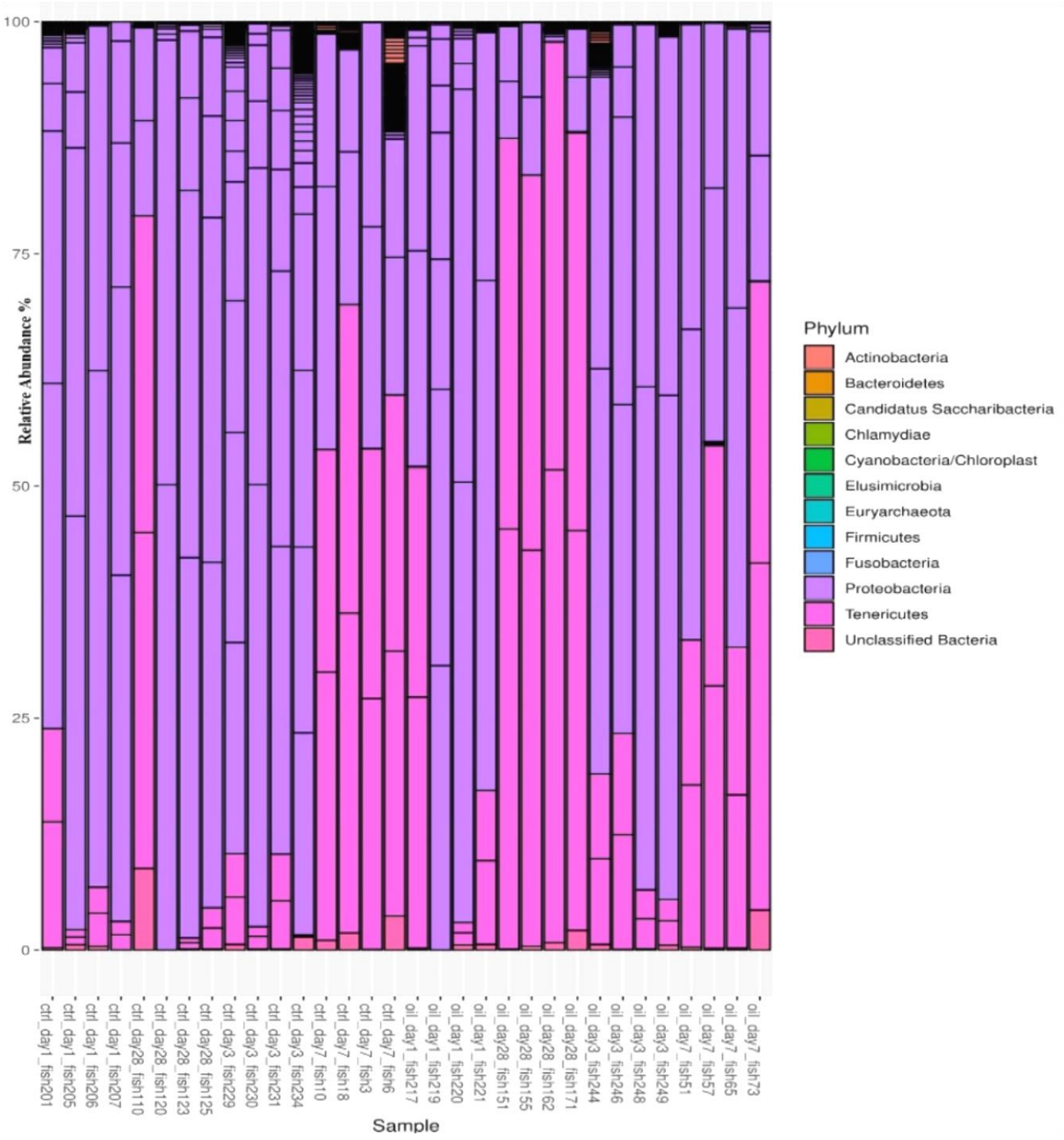

Figure 17. Composition of the 32 intestinal microbiota samples on the phylum level. Relative abundances are plotted for ASVs with > 80 % classification confidence on phylum level and relative abundance > 0.01 % in at least one sample.

The order level lot of Figure 18 was prepared for only more abundant ASVs (> 0.1 % relative abundance in at least one sample, 58 ASVs).

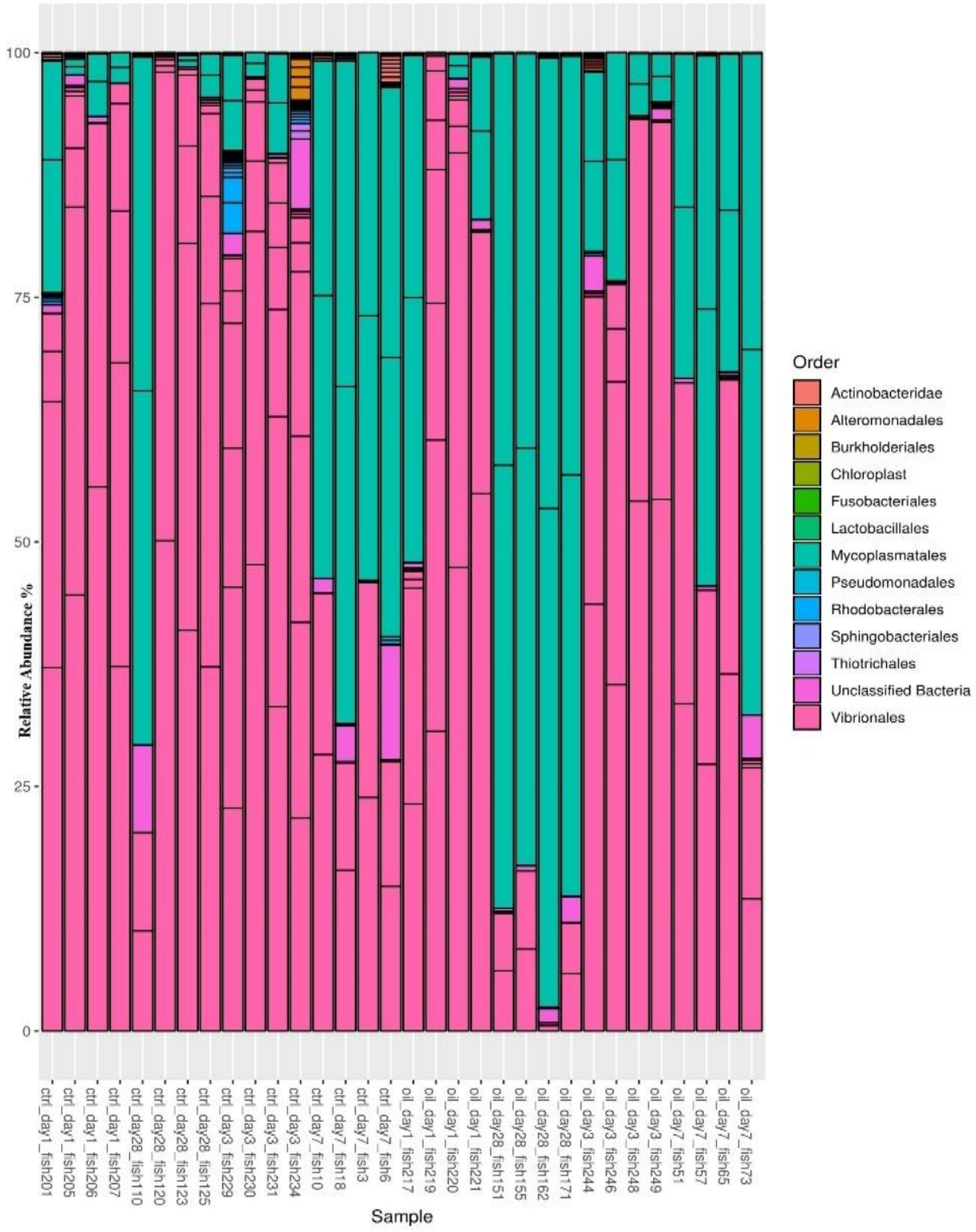

Figure 18. Composition of the 32 intestinal microbiota samples on the order level. Relative abundances are plotted for ASVs with > 80 % classification confidence on order level and relative abundance > 0.1 % in at least one sample.

