

# **The Plastisphere- difference between fresh and marine water, metadata analysis**

Keda Li  
2498449L

Supervised by  
Dr. Umer Zeeshan Ijaz

Glasgow  
August 2020

A thesis submitted in partial fulfilment of the requirements  
for the degree of  
**MASTER OF SCIENCE IN MECHANICAL ENGINEERING**

## ***Abstract***

The plastisphere which is the microplastic distributed with developing microbial community has a significant impact on the whole aquatic system, recently, its high abundance has been raising environmental attention. Previous studies focused on studying the problem of the plastisphere in the marine environment, while as one important component of the global microplastic life cycle, the issue of the Plastisphere in freshwater does not seem to be widely studied. In this work, the roles of geographic location that in shaping the microbial composition and various type of microplastics will be investigated, besides determining the main factors of distinct environmental conditions that alter the microbial composition structure of microplastic, at last, the ecological influences and microplastic degradation pathways for specific genera dominant in plastisphere finding will be explored. The Metadata Analysis of four studies was applied by utilizing high throughput 16S rRNA V4 gene sequencing throughout DADA2. The Alpha/Beta diversity analysis demonstrated the sample location and sample type were the key factors structuring bacterial community, and environmental conditions (salinity, pH, Temperature) investigated with comparing NTI/NRI and Alpha. Major vectors for bacterial assemblages of *Pseudomonadales* and *Altermonadales* were abundant for the PS and PE associated biofilm in seawater, while the *Flavobacterium* was abundant in the freshwater. These findings elucidate the environmental condition principally shapes the bacteria community, which could play a role in dealing with microplastic pollution in a distinct aquatic environment in the future.



## *Acknowledgement*

I would like to express my gratitude to Dr. Umer Zeeshan Ijaz and Dr. Ciara Keating, for providing me the opportunity to study this project. I am truly appreciated Dr. Umer for all of time on meeting at every week, you indicate clearly the points that I should focus on in my project and give me the advices on project panning. A sincerely thanks to Dr. Ciara Keating for the teaching, you dedicate lots of time on support my data analysis. Dr Ciara Keating always reply the messages on Microsoft Teams as soon as possible, provide the guidance of software application, and teaching me the tutorial of processing data (<http://www.tinyurl.com/JCBioinformatics3>) with practical demonstrations, she also share the related literatures to me to broaden my knowledge in this subject.

I would also like to thanks to my teammate Bozhen Chen and Hongyu Mu who working on the same project as me. I always give me help me when I'm confused.

## ***Table of contents***

|                                                                                       |                      |          |
|---------------------------------------------------------------------------------------|----------------------|----------|
| <b>MASTER OF SCIENCE IN MECHANICAL ENGINEERING</b>                                    | <b>Abstract.....</b> | <b>1</b> |
| <b>Acknowledgement .....</b>                                                          | <b>3</b>             |          |
| <b>Table of contents .....</b>                                                        | <b>4</b>             |          |
| <b>1 Introduction.....</b>                                                            | <b>6</b>             |          |
| 1.1 Background of Plastisphere. ....                                                  | 6                    |          |
| 1.2 Risks of Aquatic Ecosystems Microplastic Debris .....                             | 7                    |          |
| 1.3 Analysis of Microplastic Transportation Among Freshwater and Marine Ecosystems .. | 7                    |          |
| 1.4 Analysis of the Issue of Freshwater and Marine Plastic Pollution .....            | 8                    |          |
| 1.5 Comparison of Microplastics in Freshwater and Marine Environments .....           | 9                    |          |
| 1.5 Next-generation sequencing of 16S microbiome data for Plastisphere species. ....  | 10                   |          |
| 1.6 Meta-Data Analysis Overview.....                                                  | 11                   |          |
| 1.7 Existing Gaps in Knowledge .....                                                  | 11                   |          |
| 1.8 Aims of the Study .....                                                           | 11                   |          |
| <b>2 Method .....</b>                                                                 | <b>13</b>            |          |
| 2.1 Background of studies.....                                                        | 13                   |          |
| 2.2 Sequencing Analysis .....                                                         | 14                   |          |
| 2.2.1 FASTQ collection .....                                                          | 14                   |          |
| 2.2.2 Data Download and File Organization. ....                                       | 14                   |          |
| 2.2.3 Analysis using QIME2.....                                                       | 15                   |          |
| 2.2.4 DADA2 Analysis.....                                                             | 15                   |          |
| 2.3 Statistical Analysis using R-Studio.....                                          | 16                   |          |
| <b>3 Results .....</b>                                                                | <b>19</b>            |          |
| 3.1 Diversity parameters of the microbial community. ....                             | 19                   |          |
| 3.1.1 Alpha diversity analysis .....                                                  | 19                   |          |
| 3.1.2 Beta diversity analysis.....                                                    | 21                   |          |
| 3.1.3 Environmental filtering. ....                                                   | 24                   |          |
| 3.1.4 Key drivers of microbial community structure in PERMANOVA .....                 | 25                   |          |
| 3.2 Composition and structure of plastisphere community. ....                         | 26                   |          |
| <b>4 Discussion.....</b>                                                              | <b>29</b>            |          |
| 4.1 The Role of Primary/Secondary in Differentiation between Sample Groups.....       | 29                   |          |

|                                                                                                 |           |
|-------------------------------------------------------------------------------------------------|-----------|
| 4.2 Hyphomonadaceae in microplastic biofilm .....                                               | 29        |
| 4.3 Pseudomonadales, Alteromonadales in the Seawater and Flavobacterium in Freshwater.<br>..... | 30        |
| 4.4 Pathogenic Lead by Plastisphere Colonization of PS and PE.....                              | 30        |
| 4.5 Complex Carbon Degradation of Microplastics by Bacteria Community.....                      | 31        |
| 4.6 Limitation of study.....                                                                    | 31        |
| <b>5 Conclusion .....</b>                                                                       | <b>33</b> |
| <b>References.....</b>                                                                          | <b>34</b> |
| <b>Appendix.....</b>                                                                            | <b>38</b> |
| Appendix I. Steps for Data collection & Qiime2 analysis with DADA2 .....                        | 38        |
| Appendix II Statistical Analysis .....                                                          | 82        |

# ***1 Introduction***

## ***1.1 Background of Plastisphere.***

Since the mass generation of plastic in the 1950s, the plastic-associated product is becoming more widely used around the world, an estimated 300 million tons of plastic are produced each year, with a 10% increase from 2015. The persistence of plastic in the environment led it to become a common environmental pollutant, about 4.8-12.7 million tons of plastic waste entered the sea up to now, the plastic pollution is widely recognized as an increasing ecological concern. The fragmentation of plastic debris pieces derived to small plastic particles (diameter <5 mm) is so-called microplastics, which are diverse in shape, size, and density, and it could escape from the wastewater treatment process, result in high variability of distribution under spatial and temporal conditions.

The most visible impact of microplastics is on biota. The physical ingestion of the digestive system and entanglement of the body is common to kill animals, even transport the harmful compounds indirectly by changing various species' feeding habits. The chemical persistence and hydrophobic nature of plastic debris result in non-biodegradable and last for a long time in the freshwater or marine environment, which is dominated by the material of polyethylene, polypropylene (floating form) and PET, PVC, etc. (higher density, sediment form).

PD was served as different substrates for microbes, and the presence of particles could stimulate the productivity of microbial. Many studies indicate that the composition of the microbial communities is varied with substrate type (Miao et al 2018) and with geographical location (Oberbeckmann et al 2014). The interacting between microplastic and microbial contributed much to the complex impacts on plastic pollution situations. The bacteria could be identified as colonizers of plastic debris. It can be seen the filamentous, cyanobacteria, pennate diatoms, and other more complex biota-communities were observed from SEM images of microplastic pieces, Since the assemblage of distinct microbial that colonize on the outer surface of PD was refer to as "plastisphere" (Zettler et al 2013). Beyond prevalent recognized marine plastic pollution, a substantial portion of MPs was detected in the freshwater system. The terrestrial sources such as land-based activities, rivers, WWTP and inland waters act as the important transport vectors to open sea, especially the river networks could passively migrate the MPs for long-distance discharge (Duis et al 2016). Up till now, despite the research of plastic associated bacteria in the marine environment is well documented, few kinds of literature can be found on the comparison of the composition of microbial communities between fresh and seawater. It is needed to define the drivers, preference, and origin of colonization through observing the response of the microbial communities that grow in different environments. To investigate the bacterial diversity by using amplify the V4 hypervariable region of 16S rRNA gene sequencing on Illumina MiSeq platform and statistical amplicon analysis, compared the abundance of the plastisphere on PD to seawater, freshwater and sediment sample. More essential studies on the composition structure of the microbial community on PD from freshwater to marine urgently need.

## ***1.2 Risks of Aquatic Ecosystems Microplastic Debris***

The discussion of the past articles related to the challenges of microplastic pollution of aquatic ecosystems demonstrated the prevalence of a high number of risk sources related to changes in the composition of the ecosystem caused by the prevalence of microplastics. It is important to provide a more careful analysis of sources of risks of microplastics pollution for aquatic ecosystems to demonstrate the relevance of the presented study and existing gaps in prior research.

The discussion of the question about the impact of microplastics on the changes in the aquatic ecosystem led to the development of the definition of plastisphere. This term described the impact of microplastics debris on changes in the composition of inhabitants of the aquatic ecosystem, mainly bacterial (Zettler et al, 2013). The specialization of modern studies of plastisphere issues is associated with the analysis of changes in the composition of microbial communities in aquatic ecosystems. At the same time, the level of understanding in the aspect of plastisphere development among researchers is limited, since scholars cannot identify mechanisms of microplastic development in aquatic ecosystems and full scale of its impact (Zettler et al, 2013). The research by McCormick et al (2014) highlighted the fact that the researchers lack understanding of the mechanism of microbial communities' development on the microplastics in aquatic ecosystems.

Together with this, the current level of knowledge in the sphere of microbial community development in aquatic ecosystems, which lead to the lacks understanding of the role of environmental factors in the risk of dangerous microbial communities' development and spread (Oberbeckmann et al, 2018). It was found that different categories of microplastics can be considered as an attractive environment for the development of different forms of microbial communities, differing in the level of negative impact for the aquatic ecosystem in general (Pinto et al, 2019; Woodall et al, 2018). In specific conditions, the spread of microplastic in aquatic ecosystems can lead to the spread of dangerous toxic chemicals, causing severe risks for the inhabitants of the ecosystem (Engler, 2012). Taken together, analysis of the potential risks and challenges related to the spread of microplastic pollution in aquatic ecosystems and the lack of reliable knowledge about the potential impacts and means of its prevention serves as a basis for a more careful analysis of this issue in future.

## ***1.3 Analysis of Microplastic Transportation Among Freshwater and Marine Ecosystems***

Additional question that requires careful attention is related to the discussion of the means of microplastics transportation between freshwater and marine ecosystems. Due to the fact of differences in assessment of the level of understanding of microplastics community for marine and freshwater ecosystems, lack of understanding of the processes of microplastics transportation between these ecosystems is also presented as a serious knowledge gap. The analysis of the question about the connection between freshwater and marine ecosystems from the point of microplastics transportation performed by Schmidt et al (2017) demonstrated that the largest share of microplastics is delivered by a small share of large rivers. This finding means that the level of microplastic pollution for different types of freshwater ecosystems can differ, and a more careful analysis of the microplastics compositions for each type of freshwater

ecosystem is important (Roager & Sonnenschein, 2019). The correlation between the level of microplastic transportation and different types of freshwater ecosystems could be explored further in future study.

#### ***1.4 Analysis of the Issue of Freshwater and Marine Plastic Pollution***

The first question that must be considered in the context of the literature review is related to analysis and understanding of the prior findings in the plastisphere of the impact of plastic pollution on the ecosystems of the freshwater and marine environment. In this context, it is important to summarize the findings of the prior studies that considered the current level of microplastics composition and effect in the chosen ecosystems and define the mechanism of microplastic colonization connection freshwater and the marine.

A limited number of researches on the understanding of plastisphere in freshwater environments become the current knowledge gap. an analysis of current findings provided by Meng et al (2020) take the emphasis on plastisphere of freshwater microplastics pollution risks in the United Kingdom. This study identified the potential negative impact of microplastics on the ecosystem and health of its inhabitants, including the following outcomes: “inflammation of the digestive system reduced nutrient uptake and reduced growth and reproduction” (Meng et al, 2020, p.2). Meng et al (2020) highlighted the fact that while the study of the level of microplastics pollution in the marine ecosystem was quite popular in the past, low attention had been delivered to the discussion of the microplastics pollution risks for the freshwater ecosystems. The research by Meng et al (2020) demonstrated that though the set of studies on the issue of freshwater microplastics pollution was performed, each of these studies presented set of limitations that have to be solved in future, including the following: low availability of data for different regions of the world from the point of freshwater microplastics pollution; research bias in favor of analysis of marine and low attention to the analysis of other sources of freshwater; and low attention to the question of interaction between freshwater species and microplastics. The urge needs of the future study argued for a more careful and detailed analysis of the level of microplastics development in the freshwater ecosystem. Similar conclusions can be gained from the analysis of the study by Blair et al (2019). The findings of the study demonstrated the prevalence of prior studies related to the analysis of the impact of microplastics on marine ecosystems, while limited attention had been delivered to the analysis of freshwater ecosystem outcomes.

Previous studies also oriented on the discussion of the impact of plastic debris on the biological environment of the marine and freshwater ecosystems.( Strungaru et al 2019, Laist et al 2006, Kettner et al 2017) The discussion of these sources could contribute to a detailed understanding of the relevance and scale of the study problem. The research by Laist (2006) described the mechanism of the plastic debris impact on the marine ecosystem. According to the study findings, the negative impact of plastic pollution of marine ecosystem mainly has the mechanical character, leading to ingestion of small fragments of plastic materials and limitation of opportunities for the realization of specific behavior models or vital functions for animals, including eating, breathing and moving (Laist, 2006, p.320). It is important to highlight that though this study considered the negative impact of plastic pollution, it lacked discussion of the role of microplastics elements for the health of marine ecosystem inhabitants.

The relevance of studying the problem of microplastics pollution in the freshwater ecosystem was also demonstrated in the study by Kettner et al (2017). This study considered the impact of microplastics pollution on changes in the composition of communities in the aquatic ecosystems. According to the findings of the study, microplastics pollution changes the composition of species abundance in the ecosystem, leading to poorly understood outcomes for all species in the freshwater and marine environment. The research by Ogonowski et al (2018) provided quite a similar argument about the poorly understood nature of microplastics of aquatic ecosystems and its impact on different elements of the ecosystem. Ogonowski et al (2018) demonstrated that one of the main risks of microplastics pollution for the aquatic ecosystem is the active development and spread of new forms of bacterial communities inhabiting microplastic fragments. The researchers highlighted the fact of poor knowledge and understanding of the associated risks of this process and highlighted the need for careful analysis. Similar findings in aspects of bacterial communities' development on microplastics were presented by Tender et al (2015), arguing for the need for a more careful analysis of the microplastics composition level in different types of aquatic ecosystems and the definition of potential effects for the ecosystem in general.

### ***1.5 Comparison of Microplastics in Freshwater and Marine Environments***

For the comparison of the finer or coarser genera level with abundant prevalence of microplastic bacteria community for the marine and freshwater ecosystem, careful attention was given to the analysis of the samples collected from the marine environment. two types of marine environments were considered: mesocosms and marine litter. The main advantages of the mesocosms methodology application are that laboratory treated sample allows the researchers to carefully control the influence of environmental characteristics of the level of bacterial community of microplastic. In these conditions, it is important to identify what specific characteristics of the environment led to the more active spread of specific bacteria community, and what differences compared to the freshwater environment can be observed. Such an approach to the process of samples comparison and analysis explore further the difference of diversity about bacteria assemblages on microplastic surface under distinct environmental pressure. (Kesy et al 2019)

The second approach to the analysis of the marine environment samples included analysis of the marine litter, which could allow the researcher to perform a detailed analysis of bacterial communities that could be developed in the marine ecosystem environment. The comparison of the results with findings of the freshwater ecosystem assessment could demonstrate differences in the level of ecological effect caused by the spread of microplastic for different types of ecosystems. (Kesy et al 2019) At the same time, a comparison of results for different types of marine litter depending on the materials of the litter collected, that could contribute to a better understanding of the specific role of microplastics as a source of pollution-related risks for the ecosystem.

Taken together, the findings of these studies argued for the need for a careful consideration of two categories of environment for the analysis of the changes in the bacterial community - mesocosms and marine litter, and then compare them with the microbial of microplastic in freshwater. According to these findings, the specific environment elements (such as Salinity,

sample type, collection location, etc.) were considered as a basis for the collection of data about the level and impact of microbial colonization on the aquatic ecosystem.

### ***1.5 Next-generation sequencing of 16S microbiome data for Plastisphere species.***

The methodology of the study included the application of a wide range of methods for data collection and analysis. Among the means of data collection, the following instrument were applied: next-generation sequencing (NGS) for 16S ribosomal RNA (rRNA). V4 hypervariable region was targeted for all studies since this region involve the high nucleotide heterogeneity that allow high discriminatory power. (Zimmermann et al 2011) Amplicon Sequence Variants (ASVs) method was utilized, which able to gain finer resolution by recover all biological variation based on potential noise models. (Callahan et al 2017) Next-generation sequencing for the 16S ribosomal RNA (rRNA) research method was developed at the beginning of the twenty-first century and led to significant changes in the plastisphere of microbial ecology research. The main advantage of this methodology was related to a significant increase in the production of real data with a reduction of expenses for the samples collection procedure. Optimization solutions that contributed to the improvement of the quality of microbial materials analysis included the following: multiplexing sequencing, immobilization of a complex library of DNA templates, in vitro amplification. Together with this, Illumina sequencing methodology was selected as an optimal solution due to the following advantages: low error rate, reduced costs and high flexibility in data collection.

For the aims of the sequencing of samples collected from the freshwater and marine environments, the instrument of Illumina MiSeq sequencer was applied. This is the integrated instrument applied for the aims of clonal amplification, genomic DNA sequencing, and data analysis with base calling, alignment, variant calling, and reporting in a single run. The main characteristics of this sequencing method include the following: application of a double-sided, single lane flow cell, and reagent cartridge provided in the kit form. (Foley et al 2018) The procedure of sequencing using the Illumina MiSeq sequencer method includes the recording of the DNA strands synthesis in a cluster of sample templates. The procedure of recording leads to the development of newly attached bases liberating the fluorescent dye that can be excited by the diode lasers. Two digital cameras are applied for the aims of the presentation of the process to the researchers. (“Genomics and Sequencing Center, University of Rhode Island,” n.d.) The main advantages of the Illumina MiSeq sequencer include the following: flexible adjustment of reading the length of the genome due to effective sequential interrogation of bases during a single run; a high number of samples (up to 96 samples) that can be sequenced during a single run; and high capabilities of a single lane flow. Taken together, the application of the Illumina MiSeq sequencer can lead to a more productive analysis of the data and samples collected from different examples of the aquatic environment, and reliable comparison of the main characteristics of freshwater and marine environments from the point of microplastics. (Ogonowski et al 2018)



## ***1.6 Meta-Data Analysis Overview***

The multiple studies are not feasible to be compared at same time as many different methods and analysis using in different study, which described the microbial communities in different environment with different parameters (pH, salinity, temperature, colonizing plastic type, etc.). However, we can perform greater statistics with large number of samples in meta-analysis for annotated systematic review. In this way, secondary data used for the number of studies were collected from the prior research; they contributed to a complex vision of the research issue and alternative positions about the perspectives of its resolution. The main purpose of the meta-data analysis process was integrating primary data from different studies into the single research framework, boosting the effective utilization of available data and knowledge for the development of reliable findings (Ho et al., 2019; Sinclair et al., 2016). It is helpful to understand the community structure of the microbial associated with microplastics in different freshwater environments and how these differ from what has been found in these seawater samples. Taken together, the meta-data analysis was contributed to the effective management of large samples and easing the process of statistics analysis in composition differences of bacterial community to environment factors. (Koci et al., 2018).

## ***1.7 Existing Gaps in Knowledge***

Based on the analysis of the results of prior studies, it was possible to identify the significant phylogenetic tree of the bacteria colonization factor on microplastic in the future development of the aquatic ecosystem. The spread of different forms of microplastic litter can lead to the development of new forms of bacteria in the aquatic ecosystem, leading to the risk of change of the existing balance of inhabitants in the ecosystem (Parrish & Fahrenfeld, 2019). At the same time, it was found that the main gap in knowledge commons for the prior studies is related to the absence of a clear and detailed understanding of the composition of microplastic pollution for different types of aquatic ecosystems (Blair et al, 2017). One of the knowledge gaps is associated with the absence of a clear understanding of the specific influence of microplastic spreading in freshwater ecosystem. Consistently, compared with marine environment, A limited number of studies targeting the plastisphere in freshwater environments become the current existing gap. Nevertheless, the main channel of transportation of microplastic particles in marine is the freshwater ecosystem. (Schmidt et al 2017) There are lots of ecological factors such as geographical location, substrate type, environmental parameters (pH, Temperature, Salinity) may play the significant role to result the different features of plastisphere between freshwater and seawater. The metadata analysis is urged needed to analyze the community structure diversity for samples originated from different aquatic system. Moreover, the microplastic degradation pathways corresponding to the specific genera abundant in a microbial community, which is the one underexplored knowledge gap.

## ***1.8 Aims of the Study***

The main aim of the study is to investigate an environmental pollution phenomenon called the “plastisphere” through meta-data analysis. The plastic associated microbial communities are assumed those differs from freshwater and seawater environment. Through analyzing the meta-

database from existing research, the amplicon sequencing of samples was performed by utilizing the 16S rRNA gene sequence, the bioinformatic analysis of DADA2 (ASV) demonstrated abundant gene encodings of bacterial communities that colonized on plastic and organic particle (biofilm, suspension, wood, etc.). To investigate the environmental parameters that influencing bacterial assemblages, there is four objectives: 1. Investigate the influence of geographic location that affect the microbial composition associated with various type of microplastics (marine plastic litter, mesocosms, and microplastic in freshwater). To explore spatial influences accurately, the samples under the incubation experiment were compared with the samples direct from the environment. 2. Investigate the characterization and composition structure of the microbial communities on different substrate types. The diversity of the composition of bacterial communities to different plastic-type could elucidate the drivers of colonization. Also, the organism samples (biofilm, wood, surrounding freshwater, surrounding seawater) served as the biotic control, the comparison between the distinct bacterial communities on microplastic and other environmental samples was expected to be analyzed .3. Determine the main factors from distinct environment conditions (pH, salinity, temperature, etc.) that influence the microbial composition of microplastics in freshwater. 4. Explore the ecological influence and microplastic degradation pathways for the finding of specific genera dominant in the plastisphere.

## 2 Method

### 2.1 Background of studies.

Metadata table was generated for integrating primary data from four studies into one research framework: (Table 1.)

(Bio Project PRJNA325458) - The first study, “Attached biofilms and suspended aggregates are distinct microbial lifestyles emanating from differing hydraulics” (Niederdorfer et al 2016) was related to the assessment of the influence of different hydrodynamic conditions on microbial diversity in aquatic microbial lifestyles. The samples were collected from floodplain of River Danube and headwater stream of River Lunz Am See. A total number of 192 SRA experiments were analyzed in the study. Measured Environment parameter shown as: pH 8.4, Temperature 16 -20 °C, Salinity 2.5 PSU.

(BioProject PRJNA506548) - The second research, “Spatial Environmental Heterogeneity Determines Young Biofilm Assemblages on Microplastics in Baltic Sea Mesocosms” (Kesy et al 2019) investigated the role of different factors in the process of the microbial community assembly on microplastics. It was important to compare the influence of surface-specific assembly factors and environmental factors in this process. The study collected the samples from 8 different geographic locations along the eastern Baltic Sea coastline. The collecting point is recorded as varied pH, salinity (from 4.5-9 PSU), depth and temperature. Then samples were conducted incubation mesocosms experiments for 7 days. It examined the development of biofilm on types of plastic debris and other substrate (Organic matter, Wood, seston and in the free-living fraction). A total number of 301 SRA experiments were analyzed in the study.

(BioProject PRJNA242639) - The third study, “Microplastic is an Abundant and Distinct Microbial Habitat in an Urban River”, (McCormick et al 2014) used in the process of data collection, it compared the characteristics of microbial communities colonizing microplastic in the aquatic environment and the microbial communities found on natural substrates in a freshwater environment. The samples were collected in September from The North Shore Channel in Chicago, Illinois (IL), USA. Analysis of our samples compares the bacterial community composition that colonizes microplastic in comparison to those that colonize natural substrate. Samples were collected from water column upstream of a WWTP, water column downstream of a WWTP, organic material downstream of a WWTP, and microplastic (fibers, fragments, Styrofoam, pellets associated with synthetic textiles and personal care and cleaning products). The point source was measured the pH value close to neutral (pH=7.2) and less salt concentration (salinity 0.3-0.5). A total number of 16 SRA experiments were analyzed in the study.