Figure 19. shows a heatmap that was generated from the same relative abundance table for plotting the order composition. It showed that samples were clustered into 3 main groups, clearly distinguished by the relative abundance of 8 ASVs (the top 8 on the plot).

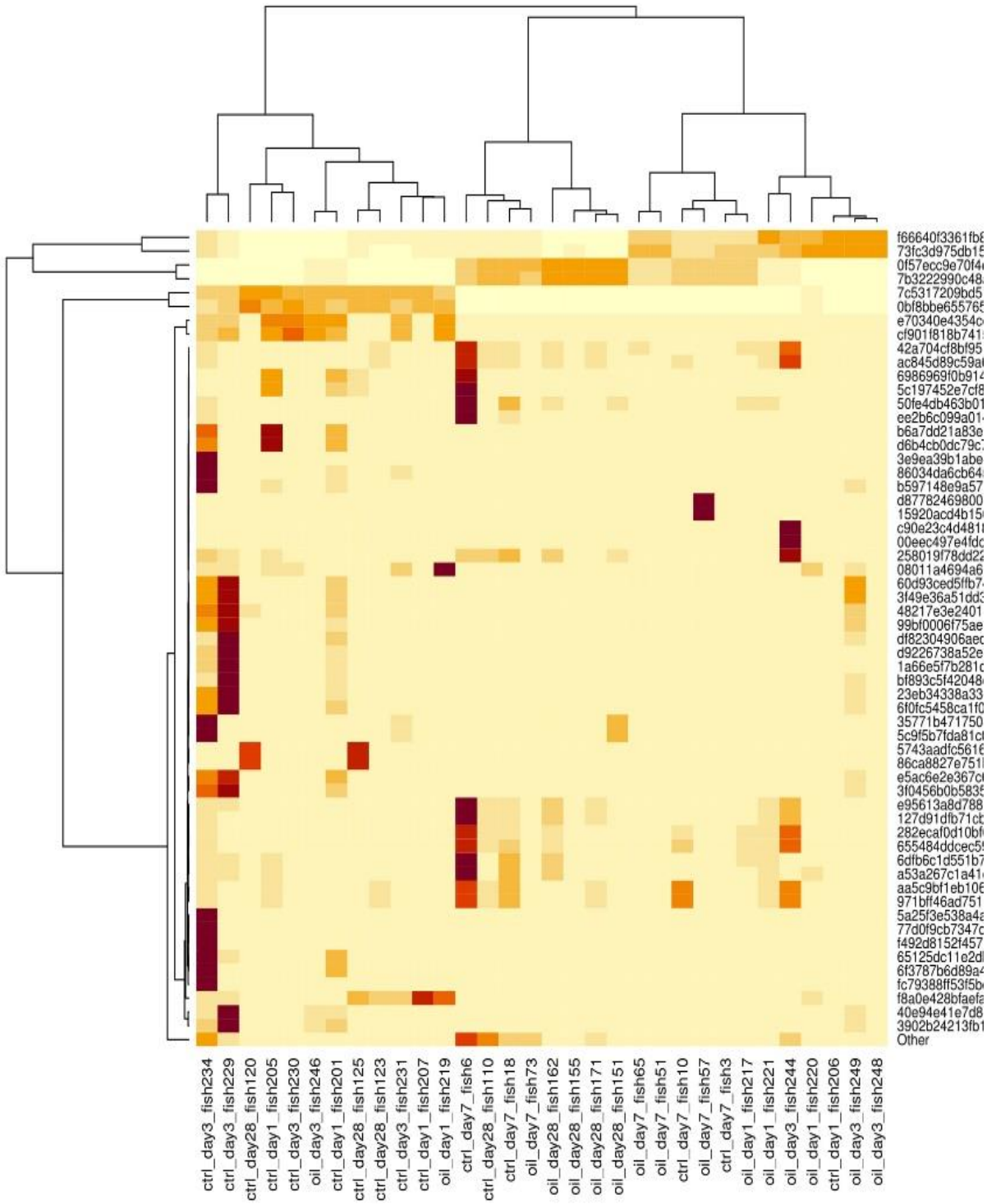

Figure 19. Heatmap of abundant ASVs (> 0.1 % relative abundance in at least one sample) that classified on order level with > 80 % confidence. The darker the color the higher the relative abundance. Sample names are shown on the x-axis while ASV identifiers are shown on the y-axis.

3.3 Upregulated and Downregulated ASVs

DESeq2 analysis identified 65 ASVs that significantly decreased in read count across all sampling times when oil exposed microbiota was compared to control (Table 3).

Table 3. The 65 significantly "downregulated" ASVs across all sampling times when oil exposed samples were compared to controlled samples (referred as controls). Only ASVs with adjusted p-value (padj) < 0.05 are included.

taxon-ASV-ID	baseMean-ReadCount	log2FoldChange	padj
65125dc11e2dbd8f05da5d3167952fa2	71.33	-27.60	0.0000
41216cd038ff3a4e00d44de212ee86a9	51.11	-27.14	0.0000
6f3787b6d89a4c9d16d094fb8cdc51f0	59.40	-26.73	0.0000
e05d90cc0f8e0a41a1780921d324bf77	39.37	-26.45	0.0000
6732545838d67ae6f1018afedafe2eef	29.60	-26.31	0.0000
5743aadfc5616db5db3739ac0ac47b1f	25.12	-26.18	0.0000
86ca8827e751b1ca08dc93dcaf23d5ea	25.91	-25.72	0.0000
d0d8d43a4c2e4d09f79cd5f08c7a2736	9.93	-24.91	0.0000
b33cabe1fe1b3a30b647c7ffe32cb075	9.08	-24.79	0.0000
b6a7dd21a83e35df6ab8bcd8742972ab	3.05	-23.34	0.0000
23eb34338a33d6b98d1f130a7f0e62df	24.31	-11.95	0.0001
e5ac6e2e367c605d6079e73801a05cc7	47.99	-11.57	0.0002
b00e23d299f36c717a12491e75f48539	24.82	-10.82	0.0026
6f0fc5458ca1f078e7e8d7350bd4eb04	26.13	-10.69	0.0007
db20d5dbc8df90dc43480197a2b91461	86.22	-10.17	0.0009
f492d8152f457a501bb03c30a0b5decb	84.39	-10.14	0.0057
1ac9d96271fcc53b0b2bff2ad7720c44	63.29	-9.72	0.0084
cf357d117c57dd881becf9be09e92ac0	45.86	-9.24	0.0140
9c7e723e834c946491ef190ef95fe074	42.73	-9.13	0.0143
9057a9b60005e2abe1aec27f045accc5	39.16	-9.06	0.0134
7734d8630f33ec16e380ebe1055a006f	39.24	-9.00	0.0157
fa316977858a5a9b95db7ffde4448644	39.30	-9.00	0.0157
c2d20db15b2cc10ee2345350b42b65e3	38.64	-9.00	0.0140
43a5ea001023b5079b6b6a8dc148e6d5	38.09	-8.96	0.0157
0f11bbd8d65e182351942cf9af5b2ed3	37.94	-8.95	0.0157
4e883e29a07921f5411b82c1bae41cb8	36.76	-8.90	0.0161
90626c1c457334d327d004ab2a791366	36.06	-8.87	0.0161
1a73c5e5ed7b0d5baa4b081841f6abaa	35.74	-8.86	0.0161
3b37390cfb4c779f8658d579d84a7951	34.67	-8.83	0.0163
64ca23bf0fb0870e83bed4f1a6be842f	33.71	-8.77	0.0165
6f4b077c7b0550de418833d92140f7cf	33.30	-8.75	0.0094
f2dfcfd2fc2fdc0b3c8e80276b614709	32.52	-8.71	0.0169
1565d76b0e6ed8952c2ef72cb515e48b	32.16	-8.69	0.0169
5c9f5b7fda81c021b48696843738a2c2	30.64	-8.63	0.0178
86034da6cb645adf9c9714feeee791b7	29.76	-8.60	0.0179
6986969f0b9147c833d7c923e15ac1fe	147.79	-8.49	0.0100
634217ed820866ed7006c65ac64e1734	28.42	-8.22	0.0252

5c197452e7cf8a95a6d37f504d1e9bba	128.89	-8.13	0.0142
215d5a088544aac86bdde4f8c6fed93b	19.09	-8.05	0.0283
10faed72813f06c31431b7a90510dcac	27.87	-8.04	0.0283
148f270e083b043990dcfa52c1ba7ecb	24.69	-7.93	0.0295
27ac7ee07d97cb490635d0f0e798bfd0	25.67	-7.80	0.0325
d128daeebd61e6e59da20f33c67a1d68	32.29	-7.80	0.0323
fc79388ff53f5bd852a779ba1fbdff7e	14.23	-7.65	0.0373
f757e151cf32ccb1f2876b3f2e5d5917	12.43	-7.47	0.0441
3bf42ab2259ae4f4e91799295a293e90	24.37	-7.38	0.0454
56616a6e1cb6275d4cb1e85b0aaf9221	27.12	-7.37	0.0289
7b5ae8668c4605738c9394026346d8c8	11.49	-7.36	0.0454
9c080f03106c35dd05b01f9014029550	30.08	-7.33	0.0289
2d933c59a572ec4a17cd454ab954b67b	32.47	-7.32	0.0454
bf893c5f42048e97f6378ae94075edcc	11.01	-7.31	0.0469
9a53bc09dbc9adf4aed8feb59fb3ddb9	10.58	-7.25	0.0491
99bf0006f75ae8147a35a5406d6b883d	12.33	-7.23	0.0251
1a66e5f7b281d263009e8bf381d65311	10.21	-7.21	0.0491
273728f4815fe3914d1fccaf6c10089e	10.25	-7.20	0.0491
39eb2998edfc0a51960bb4165d3b754d	34.34	-6.97	0.0289
7c5317209bd515028bee8955ce5e13d4	179849.56	-6.26	0.0165
0bf8bbe65576592209d168ee4178f3fa	179944.16	-6.25	0.0000
cf901f818b7415def3766a76d5821a70	15526.35	-5.71	0.0491
eb1d3607520e1526d5b0da9c54b33b33	44.95	-4.66	0.0263
6dfb6c1d551b7790bf14a961ad57e326	85.19	-4.51	0.0251
31afa51009e3e950c98ef02e6e009fe3	50.37	-4.31	0.0235
28b1256f8ab1093bbb68c759f9a0edba	44.85	-3.84	0.0444
127d91dfb71cb8901d3f6caa185fe727	194.18	-3.23	0.0142
f8a0e428bfaefafc974b83587a8b0be6	1106.06	-3.05	0.0165

From the table 3, it was observed that three ASVs with > 10000 reads were downregulated in samples exposed to oil, that belonged to Vibrionales (Photobacterium).

DESeq2 analysis identified 9 ASVs that significantly increased in read count across all sampling times when oil exposed microbiota was compared to control (Table 4).

Table 4. The 9 significantly "upregulated" ASVs across all sampling times when oil exposed samples were compared to controlled samples (referred as controls). Only ASVs with adjusted p-value (padj) < 0.05 are included.

taxon-ASV-ID	baseMean-ReadCount	log2FoldChange	padj
9fe204e261b6895081def95a2f65bf08	53.76	3.56	0.0444
664fe78c7c4749197b917b0bdcee0893	88.64	4.46	0.0283
21030bfa5893b1bfa3cd378c26ced88d	4.24	21.82	0.0000
34221d69ac967f82a270b09ea67fe947	6.40	21.98	0.0000
beb21d29c9ea790daa2b7ff70152f900	7.47	22.50	0.0000
240010b737211957fc7e53bb1e00d3a9	7.47	22.50	0.0000
04887f955cd4e7caf1b478022ec85115	7.47	22.50	0.0000
f5e0378ef178c791e5187bf09edf712e	7.47	22.50	0.0000

4 Discussion

4.1 Data Processing result analysis

The dada2 processing gave an average of 359,912±187,120 reads from the samples after quality filtration. These were further analyzed to compare the control samples with the exposed samples. The sampling days were also compared to evaluate if it had any influence on the data (figure 20).