(BioProject PRJNA338729) - The final study “Environmental Factors Support the Formation of Specific Bacterial Assemblages on Microplastics” (Oberbeckmann et al 2018), compared the characteristics of the microbial communities found on microplastics pellets (HDPE HTA 108, BASF PS 143E) to the communities on natural particles (wooden pellets, free-living water communities). The research also investigated the ecological significance of specific bacterial population on MP-associated material in aquatic environment (Marine, Freshwater) and in plastic accumulation zones (In WWTP). The samples were collected from Warnow river, coastal

Baltic sea and from freshwater area close to WWTP in Rostock, Germany. These from distinct environment were incubated for two weeks, and recorded the specific environment parameter (pH 7.6-8.1, Temperature 16 -17.8 °C, Salinity 0.5 -14.1 PSU). A total number of 154 SRA experiments were analyzed in the study.

| Study            | Accession_No | Number of Samples | Sample Types            | Sample Location                               | pH      | Temperature Range (°C) | Salinity Range (PSU) | Region Targeted | Platform       | Extraction_Method                         |
|------------------|--------------|-------------------|-------------------------|-----------------------------------------------|---------|------------------------|----------------------|-----------------|----------------|-------------------------------------------|
| Key2019          | PRJNA506548  | 301               | PE_Sewater              | Bay of Gdansk - Baltic Sea                    | 8.1     | 10-20.6                | 4.5-9                | V4              | Illumina MiSeq | Dneasy PowerSoil Kit                      |
|                  |              |                   | PE_Freshwater           | Polish coast - Baltic Sea                     |         |                        |                      |                 |                |                                           |
|                  |              |                   | Seawater                | Boleniec, Baltic Sea                          |         |                        |                      |                 |                |                                           |
|                  |              |                   | Freshwater              | Jaroslavec, Baltic Sea                        |         |                        |                      |                 |                |                                           |
|                  |              |                   | Wood Pellets            | Piasnica, Baltic Sea                          |         |                        |                      |                 |                |                                           |
|                  |              |                   | PS_Sewater              | Rostock, Germany                              |         |                        |                      |                 |                |                                           |
|                  |              |                   | PS_Freshwater           | Gulf of Riga, Baltic Sea                      |         |                        |                      |                 |                |                                           |
|                  |              |                   |                         | Narva mouth, Baltic Sea                       |         |                        |                      |                 |                |                                           |
|                  |              |                   |                         | Helinski, Baltic Sea                          |         |                        |                      |                 |                |                                           |
|                  |              |                   |                         | Klaipeda, Baltic Sea                          |         |                        |                      |                 |                |                                           |
|                  |              |                   |                         | Pomeranian Bay, Baltic Sea                    |         |                        |                      |                 |                |                                           |
|                  |              |                   |                         | Bulduri, Baltic Sea                           |         |                        |                      |                 |                |                                           |
|                  |              |                   |                         | Tallinn Bay, Baltic Sea                       |         |                        |                      |                 |                |                                           |
|                  |              |                   |                         | Sassnitz, Baltic Sea                          |         |                        |                      |                 |                |                                           |
| Niederdorfer2016 | PRJNA325458  | 192               | Freshwater              | Floodplains of the River Danube               | 6.8     | 20                     | 2.5                  | V4              | Illumina MiSeq | Allprep DNA/RNA/Protein Mini Kit (Qiagen) |
|                  |              |                   | Biofilm                 | Luna am See                                   |         |                        |                      |                 |                |                                           |
|                  |              |                   | Seawater                | Baltic Sea, Germany                           |         |                        |                      |                 |                |                                           |
| Oberbeckmann2018 | PRJNA338729  | 154               | Freshwater              | River Warnow                                  | 7.5-8.1 | 15-17.8                | 8.3-14.1             | V4              | Illumina MiSeq | QIAGEN QIAampR DNA Mini Kit               |
|                  |              |                   | Wood                    | WWTP discharge, Germany Rostock               |         |                        |                      |                 |                |                                           |
|                  |              |                   | HDPE (HTA 108)          | Alter Strom in Warnemünde, Coastal Baltic Sea |         |                        |                      |                 |                |                                           |
|                  |              |                   | BASF Polystyrol 143E    |                                               |         |                        |                      |                 |                |                                           |
| McCormick2014    | PRJNA242639  | 16                | PS                      |                                               | 7.2     | 19.15                  | 0.315                | V4              | Illumina MiSeq | MoBio Powersoil DNA extraction kits       |
|                  |              |                   | organic matter          | North Shore Channel, Chicago                  |         |                        |                      |                 |                |                                           |
|                  |              |                   | Downstream water column |                                               |         |                        |                      |                 |                |                                           |
|                  |              |                   | Upstream water column   |                                               |         |                        |                      |                 |                |                                           |

**Table 1. Simplified Meta-Data Table. Collection from four studies with 646 samples, details about: Sample Types, Sample Location, pH range, Temperature Range, Salinity Range, Region Targeted, Platform, Extraction Method.**

## 2.2 Sequencing Analysis

### 2.2.1 FASTQ collection

Four studies (646 samples) were collected from reliable prior studies to create meta-database (Table 1). Analysis were based on the high throughput 16S rRNA gene V4 region sequencing (paired end reads on Illumina MiSeq). Each of the presented studies discussed the formation of microbial and biofilm assemblages in aquatic environment, these also characterized the environmental parameters. The results of the FASTQ files collection from reliable prior studies contributed to the group classification of the samples with specific parameters: geological location, isolation source, sample type, pH ranging, temperature and salinity, detailed information of samples and bio-project ID in studies was recorded in NCBI website and prior studies. One metadata table was created organizing the parameters of samples for systematic review, and the table in CSV format was then utilized in statistics analysis in R.

### 2.2.2 Data Download and File Organization.

In order to download sequences of sample, SRA toolkit and NCBI's e-utils was activated on Orion cluster (sponsored by Dr Umer Zeeshan Ijaz). The SRR numbers associated to Bioproject PRJNA for four studies were sourced to download corresponding FASTQ files.

The project folder was created and then moved into the folder, then created sub-folder called Sequence folder and moved into the folder, the folder names were extracted from the paired-end files, Placed all FASTQ files within corresponding folders. Within each of these folders, dumped the raw sequences in a "Raw" folder.

### 2.2.3 Analysis using QIME2

The amplicon analysis was performed on QIME2 (Version 2019.7), which is an advanced workflow system able to generate the ASVs. For the realization of the QIME2 analysis procedure, the DADA2 method of analysis was used. In the first step of the analysis sequence, the abundance table of Amplicon Sequence Variants (ASVs) was constructed for the aims of the identification of several times when each exact amplicon sequence variant was observed. The fictitious barcode required to import data in the Earth Microbiome Project (EMP) format was generated. For each of the sample reads, the specific barcode was generated. To summarize primary data sourced from different studies, forward reads were assembled in a single forward.fastq file. Similarly, all the reverse reads were assembled into the single reverse.fastq file. All the FASTQ files generated in the prior steps of data analysis were integrated into the folder named “emp-paired-end-sequences.”

The following phase of analysis included the import of the folder named “emp-paired-end-sequences” into the QIME2 platform. Demultiplexing of the sequences was performed to use DADA2 analysis procedures. Quality control sets were developed using the DADA2 algorithm, Dada2 took the trimmed reads and then learned the error rates. In this case, trimmed the forward at 200, everything after 200 is discarded and trimmed the reverse at 180, everything after 180 was discarded. After that, the phylogenetic tree for the ASVs was generated. Based on the achieved result, a taxonomy for the ASVs was developed and compared with Silver Classifier (132-99-nb). In the final stage of QIME2 analysis, the generated files were exported in the format compatible with the R-analysis format: the .biom file of feature table and .nwk file of Phyloseq tree.

### 2.2.4 DADA2 Analysis

To infer data sequence variants in the process of data analysis, the DADA2 method for the amplicon sequence variants (ASVs) was used. The reliance on the DADA2 method of sequence variant analysis was explained by the advantages of this method from the perspective of the differentiation of sequencing errors from real biological variation. This method allowed the researcher to collect a consistent data sample from a wide range of prior studies while minimizing the risks of mistakes accumulation, which could affect the quality of study outcomes in the future. DADA2 analysis can be performed using two alternative modes – independent inference by sample and inference from pooled sequencing reads. However, each of these modes has specific advantages and limitations. The mode of independent inference by the sample is beneficial in terms of the assessment of computational time and flat memory requirements. At the same time, inference from pooled sequencing reads mode is associated with higher computational requirements, though this method can serve for the effective identification of rare examples in a high sample size. Taken together, inference from pooled sequencing reads was considered a more promising approach to the DADA2 method of analysis application.

The comparison of DADA2 ASVs analysis and Operational Taxonomic Units (OTUs) demonstrated the advantage of the orientation on ASVs in the realization of the research objectives due to a set of parameters. First, as compared to the OTUs approach of 97% similarity, ASVs data application provided more complex and effective incorporation of a wide range of data, including rare elements and samples. In the context of the OTUs approach, the

limited scale of data was incorporated, especially excluding sequence quality information and statistical information on the sample reads. In this context, the orientation on the ASVs data presentation approach together with the DADA2 method of analysis was found to be a sensible decision due to the higher quality of data management and reliability of potential results. (Callahan et al. 2016)

### **2.3 Statistical Analysis using R-Studio**

The diversity within samples (alpha diversity) and between samples (beta diversity) were assessed. The objective of both alpha and beta diversity methods was based on identification of the characteristics of sample species, which was supported for development of taxonomy. The difference between alpha and beta diversity methods application is the level of habitat and environment being studies. The main task of the alpha diversity analysis procedure is to identify level of generalized mean common for all species in small habitat and assess level of species diversity. Data filtering and normalization procedure were applied to identify the key characteristics of each data cluster. (Chong et al., 2020) The procedure included the following steps: data filtering, data rarefying, data scaling, and transformation. In the process of data filtering, it is important to remove low-quality data from the data sample to improve the results of downstream statistical analysis (Chong et al., 2020). This step was also used in the context of beta diversity analysis. Taken together with other steps of alpha diversity analysis, chosen methods contributed to the improvement of the statistical sample quality and effectiveness of its utilization in the context of the study. Therefore, the Measurement of alpha diversity analysis was applied for the comparison of the species richness and evenness. Different groups (Sample Location, Sample Type, Salinity, pH range, Temperature range) classified from samples was assessed, tested parameters was assumed as several potential factors in shaping of Plastisphere features, Four measurements were applied: *Fisher index*, assume the abundance of species follow logarithmic series distribution; *Pielou's evenness*, measure the evenness of community; *Richness index*, measure each rarefied one in samples (target to the minimum of library size) of species; *Shannon index*, measure the balance within the community, the higher *Shannon index* indicates higher diversity; *Simpson index*, measure the evenness of the community in alternative way. Index equals to 1 represents infinite diversity and index equals 0 represents no diversity. Taken together, the measurements above analysed the diversity using different formula in order to compare the multiple aspects of diversities.

As compared to the alpha diversity analysis, beta diversity analysis served the comparing differences between the community structure of test groups. It was applied for analysis of species composition in large number of habitats. The results of beta diversity were presented in the form of the distance matrix, which could demonstrate patterns of species distribution in different samples (Chong et al., 2020). Different forms of beta-diversity measures were applied in modern analysis frameworks. 1. Weighted UniFrac, it combined the relative abundance and phylogenetic distance together in algorithm. Rather than Unweighted UniFrac, this phylogenetic distance metric always focusses on the difference of abundance, and the phylogenetic distance was calculated by Phyloseq package. Noted the evenness of OTUs affect the distribution of clusters. 2. Unweighted UniFrac, is a computing method of Dimensionality reduction of OTUs cluster, which approximates the correlation between OTUs by calculating the distance between two samples without consider about the abundance. The distance

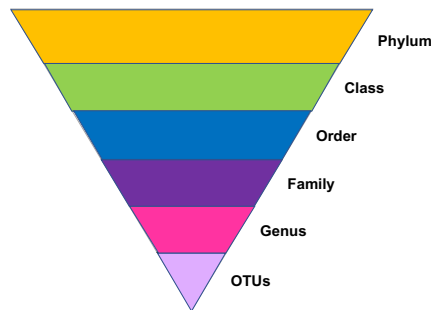
measurement depends on the percentage of the sum of unshared branch length in the sum of total branch lengths. 3. Bray- Curtis, similar to Weighted UniFrac, it is a kind of distance metric and take the abundance into account. The main objective of the beta diversity analysis methods application is related to the assessment of differences in the microbiome composition between test groups. The list of potential disadvantages of the beta diversity parameter assessment included the following: overestimation of role of rare species in the process of sample taxonomy assessment, and the application of beta diversity as the parameter of nestedness. The results of beta diversity assessment for definition of sample taxonomy might be quite contradictory, which identified need of application of both alpha and beta analysis procedures. Taken together, the application of alpha and beta diversity analysis methods contributed to the identification of the community structure in different groups, depending on the source of the sample and the procedure of data collection applied. The orientation on these methods of sample species comparison was contributing to the identification of the factor on the changes in the microbial communities' structure in different aquatic ecosystems.

R version (Version 3.5.2) was used for all statistical analysis, total 663 metadata samples were under quality control through DADA2 algorithm. QIME2 generated the .biom file of feature table that was compatible with R and with the package Phyloseq, the feature table (including the feature ID and taxonomy) and meta-data table was then utilized in R studio functions (). RColorBrewer () and gg\_color\_hue () package was used for visualization. Accordingly, The Vegan package was installed for running alpha and beta diversity. The higher value in Y co-ordinate demonstrates the more variation of abundance between distinct OTUs species. For beta diversity, the PCoA plots of OTUs was utilized by calling three difference measures for system distance between the species, the cmdscale () function in Vegan package is tool, and The ordiellipse () function in Vegan package 2.5-6 version generated the ellipses in beta plots (McKenna et al. 2020); hence the grouped samples were distinguished in better visualization. The 95% confidence interval of the standard errors of the groups were covered. (Ijaz et al. 2018). The pair-wise ANOVA or PCoA boxplots for alpha and beta diversity performed the P value as the correlation between two samples. The significance (\*) was displayed on the top of plots, the connection line means two samples were compared together. (\*:  $0.01 < p < 0.05$ ; \*\*:  $0.05 < p < 0.001$ ; \*\*\*:  $p \neq 0.001$ ), the larger value of P represents the larger phylogenetic distance between two samples.

The environmental pressure (Sample Location, Sample Type, Salinity) was examined through Alpha diversity. In addition to distance metrics (Brat-Curtis, Unweighted UniFrac, weighted UniFrac), the concept of PERMANOVA (fit linear models to distance metric) was applied and adonis () function was performed for analysis on sources of variation. The nearest taxa index (NTI) and the net relatedness index (NRI) was to determine the phylogenetic distance in consideration of host environmental pressure within the samples. The significant role of NRI/NTI is to characterize whether the microplastic samples driven by competition among OTUs or variety of environmental factors. The picante package was installed for using mpd() and ses.mpd() functions to calculate mean phylogenetic diversity and NRI index, mntd() and ses.mntd() was for calculating NTI index. Respectively, the ses.mntd() and ses.mpd() function allow the NTI/NRI output negative index display in Y-coordinate if diversity of community represent the phylogenetic overdispersion.

Based on the outcomes of alpha and beta diversity assessment for the sample species, Taxonomy annotation was developed. This measurement served for the identification of

microbial community structure of samples in distinct microplastic environment. the taxa bars were analyzed in R with the library size of 2000 to ensure collating the all OTUs, Genus, Family, Order, Class and Phylum (Figure 1), species at these levels were identified that extract the most top 25% abundant taxa that was active on microplastic samples.



*Figure 1. Major Taxonomic Ranks: phylum, class, order, family, genus, and OTUs*



### 3 Results

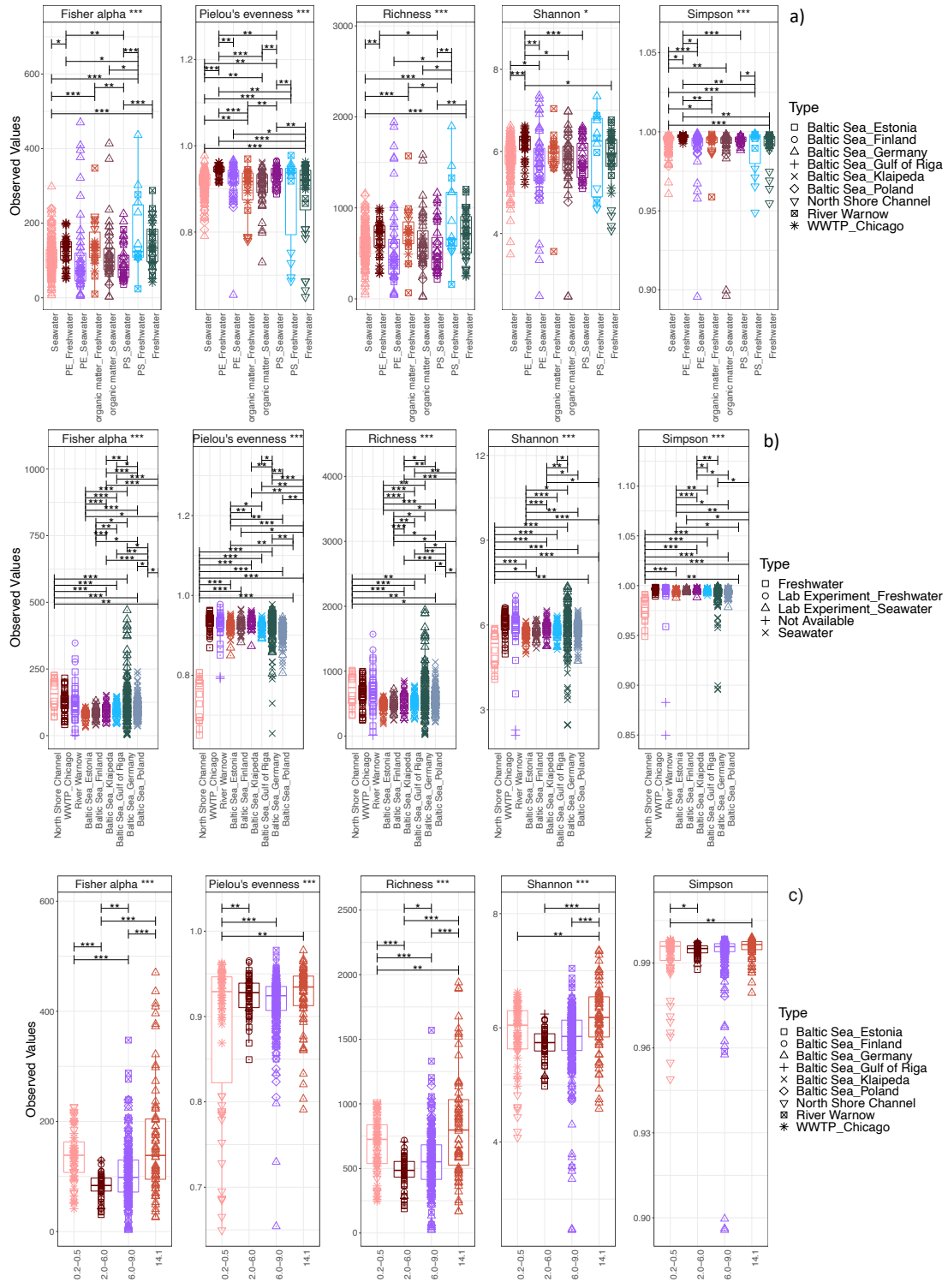
#### 3.1 Diversity parameters of the microbial community.

##### 3.1.1 Alpha diversity analysis

It is reasonable to compare the alpha diversity of the sample groups because the diversity of a single sample is not feasible to be measured. This is a valid procedure if the systematic errors and biases are consistent across different samples, which is typically the case of this study that all samples were sequenced in the same run. The more variety of species contribute, the higher the alpha diversity will be represented in the boxplot. Any group trends to have higher or lower diversity than another group, several different matrices are expected to see followed by relative higher or lower P-value. (McCormick et al, 2014) The diversity with high significance ( $P < 0.05$ ) was expected compared with the different groups.

Firstly, pair-wise ANOVA plot was performed on samples classified at different substrate types to observe diversity differences (Fig. 2a). The shapes were identified by different collection sites, and the mean value was taken for result in quartile boxplot. The analysis showed a slightly higher richness and overall diversity on PE samples in freshwater based on *Fisher*, *Pielou's evenness*, *Richness*, *Shannon*, and *Simpson* alpha. The *Simpson* index from each group of samples was significantly high (just close to 1.0), which indicates high diversity within each group of samples. Initially, we compared the difference of diversity between different types of samples but in a similar environment (both groups from seawater, both groups from freshwater). The diversity of the *Simpson* index between PS in freshwater and PE in freshwater has a high significance difference (\*\*  $p < 0.01$ ). Whereas, the diversity between PS in seawater and PE in seawater hasn't significant difference since there was no connection line comparing them. The diversity of microbiota colonizing on biofilm of microplastics in seawater may less influenced by the type of microplastic, but different type of microplastic played the significant role in freshwater. Furthermore, the diversity of samples in the different aquatic environment were compared. The *Shannon index* and *Pielou's evenness* indicated there were high significant (\*\* $p < 0.01$ ) of diversity between the PE in seawater and PE in freshwater. Consistently, *Fisher index* of diversity indicated high significance (\*\* $P = 0.001$ ) between the PS in freshwater and PS in seawater. Therefore, the colonization of bacteria community in the distinct aquatic environment was likely related to the characteristic of selection preference.

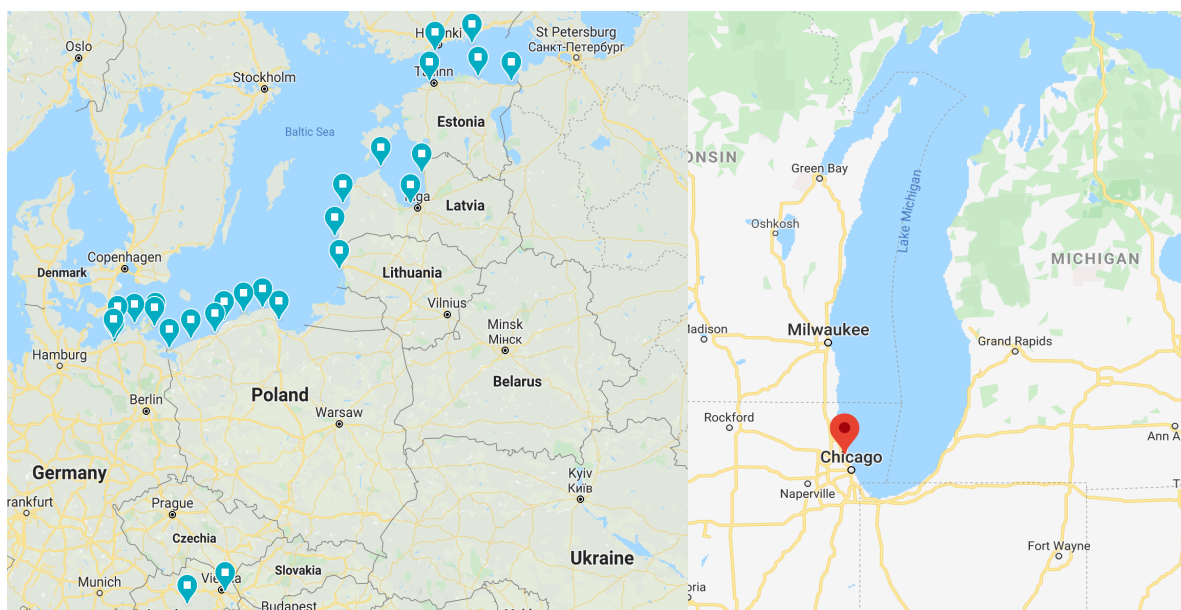
However, since there was no significant diversity difference between groups that were based on sample type, Sample location was used as new variables to create an ANOVA plot, and then more comprehensive diversity was observed. the six out of nine groups in the ANOVA plot comprise of seawater samples collected along the Baltic Sea coastline, (Fig. 2b) the shapes were identified using the source of samples (marine litter and mesocosms). *Fisher* alpha displayed the groups with River Warnow, North Shore Channel and WWTP in Chicago indicated higher alpha diversity than sample groups from the Baltic sea. Notes that due to the lack of evenness and balance on samples collected from North Shore Channel in Chicago, we just focused on to *Fisher* Alpha and *Richness* index. The significance on the top of the figure reflected the high diversity of taxa between samples from coastlines of different locations of the Baltic sea. As can be seen from the sample collection location (Fig. 3), Germany and Poland close to the estuary, where the coastline of Estonia, Klaipeda, Finland, and the Gulf of Riga



**Figure 2. (a) Alpha diversity metrics for Seawater, PE\_Freshwater, PE\_Seawater, organic matter\_Freshwater, organic matter\_Seawater, PS\_Seawater, PS\_Freshwater, Freshwater samples (b) Alpha diversity metrics for North Shore Channel, Lunz am See, River Danube, WWTP\_Chicago, River Warnow, Baltic Sea (multiple locations) using DADA2 pipeline. (c). Alpha diversity metrics for Salinity Range (0.2-0.5 PSU, 2.0-6.0 PSU, 6.0-9.0 PSU, 14.1 PSU)**

belongs to the inland sea the geographic location difference or other extrinsic parameters may cause variation of diversity of bacteria composition on plastisphere (\*\* $p=0.001$ ). Importantly,

the most noteworthy in the alpha diversity in term of geographic location (Fig. 2b) is the significant correlation of diversity between samples in freshwater and in seawater, such that the *Shannon* index displays the significant difference (\*\* $p=0.001$ ) between North Shore Channel group and Baltic sea groups.