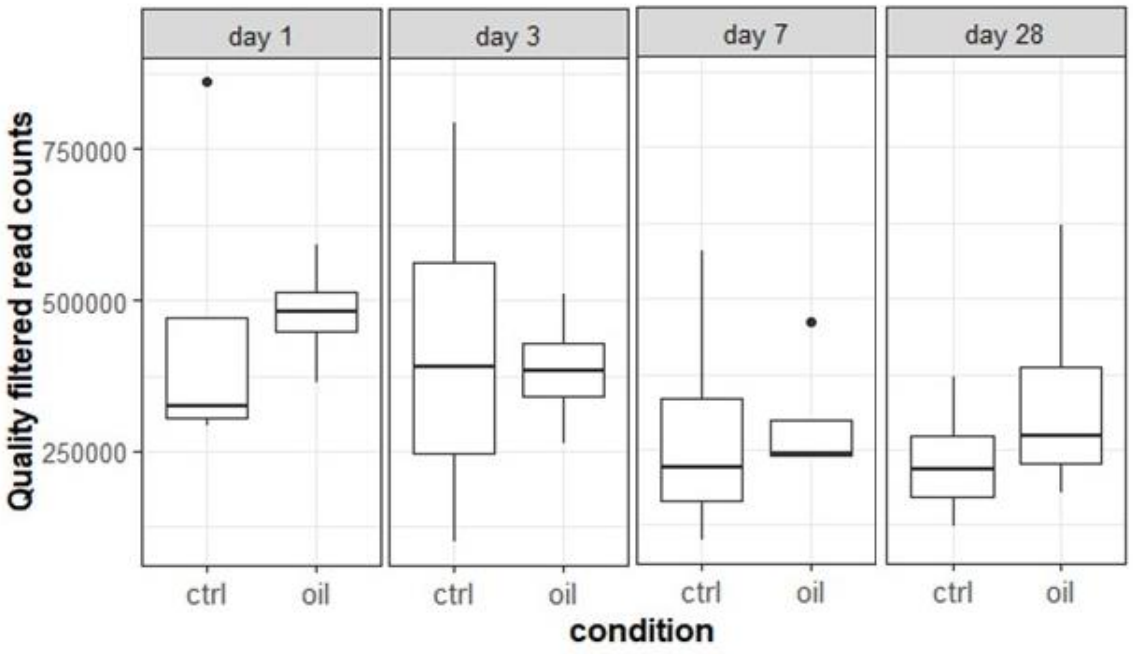

Figure 20. Boxplot representation of the read counts obtained after the dada2 processing step. Plots are grouped by sampling day and abbreviations are as follows: ctrl=control samples, oil= oil exposed samples and black dot represents outlier.

The read counts were found to be variable across all samples and it was observed that overall there was a decrease in the reads from day 1 to day 28 for both control and oil exposed samples. It was also noted that the exposed samples contained more reads as compared to the controls except for day 3 samples, where controls showed more reads than the exposed ones. Two outliers were also present, one from day 1 (control) and the other from day 7 (oil). It is possible to notice some difference between controls and oil exposed fish gut samples; however, they were not statistically significant.

4.2 Richness and Diversity analysis

Pielou's evenness and Shannon index were used to analyze the species richness and diversity in the control and exposed samples. The boxplot representing the ASV counts, OTUs observed and the diversity indices were plotted as shown in the figure 21.

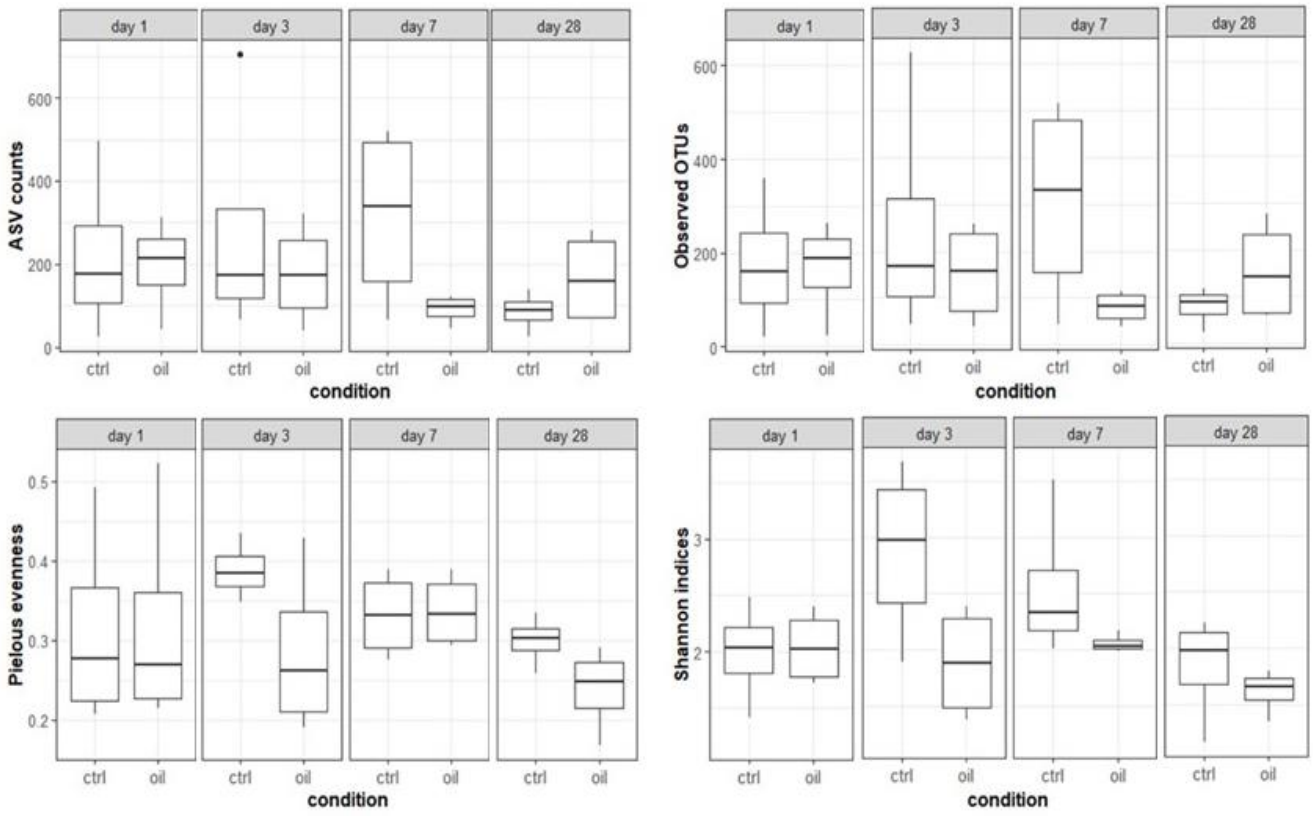

Figure 21. Boxplots summarizing ASV counts (based on dada2 output) and rarefied diversity metrics (sampling depth = 99181) grouped by sampling time (day 1, day 28, day 3 and day 7). Abbreviations are as follows: ctrl = control samples, oil = oil exposed samples and black dot represents outlier.

The following observations were derived from the data reported in figure 2:

- There was very little difference between the ASV counts for days (1 and 3) and the counts were very close in both samples (control and oil exposed fish gut samples). However, for day 7, the control samples showed an increase in the count and a difference was seen when compared with oil exposed samples. At day 28, the opposite was observed.
- A similar trend to ASV counts was seen in case of observed OTUs, controls and oil exposed samples contained almost the same number of OTUs for days (1 and 3). However, more OTUs were observed for control samples in day 7. On the contrary, exposed oil samples contained more OTUs than the control ones at day 28.

- It was also inferred that for days (1 and 7), both controls and oil exposed samples depicted same level of Pielou's evenness. On the other hand, control samples showed much higher evenness than oil treated samples for days (3 and 28). This implied that there was more evenness across species for the control samples.
- The Shannon index was higher for control samples at days (3,7 and 28). This can be related to a bigger diversity in the control samples compared to the oil exposed ones. However, at day 1, both controls and oil exposed samples had similar index values.

4.3 Abundant Orders in the Microbe Community

The order composition for the 29 most abundant ASVs (relative abundance > 0.1% in at least one sample) was investigated (figure 22).