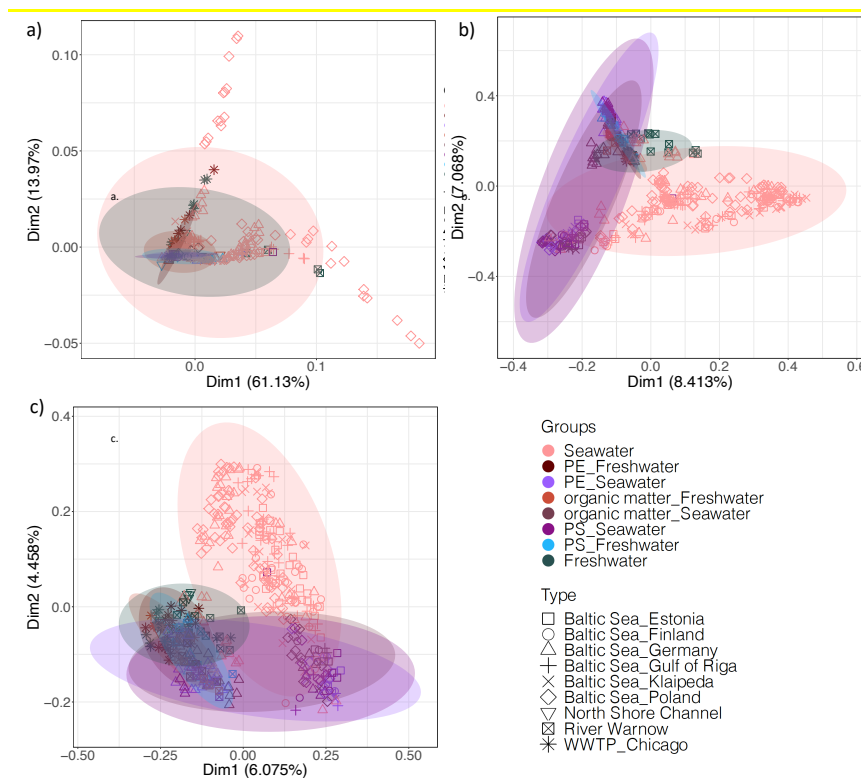
**Figure. 3 Geographic Location display the distribution of samples from studies. (Cited from Google Map Application)**

The average salinity in the different aquatic environments may be one of the primary clues for the high  $p$ -value of diversity between groups (Fig. 2c). The samples were collected from the different geographic locations of the Baltic Sea and from freshwater. The groups with low salt concentration (0.2-0.5 PSU) indicated samples collected only in freshwater (River Danube, North Shore Channel, and River Lunz Am See), which contained the microplastic particles and organic matters. The high alpha diversity was represented in the ANOVA plot (Fig. 2c), the significant correlation (\*\* $p=0.001$ ) was observed between the lower salinity group (0.2-0.5 PSU) and high salinity groups (6.0-9.0, 14.1 PSU) calculated by *Pielou's Evenness* and *Richness index*. When put the emphasis on the shapes of diversity distribution, the samples collected from the Baltic sea along the coastline of Gulf of Riga, Poland, Germany were found under high salinity condition (6.0-14.1 PSU); Samples collected from the Baltic sea along the coastline of Estonia, Finland were found under less salinity condition (2.0-9.0). Interestingly, *Shannon* and *Fisher* were performed significant differences (\*\* $p=0.001$ ) between groups in spite of all those samples were collected from the Baltic sea. The difference of salt concentration elucidated such an important factor affecting the diversity of the microbial community on microplastics. (Yang et al. 2020) The correlation between the salinity and diversity of the microbial community on microplastic is urged to investigate further since the higher diversity was represented in the low salinity group (0.2-0.5 PSU) that sourced from freshwater.

### 3.1.2 Beta diversity analysis

Beta diversity was performed using three measurements referred to as the Method part. The sum of 71700 OTUs with the top 25% abundant genera was analysed by R and obtained three PCoA plots by Bray-Curtis, weighted UniFrac, and Unweighted UniFrac. Firstly, considering the phylogenetic distance associated with the type of samples PE, PS, organic matter, seawater, and freshwater, (Fig. 4) the shapes of samples were classified as Geographic Location. Since the drawing ellipse was based on standard errors, the high significant superposition of ellipses between different groups was conducted by weighted UniFrac, this measure was not considered in this case. The Bray-Curtis demonstrated samples seawater represents a more pressing dispersion by measuring the phylogenetic distance. However, PCoA plot displays that no apparent separation of groups by comparing either different plastic material or the samples in a different environment (freshwater and seawater). More precisely, the seawater group was observed having relatively higher diversity as compared to Group freshwater and Group PE/PS in waters. Accordingly, Unweighted UniFrac indicated the diversity of the microplastic bacteria community on PE/PS was still less pronounced in freshwater and in seawater. From three Plots, the clear difference was only observed on clusters related to seawater.

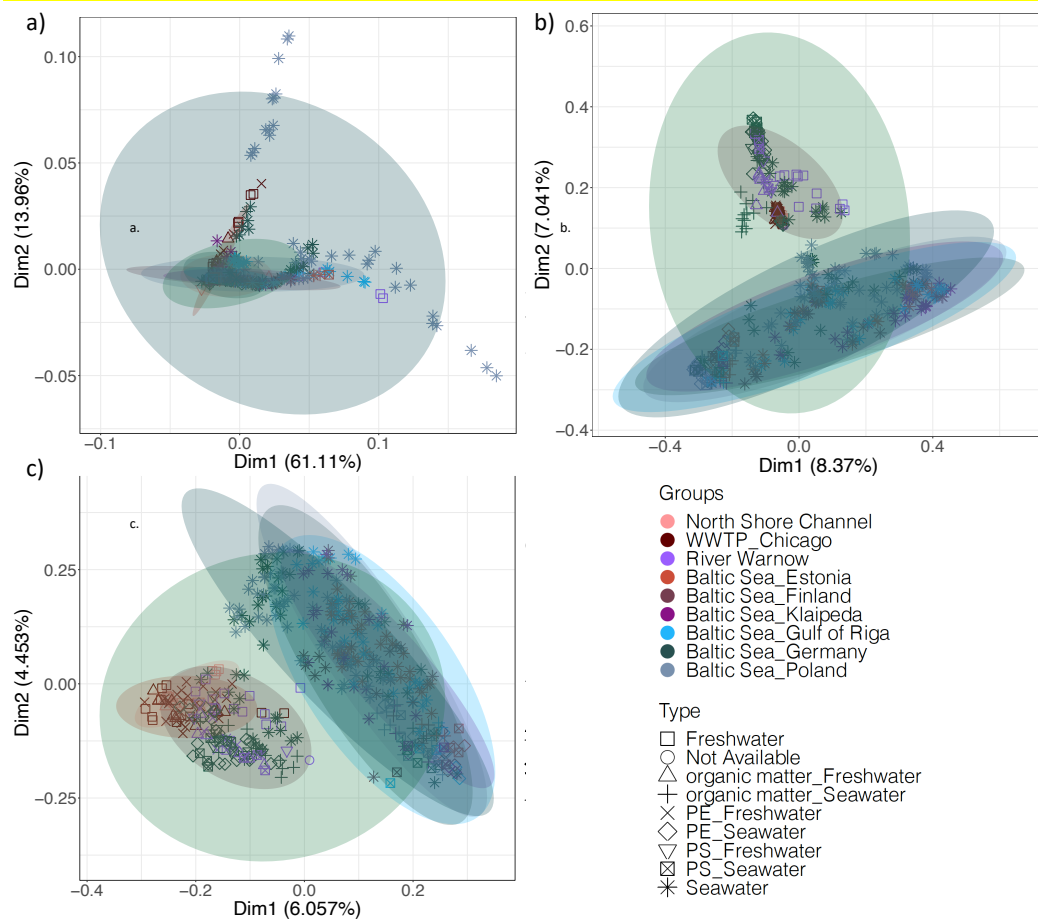
In term of considering Geographic location effect, (Fig. 5) Samples from Baltic sea (Germany) and Baltic sea (Finland) showed larger distance than the groups in freshwater, the



**Figure 4. PCoA of beta diversity was measured for Sample Type groups using: (a) weighted UniFrac. (b) Bray-Curtis. (c) Unweighted UniFrac. Two samples if similar lie very close to each other. The ellipses represent the standard error in terms of grouping variations.**

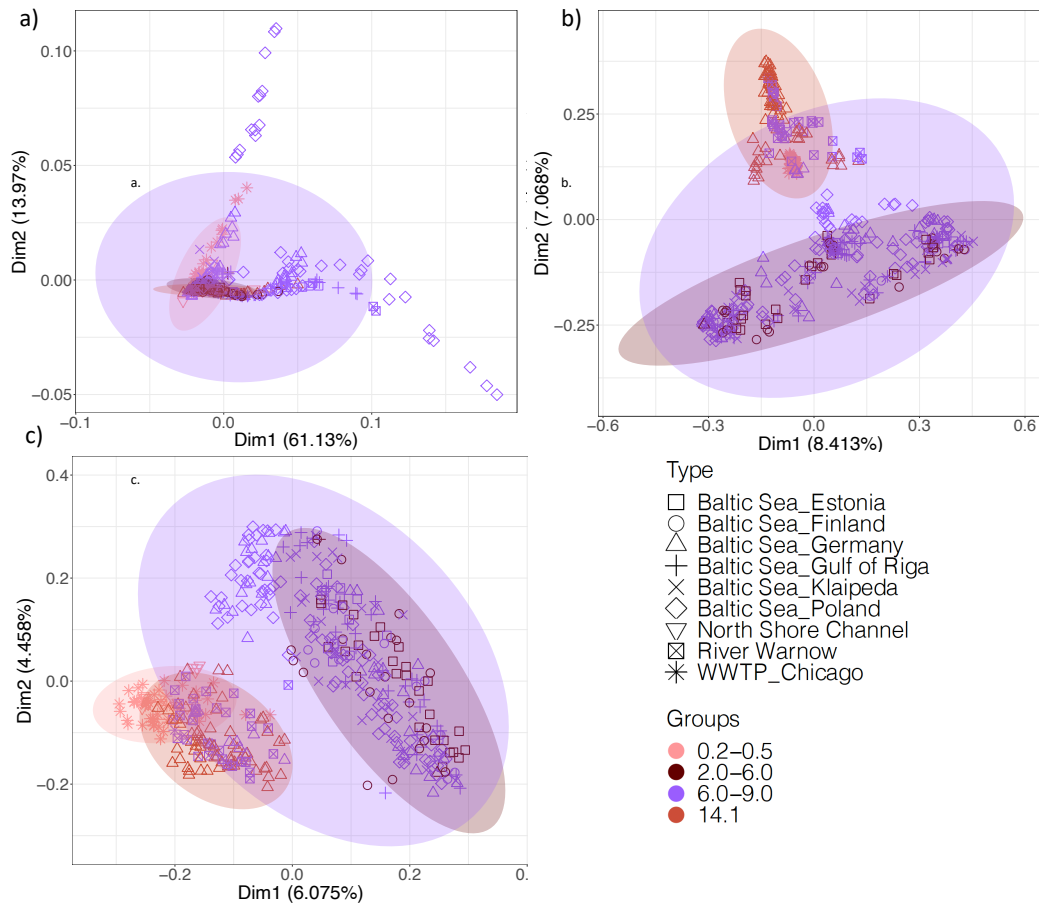
difference between River Warnow and Baltic sea (Poland) and Baltic Sea (Gulf of Riga) was not clear since there was a large proportion of overlap of the ellipses. UniFrac and Bray-Curtis distance measures indicated the spreading of samples from the Baltic sea (Germany) mainly

distributed into two clusters, one of the clusters was no clear separation with the samples from the Baltic sea (Finland) and North shore channel of Chicago. The other cluster had a wider spreading and overlapped a lot with the samples from different locations of the Baltic Sea and freshwater.