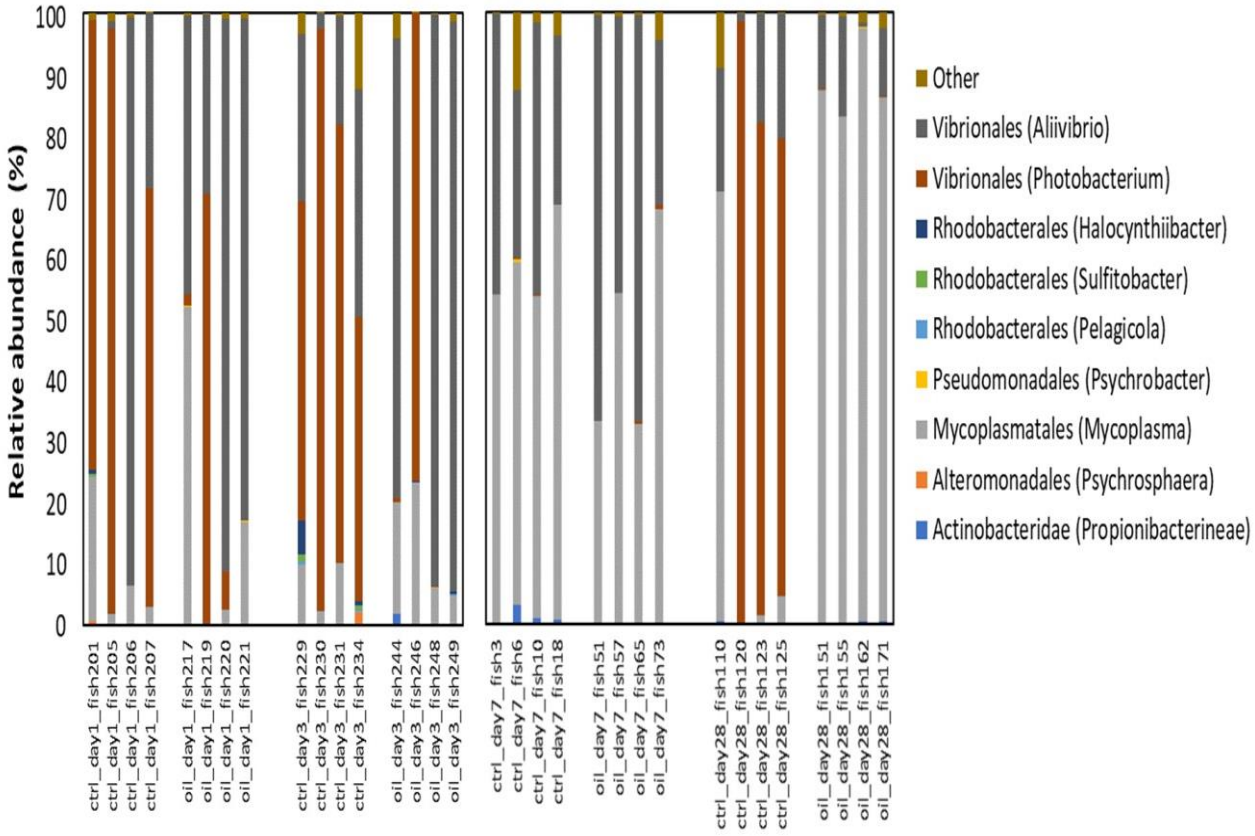

Figure 22. Composition of the 32 intestinal microbiota samples on the order level. Relative abundances are plotted for the 29 most abundant (relative abundance > 0.1 % in at least one sample) ASVs from the first exposure (left) and the second exposure (right). The legend shows the name of each order and genus in parenthesis.

Microbes of the order Vibrionales, Mycoplasmatales, Actinobacteridae, Alteromonadales, Rhodobacterales and Pseudomonadales were present in significant proportions in the Atlantic cod gut. Out of these orders, one order that was clearly dominating the whole community was Vibrionales. This finding is in agreement with the previous studies (Star et al., 2013; Bagi et al., 2018; Riiser et al., 2018 and 2019). When looked at the genus level, the following three genera Photobacterium, Aliivibrio and Mycoplasma dominated the dataset. This could have been the result of using juvenile cod with a less developed microbiota. Riiser et al., 2018 in their investigation of intestinal microbiome of cod, also reported that 78% of all Vibrionales

reads were represented by Photobacterium. A similar finding was documented in Bagi et al. (2018), where more than 70% of Vibrionales belonged to photobacterium genus.

As discussed above, earlier studies on Atlantic cod have shown presence of Vibrionales in high abundance, however orders of Bacteriodales, Clostridiales, Alteromonadales, Fusobacteriales and Desulfovibrionales were reported as well from those studies. The results from the present thesis work also show the presence of these orders; even if they were present in a relatively very low abundance when compared to previous studies.

It was also interesting to note that Bagi et al. (2018) recorded an order called Deferribacterales that had a high relative abundance in samples exposed to high concentration of oil. Star et al., 2013 also reported variable amount of this order in their samples, that were not given any oil treatment. However, according to the findings of present work, this bacterial order was neither found in control nor in oil samples.

The most abundant orders in the samples were further analyzed for each condition and sampling day (figure 23).