**Figure 5. PCoA of beta diversity was measured for Sample Location groups using: (a) weighted UniFrac. (b) Bray-Curtis. (c) Unweighted UniFrac. Two samples if similar lie very close to each other. The ellipses represent the standard error in terms of grouping variations.**

At salinity range (Fig. 6), the Bray-Curtis and UniFrac showed clear difference between the group with salinity (0.2-0.5 PSU) and the groups with salinity (2.0-6.0, 6.0-9.0 PSU). Interestingly, the cluster with low salinity (0.2-0.5 PSU) was close to cluster with high salinity (14.1), significant overlapped group of salinity with (2.0-6.0 PSU) displayed no apparent diversity difference compared with group of salinity with (6.0-9.0 PSU). Moreover, two clusters separation of samples from salinity (14.1 PSU) group was observed, which spread to bigger cluster overlapped with samples of salinity (2.0-6.0 PSU), the smaller cluster overlapped with samples of salinity (14.1 PSU) and samples of salinity (0.2-0.5 PSU). However, the clear difference of microbial community living on microplastics in freshwater and seawater was not well represented by the PCoA diagram.



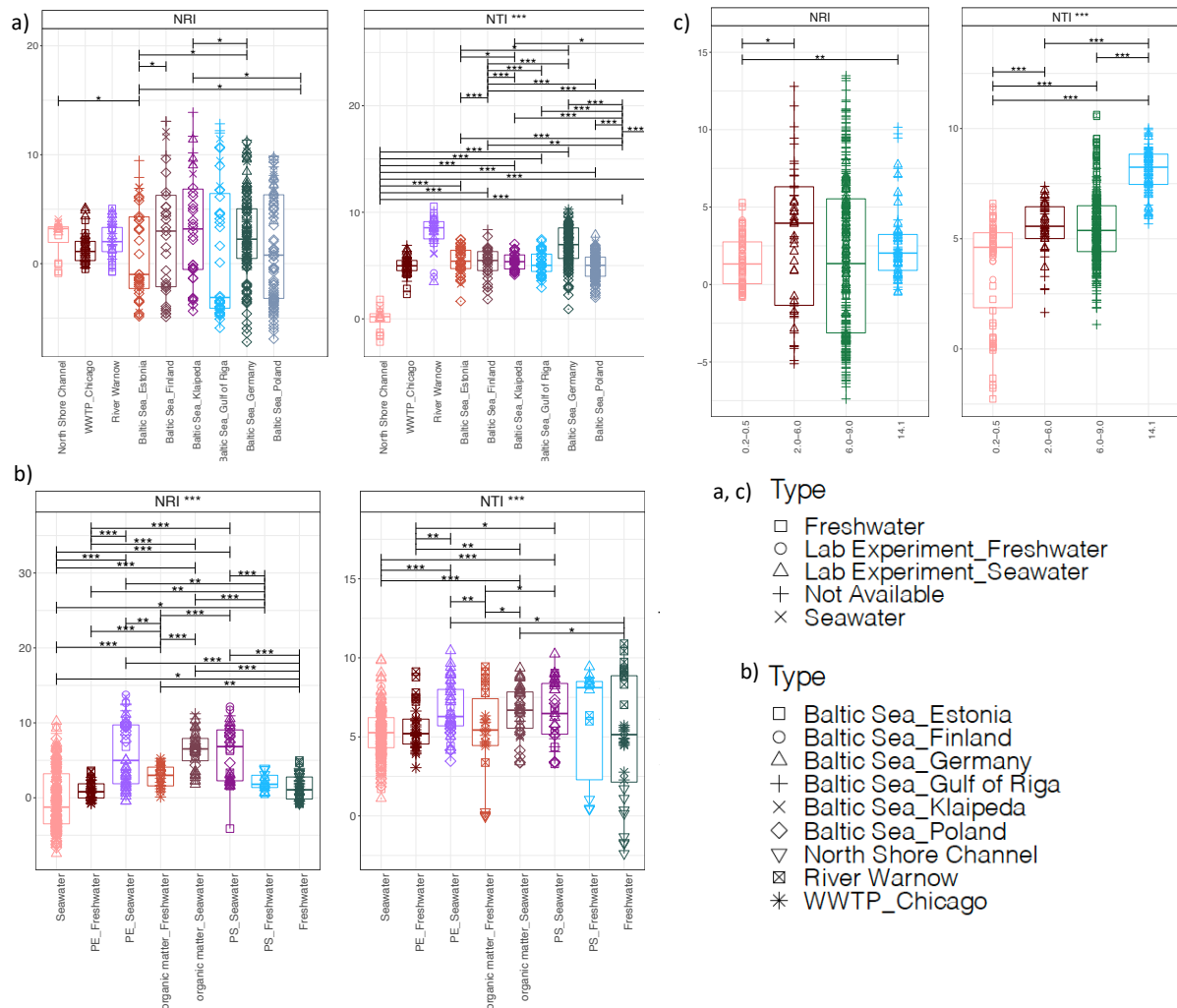
**Figure 6. PCoA of beta diversity was measured for Salinity groups using: (a) weighted UniFrac. (b) Bray-Curtis. (c) Unweighted UniFrac. Two samples if similar lie very close to each other. The ellipses represent the standard error in terms of grouping variations.**

### 3.1.3 Environmental filtering.

Based on the limitation of beta diversity, ecological factors driving the variation of the microbial community were explored further by utilizing NRI/NTI index. PCoD diagrams (Fig. 7) were generated to compare with the ANOVA plots (Fig. 2). NRI/NTI index with positive value shows powerful phylogenetic clustering that was influenced by certain extrinsic parameters. A negative value indicates the weak influence of environmental pressure. *Shannon* Alpha showed the diversity difference within groups of sample location, the variation in diversity was then observed applying NRI/NTI with Bray-Curtis measure. (Fig. 7a) The diversity of Sample from River Warnow (freshwater) and Sample from the Baltic Sea (Germany) was significantly increased. Sample from North Shore Channel (freshwater) with nearly zero value indicated phylogenetic overdispersion. Salinity was charged to be an extrinsic parameter. the diversity of samples with salinity 2.0-6.0 PSU was increased (Fig. 7c) compared with *Shannon* Alpha, and the reduced value on the group with salinity (0.2-0.5 PSU) was also observed. Taken together, the dispersion between *Shannon* alpha and NTI/NRI boxplot was compared relatively, the observed increasing value implied raising environmental influence.



However, other than extrinsic environmental parameters such as Salinity and geographic location, the substrate type may be taken as a significant advantage in affecting diversity. (Fig. 7b) The diversity of Sample PE in seawater and Sample PS in seawater was increased respectively, which implies strong phylogenetic clustering. The diversity of Sample PS in freshwater and Sample PE in freshwater was significantly reduced using NTI. The plastisphere in freshwater with decreasing value of NTI/NRI indicates lower host environmental pressure relate to Plastic material.



**Figure 7. PC\_vs\_OD diagrams (NTI/NRI) by Bray-Curtis: a). Group Classified by: North Shore Channel, Lunz am See, River Danube, WWTP(Chicago), River Warnow, Baltic Sea (multiple locations). (b). Group Classified by: Seawater, PE (Freshwater), PE (Seawater), organic matter (Freshwater), organic matter (Seawater), PS (Seawater), PS(Freshwater), Freshwater c). Group Classified by: Salinity Range (0.2-0.5 PSU, 2.0-6.0 PSU, 6.0-9.0 PSU, 14.1 PSU)**

### 3.1.4 Key drivers of microbial community structure in PERMANOVA

Simultaneous, Permutational Multivariate Analysis of Variance (PERMANOVA) was performed using the permutation test with pseudo-F ratios to see the loss of variation and compare the sources of variation affecting the microbial community structure in order. R2 (quantized how much variation in these distances explained by a variable) and P-value was

found by PERMANOVA, which indicates a significant difference between groups. The P-value for all three parameters displays significant diversity using PERMANOVA. In term of R<sup>2</sup>, the most significant influence is Sample location (R<sup>2</sup> = 0.14323, P = 0.001 (Bray-Curtis); R<sup>2</sup> = 0.10598, P = 0.001 (UniFrac) and R<sup>2</sup> = 0.15744, P = 0.03 (Weighted UniFrac), indicating 15.744% of the variation in Phylogenetic distance is explained by sample location using weighted UniFrac. Using both Bray-Cutis and UniFrac indicates also similar trends the Sample Location taking the greatest influence on community diversity (14.323% and 10.598% respectively). On the other hand, R<sup>2</sup> figure out the second and third key parameters were Sample Type (14.013% variability) and Salinity (7.753% variability) in Weighted UniFrac. A similar trend was indicated using both Bray-Cutis and UniFrac. Therefore, the vague separation between groups in the PCoA plot was supported by observing Table 2. the phylogenetic distance between the centroids was represented by R<sup>2</sup>, which elucidates the Sample Location (freshwater seawater) and Sample Type (plastic-type, when we excluding the reference sample such as organic matter) had the most significant effect on microbial community structure.

|                         | Df | Sum of Sqs | Mean Sqs  | F. Model | R <sup>2</sup> | P <sub>r</sub> | Significance |
|-------------------------|----|------------|-----------|----------|----------------|----------------|--------------|
| <b>Bray_Curtis</b>      |    |            |           |          |                |                |              |
| Sample_Location         | 8  | 29.357     | 3.6696    | 9.3406   | 0.14323        | 0.001          | ***          |
| Sample_Type             | 7  | 23.696     | 3.3851    | 8.3793   | 0.11623        | 0.001          | ***          |
| Salinity_Range          | 3  | 17.965     | 5.9884    | 14.495   | 0.08812        | 0.001          | ***          |
| <b>Weighted Unifrac</b> |    |            |           |          |                |                |              |
| Sample_Location         | 8  | 0.11829    | 0.0147862 | 10.44    | 0.15744        | 0.001          | ***          |
| Sample_Type             | 7  | 0.10516    | 0.0150226 | 10.383   | 0.14013        | 0.001          | ***          |
| Salinity_Range          | 3  | 0.05818    | 0.0193937 | 12.607   | 0.07753        | 0.001          | ***          |
| <b>Unifrac</b>          |    |            |           |          |                |                |              |
| Sample_Location         | 8  | 20.009     | 2.50111   | 6.6235   | 0.10598        | 0.001          | ***          |
| Sample_Type             | 7  | 15.016     | 2.1451    | 5.5375   | 0.07996        | 0.001          | ***          |
| Salinity Range          | 3  | 12.957     | 4.3191    | 11.117   | 0.069          | 0.001          | ***          |

*Table .2 PERMANOVA using different measures in beta diversity analysis*

### **3.2 Composition and structure of plastisphere community.**

The complex community structure of the attached microbial community is able to be revealed from the taxonomic analysis. The top 25% of most genera in each sample were assigned to bacteria. prior to analysis, the samples were identified from eight sample types: freshwater, seawater, organic matter (freshwater/seawater), PE/PS (freshwater/seawater). The differences in bacterial community composition between microplastics in freshwater or seawater were expected, different species of the substrate have specific populations in certain aquatic environments. the microbial community composition on various organic matter (such as wood particle, biofilm, and suspension) represented the difference compared with the plastisphere community.