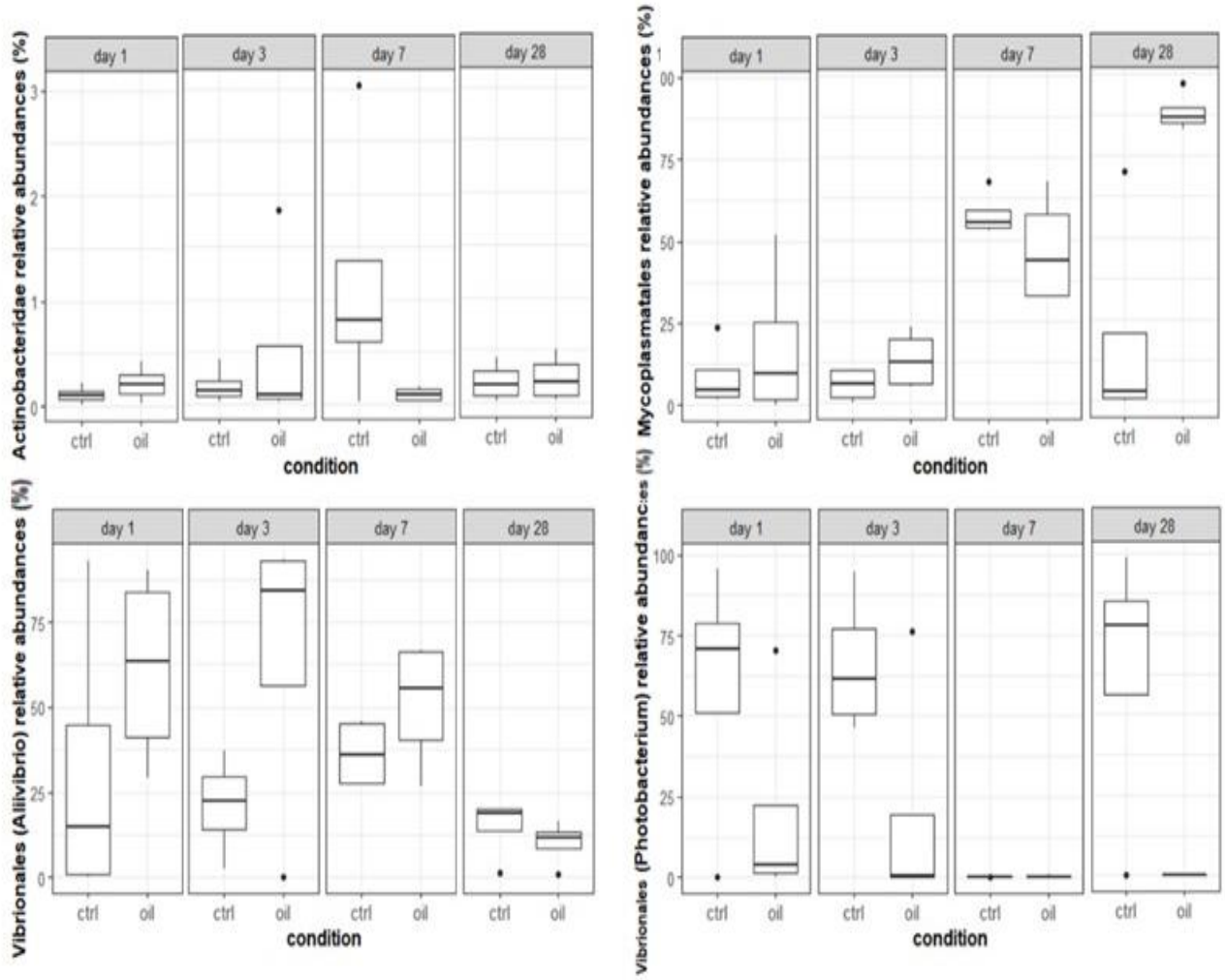

Figure 23. Boxplot representation of selected abundant orders, grouped by sampling time (day 1, day 28, day 3 and day 7). Abbreviations are as follows: ctrl = control samples, oil = oil exposed samples and black dot represents outlier.

The attained results trigger the following discussion:

- The relative abundance of Actinobacteridae was overall very low in both control and oil exposed samples, but it was slightly present more after 1 day of exposure in the oil treated samples. Samples at day 3 did not show any difference in the abundance levels. Control samples were more abundant in the Actinobacteridae order after 7 days. While at day 28, there was no variability in the abundance levels between control and oil exposed samples.
- Mycoplasmatales were more abundant in oil exposed samples for day 1 and 3, and highly abundant for day 28. In contrast, for day 7, control samples displayed more abundance than oil exposed ones.
- The relative abundance of Vibrionales (genus- Aliivibrio) was very high in the oil exposed samples after 1, 3 and 7 days. Although at day 28, control samples were found to have more abundance than oil exposed ones.

- For Vibrionales (Photobacterium), large abundance was seen in control samples after 1, 3 and 28 days. While at day 7, this order seemed to be absent in both control and oil exposed samples. This absence was also noted in oil exposed samples at day 28.

5 Conclusions

This research work aimed at addressing the question: Can gut microbial community changes in Atlantic cod be used as biomarker of water pollution? The findings from the study lead to an affirmative answer.

The experimental work was systematically planned to mimic a marine polluted environment where cod samples were exposed to dispersed crude oil. Control samples with no oil treatment were used to draw comparisons with oil treated samples. A method was optimized for DNA extraction of the fish gut and after ensuring the quality of DNA, the samples were sequenced. The results from 32 samples, including both control and oil treated samples were further processed with advanced bioinformatic tools. Quality control of sequences, identification of amplicon sequence variants and taxonomic analysis were carried out.

The non-metric multidimensional scaling plots displayed that the samples were scattered in space with no clear clustering of the groups. However, the heatmap gave more insights into the order composition of the samples and boxplot representation of these orders verified the pattern seen in the heatmap. It was observed that orders Vibrionales (genus Aliivibrio) and Mycoplasmatales were represented more in oil treated samples and order Vibrionales (genus Photobacterium) were comparatively present more in control samples. Also, Vibrionales (genus Photobacterium) were found to be downregulated in oil treated samples across all sampling days. The values for diversity metrics also indicated that the microbial diversity was reduced in oil treated samples. When comparing the obtained results with previous studies, it was confirmed that Vibrionales dominated the gut in Atlantic cod. Based on the outcome of this thesis study, Vibrionales (genus Aliivibrio) and Mycoplasmatales bacterial orders could be used as biomarker of water pollution in Atlantic cod.

6 Scope of further study

Since the 16s rRNA sequencing has limitation to perform species-level classification, further research studies could be focused that task.

Alternative methods to explore diversity changes, for example, whole genome sequencing can provide the species or strain level classification. It has significant potential to provide additional details on genes other than 16s. This may reveal more information regarding the microbes activities related to the dispersed crude oil exposure.

Omics studies could also be applied to analyze mRNA transcripts or gene products and the presence of various metabolites. The field of genome sequencing has tremendous scope for analysis of microbial changes in Atlantic cod gut samples.

7 References

1. Røe Utvik, T.I., Durell, G.S., Johnsen, S., 1999. Determining produced water originating polycyclic aromatic hydrocarbons in North Sea waters: comparison of sampling techniques. Mar. Poll. Bull. 38 (11), 977–989.

2. Latimer JS, Zheng J (2003) The sources, transport and fate of PAHs in the marine environment. In: Doube PET editor. PAHs: an ecotoxicological perspective. Wiley, West Sussex, pp. 9-34.

3. Pampanin, D.M., and Sydnes M.O. (2013). Polycyclic aromatic hydrocarbons a constituent of petroleum: presence and influence in the aquatic environment. In: Vladimir K, Kolesnikov A, (Ed.), Hydrocarbon. Rijeka: InTech; doi:10.5772/48176.

4. Sanni, S., Lyng, E., & Pampanin, D. M. (2017). III: Use of biomarkers as Risk Indicators in Environmental Risk Assessment of oil based discharges offshore. Mar Environ Res, 127, 1-10. doi:10.1016/j.marenvres.2016.12.004.

5. Peakall, D.B. & Walker, C.H. Ecotoxicology (1994) 3: 173. doi:10.1007/BF00117082.

6. Depledge, M.H., Fossi, M.C., 1994. The role of biomarkers in environmental assessment (2). Invertebrates. Ecotoxicology 3 (3), 161e172.

7. Fossi, C. and Leonzio, C. (1993). Nondestructive biomarkers in Vertebrates. Boca Raton, FL:Lewis.

8. Viarengo, A., Lowe, D., Bolognesi, C., Fabbri, E., & Koehler, A. (2007). The use of biomarkers in biomonitoring: a 2-tier approach assessing the level of pollutant-induced stress syndrome in sentinel organisms. *Comp Biochem Physiol C Toxicol Pharmacol, 146*(3), 281-300. doi:10.1016/j.cbpc.2007.04.011.

9. Nahrgang, J., Brooks, S. J., Evenset, A., Camus, L., Jonsson, M., Smith, T. J., Lukina, J., Frantzen, M., Giarratano, E., Renaud, P. E. (2013). Seasonal variation in biomarkers in blue mussel (Mytilus edulis), Icelandic scallop (Chlamys islandica) and Atlantic cod (Gadus morhua): implications for environmental monitoring in the Barents Sea. *Aquat Toxicol, 127*, 21-35. doi:10.1016/j.aquatox.2012.01.009.

10. Pampanin, D. M., Le Goff, J., Skogland, K., Marcucci, C. R., Oysaed, K. B., Lorentzen, M., Jorgensen, K.B., Sydnes, M. O. (2016). Biological effects of polycyclic aromatic hydrocarbons (PAH) and their first metabolic products in in vivo exposed Atlantic cod (Gadus morhua). *J Toxicol Environ Health A, 79*(13-15), 633-646. doi:10.1080/15287394.2016.1171993.

11. Sundt, R. C., Ruus, A., Jonsson, H., Skarphéðinsdóttir, H., Meier, S., Grung, M., Beyer, J., Pampanin, D. M. (2012). Biomarker responses in Atlantic cod (Gadus morhua) exposed to produced water from a North Sea oil field: Laboratory and field assessments. *Marine Pollution Bulletin, 64*(1), 144-152. doi:10.1016/j.marpolbul.2011.10.005.

12. Bagi, A., Riiser, E. S., Molland, H. S., Star, B., Haverkamp, T. H. A., Sydnes, M. O., & Pampanin, D. M. (2018). Gastrointestinal microbial community changes in Atlantic

cod (Gadus morhua) exposed to crude oil. *BMC Microbiol*, *18*(1), 25. doi:10.1186/s12866-018-1171-2.

13. Ghanbari, M., Kneifel, W., and Domig, K.J. (2015). A new view of the fish gut microbiome: Advances from next-generation sequencing. *Aquaculture 448*, 464-475. doi:10.1016/j.aquaculture.2015.06.033.

14. Johny, T. K., Saidumohamed, B. E., Sasidharan, R. S., & Bhat, S. G. (2018). Inferences of gut bacterial diversity from next-generation sequencing of 16S rDNA in deep sea blind ray - Benthobatis moresbyi. *Ecological Genetics and Genomics*, *9*, 1-6. doi:10.1016/j.egg.2018.07.001.

15. Kumar, G., & Kocour, M. (2017). Applications of next-generation sequencing in fisheries research: A review. *Fisheries Research, 186 (1)*, 11-22. doi:10.1016/j.fishres.2016.07.021.

16. Behjati, S., & Tarpey, P. S. (2013). What is next generation sequencing? *Arch Dis Child Educ Pract Ed*, *98*(6), 236-238. doi:10.1136/archdischild-2013-304340.

17. Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends Genet*, *24*(3), 133-141. doi:10.1016/j.tig.2007.12.007.

18. Van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends Genet*, *30*(9), 418-426. doi:10.1016/j.tig.2014.07.001.

19. Gargis, A. S., Kalman, L., & Lubin, I. M. (2016). Assuring the Quality of Next-Generation Sequencing in Clinical Microbiology and Public Health Laboratories. J Clin Microbiol, 54(12), 2857-2865. doi:10.1128/JCM.00949-16.

20. Ambardar, S., Gupta, R., Trakroo, D., Lal, R., & Vakhlu, J. (2016). High Throughput Sequencing: An Overview of Sequencing Chemistry. Indian J Microbiol, 56(4), 394-404. doi:10.1007/s12088-016-0606-4.

21. Loman, N. J., Constantinidou, C., Chan, J. Z. M., Halachev, M., Sergeant, M., Penn, C. W., Robinson, E.R., Pallen, M. J. (2012). High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nature Reviews Microbiology, 10*, 599-606. doi:10.1038/nrmicro2850.

22. Shokralla, S., Spall, J. L., Gibson, J. F., & Hajibabaei, M. (2012). Next-generation sequencing technologies for environmental DNA research. *Mol Ecol*, *21*(8), 1794-1805. doi:10.1111/j.1365-294X.2012.05538.x.

23. Breitwieser, F. P., Lu, J., & Salzberg, S. L. (2017). A review of methods and databases for metagenomic classification and assembly. *Brief Bioinform*. doi:10.1093/bib/bbx120.

24. Marchesi, J. R., & Ravel, J. (2015). The vocabulary of microbiome research: a proposal. *Microbiome*, *3*, 31. doi:10.1186/s40168-015-0094-5.

25. Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis, The ISME Journal 11, 2639–2643. doi.org/10.1038/ismej.2017.119

26. Sanni, S., Øysæd, K. B., Høivangli, V., & Gaudebert, B. (1998). A continuous flow system (CFS) for chronic exposure of aquatic organisms. *Marine Environmental Research*, *46*(1-5), 97-101. doi:10.1016/S0141-1136(97)00086-X.

27. Lambert, Y., & Dutil, J.-D. (1997). Can simple condition indices be used to monitor and quantify seasonal changes in the energy reserves of cod (Gadus morhua)? *Canadian Journal of Fisheries and Aquatic Sciences, 54*, 104-112. doi:10.1139/f96-149.

28. Han, Z., Sun, J., Lv, A., Sung, Y., Sun, X., Shi, Hu, X.,Wang, A., Xing, K. (2018). A modified method for genomic DNA extraction from the fish intestinal microflora. *AMB Express*, *8*(1), 52. doi:10.1186/s13568-018-0578-3.

29. AllPrep PowerFecal DNA/RNA Kit (2019, May 13). Retrieved from https://www.qiagen.com/no/shop/sample-technologies/combined-sample-technologies/preparation/allprep-powerfecal-dna-rna-kit/#productdetails.

30. Monarch DNA Gel Extraction Kit (2019, May 13). Retrieved from https://www.neb.com//media/catalog/Long%20Description/T1020_GelExtractionKitWorkflow.png.

31. QIAmp Fast DNA Stool Mini Kit (2019, May 13).Retrieved from https://www.qiagen.com/no/shop/sample-technologies/dna/genomic-dna/qiaamp-fast-dna-stool-mini-kit/#productdetails.

32. GNU Wget. Introduction to GNU Wget (2019, May 29). Retrieved from https://www.gnu.org/software/wget/.

33. Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I.,Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J., Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. Nature Methods, 7, 335. doi:10.1038/nmeth.f.303.

34. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet C, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Titus Brown C, Callahan BJ, Caraballo Rodríguez AM, Chase J, Cope E, Da Silva R, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley G, Janssen S, Jarmusch AK, Jiang L, Kaehler B, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, et al. (2018). QIIME 2: reproducible, interactive, scalable, and extensible microbiome data science. PeerJ PrePrints 6:e27295v2. https:// doi.org/10.7287peerj.preprints/27295v2/.

35. Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods, 13, 581. doi:10.1038/nmeth.3869.

36. QIIME 2 docs. Core-metrics-phylogenetic: core diversity metrics (phylogenetic and non-phylogenetic) (2019, May 29). Retrieved from https://docs.qiime2.org/2019.4/plugins/available/diversity/core-metrics-phylogenetic/.

37. Kruskal, W. H., & Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. Journal of the American Statistical Association, 47(260), 583-621. doi:10.1080/01621459.1952.10483441.

38. Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. doi:10.1007/978-0-387-98141-3.

39. Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs, E., Wagner, H. (2019). Community Ecology Package. https://cran.r-project.org/web/packages/vegan/vegan.pdf.

40. McMurdie, P. J., & Holmes, S. (2013). phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS One*, *8*(4), e61217. doi:10.1371/journal.pone.0061217.

41. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, *15*(12), 550. doi:10.1186/s13059-014-0550-8.

42. Star, B., Haverkamp, T. H., Jentoft, S., & Jakobsen, K. S. J. B. M. (2013). Next generation sequencing shows high variation of the intestinal microbial species composition in Atlantic cod caught at a single location. *13*(1), 248. doi:10.1186/1471-2180-13-248.

43. Riiser, E. S., Haverkamp, T. H. A., Varadharajan, S., Borgan, O., Jakobsen, K. S., Jentoft, S., & Star, B. (2019). doi:10.1101/545889.