At the phylum level, the taxa plot (*Appendix II Fig 8.*) illustrated the distributions of microorganism under different medium, the dominant phylum of bacteria detected in all medium were varied. While the abundance of Proteobacteria (80-97% in average) and



*Bacteroidota* (60-80% on average) account for the dominant proportion for all groups, the other main bacteria genera in Seawater were *Cyanobacteria* (30-70% on average), *Planctomycetota* (25-50% in average), *Actinobacteriota* (20-35% on average) and *Verrucomicrobiota* (5-18% in average). The high proportion of *Proteobacteria* in all samples represents this aggregate could be the early colonizers in the aquatic environment. However, *Cyanobacteria*, *Actinobacteriota*, and *Verrucomicrobiota* decreased significantly in the other medium. The samples PE and PS from either freshwater and seawater, Organic matter from either freshwater and seawater, the dominant phylum was *Proteobacteria* and *Bacteroidota*. *Cyanobacteria* (30-70% on average) were the main phylum detected in the Seawater. Therefore, Seawater possessed the greater dominant phylum compared with the other medium. To be more specific, when put the emphasis on samples of microplastic, *Patescibacteria*, and *Actinobacteriota* were more enriched on PE in freshwater. However, the abundance of *Cyanobacteria* on PE from seawater was higher than from freshwater. Besides, *Desulfobacterota* and *Campilobacterota* were more abundant in PS samples in freshwater. Under the visual cue, relatively fewer phyla were represented on the microplastic sample in seawater. Constantly, the difference in the aquatic environment condition may influence the microbial community structure.

At the order level, there were a total of 26 recognized orders represented. (*Appendix II Figure 9*) *Flavobacteriales* (phylum: *Bacteroidota*,) and *Burkholderiales* (phylum: *Proteobacteria*) were more abundant in seawater and freshwater, but *Burkholderiales* in freshwater were more enrich., *Rhodobacterales* (phylum: *Proteobacteria*) in seawater were more abundant, *Chloroplast* (phylum: *Cyanobacteria*) were abundant to found in seawater as well. To compare the community of water and microplastic, *Pseudomonadales* (phylum: *Proteobacteria*) were more enriched on PE and PS in seawater, and *Chitinophagales* (phylum: *Bacteroidota*) were more abundant on PE in freshwater when compare with PE in seawater. At the lower abundance level of microplastic, the abundance of bacteria on microplastic is distributed uniformly. *Alteromonadales* (phylum: *Proteobacteria*) with moderate abundance (about 25-55%) was found on the microplastic sample in seawater.

At the genus level, the dominant taxonomic groups were clearer between groups. (*Appendix II Fig.10*) The significant abundance represented on PE and organic matter in the freshwater group was *Flavobacterium*. Whereas, PE and PS in seawater were highly abundant in *Pseudomonas*. A high proportion of *Chloroplast* on seawater samples was greater than freshwater samples. In addition to the main taxa group compared with seawater and freshwater above, the abundance of other groups on freshwater samples was uniform. The typical bacteria genera such as *Pseudomonas* were highly abundant in microplastic samples. However, *Pseudomonas* genus is fish pathogens and it was one of the sources bacteria that could degrade the plastic polymer. (McCormick et al. 2016). the aquatic animal ingests a large amount of microplastic with pathogens bacteria such as *Pseudomonas* may cause the disease and even affect human health. (McCormick et al. 2016).

In more detail, the difference of microbial community colonizing on microplastic between freshwater and seawater was estimated to be related to complex environmental pressure such as pH and temperature of water. As can be seen from the taxonomic plot at phylum level (*Appendix II Fig. 11a,11b*), the abundant level of taxa represents variation under a distinct range of temperature and pH conditions. The abundance of *Cyanobacteria* decreases gradually from about 75% to 48% as the temperature decreasing from (>20 °C) to (< 15°C) (*Appendix II*

Fig. 11b), and the abundance of *Bacteroidota* also increase from 65% to more than 80% if the temperature of aquatic environment raised from less than 15°C to more than 20 °C. Taking alkaline/acid environment condition into account, the sum of 26 taxa at phylum level were represented on samples. (Appendix II Fig. 11a) In the condition of pH 7.2-7.6 (nearly neutral), a greater number of bacteria phyla with relative higher abundance and a uniform distribution was represented. within the range of pH greater than 8.0, the abundance of *Proteobacteria* and *Bacteroidota* were not influenced to a great extent, respectively, but *Cyanobacteria* become more enriched that possibly was resulted from influence of alkaline environment. More precisely, the main phyla at the most top 25% abundant level of samples was tested by classified by different collecting locations (Appendix II Fig. 11c). The samples were collected from the coastal line of the Baltic Sea was more enriched at proportion of *Cyanobacteria* than freshwater sources.

Taken together, the abundance of *Proteobacteria* (80-97% in average) and *Bacteroidota* (75-90% on average) account for the dominant proportion for all groups at Phylum level. *Patescibacteria* and *Actinobacteriota* on PE were enriched in freshwater, *Desulfobacterota* and *Campilobacterota* on PS were enriched in freshwater as well, but the abundance of *Cyanobacteria* on PS samples was enriched in seawater with more than 70%, which is the third highest abundance phylum within these samples. In the freshwater, there were not a significant advantage of abundance in *Cyanobacteria*. further considering the finer level of taxa. When down finer phylum to genus level, *Flavobacterium* (about 75-80%) was more enriched on PS and PE substrates in freshwater. *Pseudomonas* (about 80-95%) and *Chloroplast* (about 92-95%) distributed higher abundance on PS and PE substrate in seawater.

Through the comparison of beta diversity in three measurements of PCoA plots with different classified groups, a higher abundance of *Chloroplast* in the seawater group and a higher abundance of *Comamonadaceae* in the freshwater group were performed at family level showed a higher degree of variance in diversity. In addition, the heat tree algorithm was supported to express the different microbial community structure of plastisphere between seawater and freshwater, different colored clades indicated the phylogenetic overdispersion between two communities (Appendix II Fig. 12).

## 4 Discussion

### 4.1 The Role of Primary/Secondary in Differentiation between Sample Groups

In this study, the microbial community structure on microplastics and diversity tested by multiple parameters were systematically analysed. The method including alpha/beta diversity, NTI/NRI environmental filtering, TAXAplot, and differential heat tree, the analysis shows the distinct community structure of the Plastisphere in freshwater comparing with those bacteria communities colonizing on microplastic in the marine environment. Few parameters were tested to be the significant factor that clustering or over disperse the phylogenetic distance among the biota community. The availability of nutrients influences biofilm structure and the formation of microplastics (Yang et al., 2020). It means that the differences in composition between sample groups arise from the difference in nutrients. As can be seen from the alpha diversity comparison in the result part, PE samples from different location along the Baltic Sea coastline was picked to test the difference in composition structure in a similar environment, the *Shannon* showed that diversity varied widely may due to their nutrient composition. The abundance of microorganisms (*Patescibacteria*, *Actinobacteria*, *Desulfobacterota*, and *Campilobacterota*) in freshwater could be resulted from nutrient availability. Other factors, such as salinity influencing the bacterial community of microplastics, especially in the seawater. The evidence was by considering the alpha analysis with NTI/NRI, the diversity of groups with salinity 2.0-6.0 PSU was remarkable increased (Fig. 7c) as compared with *Shannon* Alpha (Fig. 2c). The increased salinity and the diminished levels of nutrients in seawater cause the specification of the biofilms that attach to the substrate (Yang et al., 2020). The abundant presence of *Cyanobacteria* in seawater indicates that different environmental factors result in a diversified microbial community in different aquatic environments. Evidence can be found from the alpha diversity with NTI/NRI (Fig. 7b). The diversity of Sample PE/PS in seawater was increased, which implies strong phylogenetic clustering because of the environmental pressure of seawater. The diversity of Sample PE/PS in freshwater was significantly reduced. indicates lower host environmental pressure. Therefore, the primary factors cause the differences in microbial community structure largely account by the spreading of sample location. Consistently, the microorganisms can only thrive well where there are conducive growth conditions, the surrounding water's productivity influences the formation of biofilm on plastic (Kershaw & Rochman, 2015). Secondary, the significant effect of plastic material causes the diversity of microbial community structure was concluded through the application of alpha/beta diversity and PERMANOVA (14.013% variability). Evidently, *Simpson* indicates a clear difference in diversity (\*\*  $p < 0.01$ ) on PS and PE in freshwater, but the results are limited because only two plastic types (PE and PS) were found in metadata.

### 4.2 Hyphomonadaceae in microplastic biofilm

At the family level, PS and PE colonizing communities are made up of many aquatic colonizers. Several families are found on the microplastics biofilm. However, the results showed that *Hyphomonadaceae* was the presence in microplastic biofilms associated to with lower abundance (about 18%). These findings were similar to what Zetter et al. (2013) found in the Sargasso Sea, where it was opined that OTUs associated with these families were plentifully

related to microplastics. Likewise, members of these families formed a part of the primary microbiome that De Tender et al. (2017) defined for biofilms on PE ropes and sheets found incubated at the North Sea seafloor. The *Hyphomonadaceae* are known to be *prosthecae* bacteria that manufacture the polysaccharide holdfast that allows them to attach to the primary colonizers' surfaces. It is further extended by the ability of the *prosthecae* thus enabling them to use nutrients in a wide surrounding area (Oberbeckmann et al., 2018). The *Hyphomonadaceae*'s effective use of nutrients and its capability to attach to smooth surfaces make them able to colonize microplastics.

#### ***4.3 Pseudomonadales, Alteromonadales in the Seawater and Flavobacterium in Freshwater.***

The study established that *Pseudomonadales* and *Alteromonadales* were more enriched on PS and PE in seawater, while the *Flavobacterium* was abundant in the freshwater. The presence of *Pseudomonadales* and *Alteromonadales* in the seawater is associated with the carbon richness of seawater. The two orders belong to the phylum *Proteobacteria* and can carry out nitrite and nitrate degradation thus making them versatile in aerobic and anaerobic environments (Nagaraj et al., 2017). The presence of the two genera in seawater relates to their ability to coexist together thus maintaining an ecological balance and ensuring an optimal substrate utilization for survival in the harsh environment of seawater. The findings on the two suggest that the seawater bacterial community either indirectly or directly influences the ecological structure of the bacterial community structure on microplastics.

On the other hand, *Flavobacterium* is known to proliferate at optimal oxygen and salinity levels, which can be found in freshwater. Therefore, their abundance in freshwater MP biofilm is predictable (Nagaraj et al., 2017). The differences in the plastisphere composition in the seawater and freshwater suggest the significant influence of environmental conditions. Ultimately, the presence of specific bacteria genera enriched on MP biofilms could promote complex ecological problems. Alternations in ecological variables contribute to the changing structure of the bacterial community in both the freshwater and seawater. (Suh et al., 2015).

#### ***4.4 Pathogenic Lead by Plastisphere Colonization of PS and PE***

The study showed that the potentially pathogenic *Proteobacterium* (*Pseudomonadales* and *Alteromonadales*) and *Flavobacterium* were found to be present in the PS and PE biofilm. Similar studies have also indicated the presence of members of the same family in plastic debris at a beach in Scotland (Quilliam et al., 2014) and on floating microplastics in the Baltic Sea (Kirstein et al., 2016). This study indicated that *Pseudomonadales* were more enriched in PS and PE in the seawater while the *Flavobacterium* had a significant abundance. It shows that the presence of these microorganisms has a potential role of selection preference colonizing on the microplastic. The pathogenic colonization of the PE and PS was the most abundant. For example, the *Proteobacteria* (*Pseudomonadales* and *Alteromonadales*) can degrade microplastics and use them as the sole source of carbon (Jacquin et al., 2019). Some *Pseudomonadales* members are known to be opportunistic and can cause such diseases as pneumonia. The PE and PS surfaces provide a more stable surface than the naturally occurring

surfaces of wood or organic matter allow for the abundant formation of these microorganism's biofilms. According to Hoellein et al. (2014), the biofilms formed on microplastics stay intact for a longer period and allow for more extensive transportation to new areas as compared to natural surfaces. *Flavobacterium* has been reported to cause diseases in freshwater fish, and the analysis of this study found their high abundance on the Plastisphere of freshwater. This fact puts the marine animals in danger since they consume the available microplastics. Therefore, the presence of the PE and PS in both the freshwater and seawater provides an avenue for breeding pathogenic microorganisms that thrive well on these surfaces.