8 Abbreviations

ASV : Amplicon sequence variants

CAT : Catalase

CFS : Continuous flow system

CI : Condition Index

CTAB : Hexadecyl trimethyl ammonium bromide

EROD : Ethoxyresorufin O-deethylase

GST : Glutathione S-transferases

HSI : Hepatosomatic Index

ITS : Internal Transcribed Spacer

NGS : Next generation sequencing

NMDS : Non-metric multidimensional scaling

OTU : Operational Taxonomic Units

PW : Produced Water

PAH : Polycyclic Aromatic Hydrocarbons

9 Appendix

QIIME2 commands for processing the sequencing data:

installing QIIME2 according to: https://docs.qiime2.org/2019.4/install/virtual/virtualbox/

downloading the raw data for 32 x 2 files from UCR

wget-r--no-parent
http://illumina.bioinfo.ucr.edu/illumina_runs/1102/190514_M02457_0374_000000000-CGGVK

##place sequences (only the sequences) in the shared folder

media/sf_Desktop/data/

preparing metadata file in an Excel sheet and savind as tab separated file

media/sf_Desktop/sample-metadata.tsv

check where you are

$ pwd

/media/sf_Desktop

##################################

Initial data processing steps

##################################

import sequences into a qiime artifact

qiime tools import --type 'SampleData[PairedEndSequencesWithQuality]' --input-path data --input-format CasavaOneEightSingleLanePerSampleDirFmt --output-path demux.qza

summarize the results of the import step

qiime demux summarize --i-data demux.qza --o-visualization demux.qzv

view the summary info about the sequences

qiime tools view demux.qzv

##

Filtering-trimming-denoising-merging sequences using dada2

##

first attempt - based on quality score based filtering q = 20

this retained 1% of the original reads and produced 94 ASVs

qiime dada2 denoise-paired --i-demultiplexed-seqs demux.qza --p-trim-left-f 35 --p-trim-left-r 35 --p-trunc-q 20 --p-trunc-len-f 300 --p-trunc-len-r 300 --o-table table.qza --o-representative-sequences rep-seqs.qza --o-denoising-stats denoising-stats.qza

second attempt based on truncating to a minimum viable length

this retained large portion of the raw reads and produced 3433 ASVs

qiime dada2 denoise-paired --i-demultiplexed-seqs demux.qza --p-trim-left-f 35 --p-trim-left-r 35 --p-trunc-len-f 285 --p-trunc-len-r 250 --o-table table.qza --o-representative-sequences rep-seqs.qza --o-denoising-stats denoising-stats.qza

third attempt based on truncating to a shorter length (425 bp without overlap)

default overlap in dada2 is 12 bases - could try --p-trunc-len-r 225 next

this worked well!

qiime dada2 denoise-paired --i-demultiplexed-seqs demux.qza --p-trim-left-f 35 --p-trim-left-r 35 --p-trunc-len-f 285 --p-trunc-len-r 210 --o-table table.qza --o-representative-sequences rep-seqs.qza --o-denoising-stats denoising-stats.qza

summarize denoising statistics and create visualiztion

qiime metadata tabulate --m-input-file denoising-stats.qza --o-visualization denoising-stats.qzv

##

Processed data - ASV table and representative sequences

##

summarizing and visualizing feature table and representative sequences

a lot of analysis can be done on these once exported

also a lot of files can be saved from the visualization website

qiime feature-table summarize --i-table table.qza --o-visualization table.qzv --m-sample-metadata-file sample-metadata.tsv

qiime feature-table tabulate-seqs --i-data rep-seqs.qza --o-visualization rep-seqs.qzv

exporting the raw results

it gives an error but still exports the files

qiime tools export --input-path table.qza --output-path exported-table

qiime tools export --input-path rep-seqs.qza --output-path exported-rep-seqs

convert the exported .biom file into something that we can work with

biom convert -i exported-table/feature-table.biom -o ASVtable.txt --to-tsv

############################

Some diversity analysis

############################

creating phylogenetic trees of the unique sequences for downstream analysis

qiime phylogeny align-to-tree-mafft-fasttree --i-sequences rep-seqs.qza --o-alignment aligned-rep-seqs.qza --o-masked-alignment masked-aligned-rep-seqs.qza --o-tree unrooted-tree.qza --o-rooted-tree rooted-tree.qza

computing beta diversity metrics based on 99181 randomly selected sequences from each sample

qiime diversity core-metrics-phylogenetic --i-phylogeny rooted-tree.qza --i-table table.qza --p-sampling-depth 99181 --m-metadata-file sample-metadata-ver2.tsv --output-dir core-metrics-results

testing for associations between categorical metadata columns and alpha diversity data

qiime diversity alpha-group-significance --i-alpha-diversity core-metrics-results/faith_pd_vector.qza --m-metadata-file sample-metadata-ver2.tsv --o-visualization core-metrics-results/faith-pd-group-significance.qzv

qiime diversity alpha-group-significance --i-alpha-diversity core-metrics-results/shannon_vector.qza --m-metadata-file sample-metadata-ver2.tsv --o-visualization core-metrics-results/shannon-group-significance.qzv

qiime diversity alpha-group-significance --i-alpha-diversity core-metrics-results/observed_otus_vector.qza --m-metadata-file sample-metadata-ver2.tsv --o-visualization core-metrics-results/observed-otus-group-significance.qzv

qiime diversity alpha-group-significance --i-alpha-diversity core-metrics-results/shannon_vector.qza --m-metadata-file sample-metadata-ver2.tsv --o-visualization core-metrics-results/shannon-group-significance.qzv

qiime diversity alpha-group-significance --i-alpha-diversity core-metrics-results/evenness_vector.qza --m-metadata-file sample-metadata-ver2.tsv --o-visualization core-metrics-results/evenness-group-significance.qzv

since we were working with rarefied data - let's plot the rarefaction curve

qiime diversity alpha-rarefaction --i-table table.qza --i-phylogeny rooted-tree.qza --p-max-depth 99181 --m-metadata-file sample-metadata-ver2.tsv --o-visualization alpha-rarefaction.qzv

##

Classifying sequences using scikitlearn and Silva database

##

NOTE: due to plugin error - this was not completed!

downloaded the full length trained Silva database ver 132

from: https://docs.qiime2.org/2018.11/data-resources/

moved the silva-132-99-nb-classifier.qza into /media/sf_Desktop folder

executed the following command to classify the 3433 ASVs

executed the following command according to the instructions on the above website

conda install --override-channels -c defaults scikit-learn=0.19.1

this did not work - it removed the feature-classifier plugin which I could not reinstall

downloaded a QIIME recommended Silva-132 version archive database

selected the representative sequences for99% clustered OTUs

selected the taxonomy file for all levels consensus

importing these two files

qiime tools import --type 'FeatureData[Sequence]' --input-path silva_132_99_16S.fna --output-path silva_132_99_otus.qza

qiime tools import --type 'FeatureData[Taxonomy]' --input-format HeaderlessTSVTaxonomyFormat --input-path consensus_taxonomy_all_levels.txt --output-path ref-taxonomy.qza

without trimming as recommended training the classifier

qiime feature-classifier fit-classifier-naive-bayes --i-reference-reads silva_132_99_otus.qza --i-reference-taxonomy ref-taxonomy.qza --o-classifier silva-full-length-classifier.qza