#### ***4.5 Complex Carbon Degradation of Microplastics by Bacteria Community***

The biodegradation of microplastics is a process that ends up in the partial or total conversion of the organic carbon into biomass and biogas. Therefore, degradation of microplastic is associated with the activity of the bacteria community that ingest microplastic as the carbon source. However, different conditions found in the freshwater and seawater provide for the type of degradation to occur. The *Pseudomonadales* and *Alteromonadales* found to be abundant in PE and PS can utilize microplastics as sole carbon sources. This degradation occurs alongside the chemical degradation through the weakening of the polymers, evidenced by the molecular changes and roughness of the microplastics (Jacquin et al., 2019). The complex carbon degradation has been explained in four steps by Dussud and Ghiglione (2014). The growing biofilm on microplastics results in the first step that is called biodeterioration, which increases the pore size and weakens the physical properties of microplastics. The second step is bio-fragmentation, which entails the action of extracellular enzymes produced by the bacteria. The enzymes act on the polymers and release oligomers and then monomers that are acted upon by the biofilm cells. The third step is assimilation, which allows the oligomers to be integrated into the cells and be used as a carbon source thus increasing the microbial biomass. The microplastic polymers are broken down into oxidized metabolites through the last step, which is mineralization. This complete biodegradation of the microplastic polymer marks the end of the carbon degradation.

#### ***4.6 Limitation of study***

The meta-analysis in this study contained only four studies. The freshwater sample sets were collected from two studies, and the others two sample sets from seawater (mesocosms and marine litter) compared freshwater type environment to marine type environments. The missing of analysis for one study that data sourced from the freshwater environment was due to the failure of statistical analysis in R. The absence of freshwater samples greatly limits the analysis results and increase the randomness and bias of the test, which caused only small diversity of microbial community was observed between freshwater and seawater. Since this study aims to investigate the difference of the Plastisphere between freshwater and seawater, more samples from different aquatic environments are inspired to collect. However, two sample sets from seawater were all from the Baltic Sea, the similar environmental conditions cause the limitation of analysis when comparing the parameters that influence the microbial community structure. Besides, poor research of the Plastisphere in the freshwater environment at present causes the dilemma of collecting samples. The study (McCormick2014) about freshwater environment provides only 16 samples (including organic matter, water column,

and PS particle), the limited sample numbers in this study raise the limitation of meta-analysis. Therefore, the imperfection of metadata resulted in the differences between microbial community structure in diversity analysis not to be prominent, and less attractive for phylogenetic distance comparison. Moreover, the higher abundance of *Pseudomonadales* was found in both PE samples and organic matters in the seawater environment. The lack of preference identification of *Pseudomonadales* on the seawater plastisphere was due to the limited number of plastic types in the test, more types of microplastic material are inspired to compare for investigating this project. In the analysis of taxa composition, the bacteria community at phylum, order, and genus level were mainly studied, since the taxa figure at OTUs level was not able to provide finer distribution of community structure.

On the other side, the potential problem always exists as the microbial community was described by using the 16S rRNA gene sequence since the taxonomic analysis is limited to be resolved across only the V4 hypervariable region. Also, the methodology of 16S provided a limitation on sequencing analysis. Firstly, although the similarity of rRNA gene sequence support to define the boundary of different genera., the threshold was still unclear on distinguishing between species. Secondly, the overrating of diversity between distinct communities could be made by horizontal gene transfer in the process of amplicon sequencing. (Acinas et al., 2004) Thirdly, the presence of bias and accuracy of amplicon sequencing in the V4 Region. (Hong et al., 2009, Sharpton et al., 2011).

## 5 Conclusion

Based on the diversity analysis and Taxonomic annotation, the study demonstrates the significant role of intrinsic and extrinsic parameters influencing the bacteria community structure in freshwater and seawater. The collection location of samples was performed as the primary factor and the material of microplastics was identified as the secondary factor followed by the *PERMANOVA* test model. In comparison with organic matter, *Pseudomonadales* and *Altermonadales* with a higher prevalence of plastic material in seawater and abundant proportion of *Flavobacterium* in freshwater elucidate the mechanisms of selection for bacteria community assemblage under different environmental pressure. In contrast, the alpha and beta diversity indicates the relative correlation of community structure in freshwater and seawater, the connection between two ecosystems cause the transportation of microplastic with potentially pathogenic bacteria. Accordingly, involve the investigation of the correlation between pathogenic colonization and carbon degradation on microplastic is expected, the weakening of the polymers in the aquatic environment causes roughness, and molecular changes of the surface could inspire the shaping of plastisphere. In the future study, the larger sample size for the plastisphere developing in freshwater will take advantages, and extensive methods are wished to be utilized in comparison with microplastics in seawater.

## References

- Amanda McCormick, Timothy J. Hoellein, Sherri A. Mason, Joseph Schluep and John J. Kelly (2014). Microplastic is an Abundant and Distinct Microbial Habitat in an Urban River. *Environmental Science and Technology*, 48, pp. 11863-11871.
- Acinas, S. G., Marcelino, L. A., Klepac-Ceraj, V., & Polz, M. F. (2004). Divergence and redundancy of 16S rRNA sequences in genomes with multiple *rrn* operons. *Journal of bacteriology*, 186(9), 2629-2635.
- Benjamin J Callahan, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson. & Susan P Holmes (2016). DADA2: High-resolution sample inference from Illumina amplicon data.
- Caroline A. De Tender, Lisa I. Devriese, Annelies Haegeman, Sara Maes, Tom Ruttink and Peter Dawyndt (2015). Bacterial Community Profiling of Plastic Litter in the Belgian Part of the North Sea. *Environmental Science and Technology*, 49, pp. 9629-9638
- Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME journal*, 11(12), 2639-2643.
- Christian Schmidt, Tobias Krauth and Stephan Wagner (2017). Export of Plastic Debris by Rivers into the Sea. *Environmental Science and Technology*, 51, pp. 12246-12253.
- Chong, J., Liu, P., Zhou, G., & Xia, J. (2020). Using MicrobiomeAnalyst for comprehensive statistical, functional, and meta-analysis of microbiome data. *Nature Protocols*, 15, 799-821. <https://doi.org/10.1038/s41596-019-0264-1>
- Dussud, C., & Ghiglione, J. F. (2014). Bacterial degradation of synthetic plastics. In *CIESM Workshop Monogr* (Vol. 46, pp. 49-54).
- Duis, K., & Coors, A. (2016). Microplastics in the aquatic and terrestrial environment: sources (with a specific focus on personal care products), fate and effects. *Environmental Sciences Europe*, 28(1), 2.
- Engler, R. (2012). The Complex Interaction between Marine Debris and Toxic Chemicals in the Ocean. *Environmental Science and Technology*, 46, pp. 12302-12315.
- Erik R. Zettler, Tracy J. Mincer and Linda A. Amaral-Zettler (2013). Life in the “Plastisphere”: Microbial Communities on Plastic Marine Debris. *Environmental Science and Technology*, 47, pp. 7137-7146.
- Foley, B. A. (2018). *Use of ATCC [R] MSA-2002 [TM] for Validation of Extraction and Amplification Techniques in 16S Microbial Community Profiling* (Doctoral dissertation).
- Gao, X., Gao, W., Cui, Z., Han, B., Yang, P., Sun, C., & Zheng, L. (2015). Biodiversity and degradation potential of oil-degrading bacteria isolated from deep-sea sediments of South Mid-Atlantic Ridge. *Marine Pollution Bulletin*, 97(1-2), 373-380.
- Ho, N., Li, F., Wang, S., & Kuhn, L. (2019). MetamicrobiomeR: An R package for analysis of microbiome relative abundance data using zero-inflated beta GAMLSS and meta-analysis



across studies using random effects models. *BMC Bioinformatics*, 20(188). <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2744-2>

HONG, S., BUNGE, J., LESLIN, C., JEON, S. & EPSTEIN, S. S. 2009. Polymerase chain reaction primers miss half of rRNA microbial diversity. *The ISME journal*, 3, 1365-1373.

Hoellein, T., Rojas, M., Pink, A., Gasior, J., & Kelly, J. (2014). Anthropogenic litter in urban freshwater ecosystems: distribution and microbial interactions. *PloS one*, 9(6), e98485.

Ijaz, U. Z., Sivaloganathan, L., McKenna, A., Richmond, A., Kelly, C., Linton, M., ... & Dorrell, N. (2018). Comprehensive longitudinal microbiome analysis of the chicken cecum reveals a shift from competitive to environmental drivers and a window of opportunity for *Campylobacter*. *Frontiers in microbiology*, 9, 2452.

Jacquin, J., Cheng, J., Odobel, C., Pandin, C., Conan, P., Pujo-Pay, M., ... & Ghiglione, J. F. (2019). Microbial ecotoxicology of marine plastic debris: a review on colonization and biodegradation by the "plastisphere" *Frontiers in microbiology*, 10, 865.

Katharina Keszy, Sonja Oberbeckmann, Bernd Kreikemeyer and Matthias Labrenz (2019). Spatial Environmental Heterogeneity Determines Young Biofilm Assemblages on Microplastics in Baltic Sea Mesocosms. *Front. Microbiol.* 10

Koci, O., Logan, M., Svolos, V., Russell, R., Gerasimidis, K., & Ijaz, U. (2018). An automated identification and analysis of ontological terms in gastrointestinal diseases and nutrition-related literature provides useful insights. *PeerJ*. <https://doi.org/10.7717/peerj.5047>

Kathleen Parrish and N. L. Fahrenfeld (2019). Microplastic biofilm in fresh- and wastewater as a function of microparticle type and size class. *Environmental Science: Water Research & Technology*, 5.

Kershaw, P. J., & Rochman, C. M. (2015). Sources, fate, and effects of microplastics in the marine environment: part 2 of a global assessment. *Reports and Studies-IMO/FAO/Unesco-IOC/WMO/IAEA/UN/UNEP Joint Group of Experts on the Scientific Aspects of Marine Environmental Protection (GESAMP) Eng No. 93*.

Kirstein, I. V., Kirmizi, S., Wichels, A., Garin-Fernandez, A., Erler, R., Löder, M. & Gerdt, G. (2016). Dangerous hitchhikers? Evidence for potentially pathogenic *Vibrio* spp. On microplastic particles. *Marine environmental research*, 120, 1-8.

Laist, D. (2006). Overview of the Biological Effects of Lost and Discarded Plastic Debris in the Marine Environment. *Marine Pollution Bulletin*, 18 (6), pp. 319-326.

Line Roager and Eva C. Sonnenschein (2019). Bacterial Candidates for Colonization and Degradation of Marine Plastic Debris. *Environmental Science and Technology*, 53, pp. 11636-11643.

Lucy C. Woodall, Anna D. Jungblut, Kevin Hopkins, Andie Hall, Laura F. Robinson, Claire Gwinnett and Gordon L. J. Paterson (2018). Deep-sea anthropogenic macrodebris harbours rich and diverse communities of bacteria and archaea. *PLoS One*, 13 (11).

- Maria Pinto, Teresa M. Langer, Thorsten Huffer, Thilo Hofmann and Gerhard J. Herndl (2019). The composition of bacterial communities associated with plastic biofilms differs between different polymers and stages of biofilm succession. *PLoS One*, 14 (6)
- Marie Therese Kettner, Keilor Rojas-Jimenez, Sonja Oberbeckmann, Matthias Labrenz and Hans-Peter Grossart (2017). Microplastics alter composition of fungal communities in aquatic ecosystems. *Environmental Microbiology*, 19(11), pp. 4447–4459
- Martin Ogonowski, Asa Motiei, Karolina Ininbergs, Eva Hell, Zandra Gerdes, Klas I. Udekwu, Zoltan Bacsik and Elena Gorokhova (2018). Evidence for selective bacterial community structuring on microplastics. *Environmental Microbiology*, 20(8), pp. 2796–2808
- Meng, Y., Kelly, F. & Wright, S. (2020). Advances and challenges of microplastic pollution in freshwater ecosystems: A UK perspective. *Environmental Pollution*, 256
- McKenna, A., Ijaz, U. Z., Kelly, C., Linton, M., Sloan, W. T., Green, B. D., ... & Corcionivoschi, N. (2020). Impact of industrial production system parameters on chicken microbiomes: mechanisms to improve performance and reduce *Campylobacter*. *bioRxiv*.
- Nagaraj, V., Skillman, L., Ho, G., Li, D., & Gofton, A. (2017). Characterization and comparison of bacterial communities on reverse osmosis membranes of a full-scale desalination plant by bacterial 16S rRNA gene metabarcoding. *NPJ biofilms and microbiomes*, 3(1), 1-14.
- Oberbeckmann, S., Loeder, M. G., Gerds, G., & Osborn, A. M. (2014). Spatial and seasonal variation in diversity and structure of microbial biofilms on marine plastics in Northern European waters. *FEMS microbiology ecology*, 90(2), 478-492.
- Oberbeckmann, S., Kreikemeyer, B., & Labrenz, M. (2018). Environmental factors support the formation of specific bacterial assemblages on microplastics. *Frontiers in Microbiology*, 8, 2709.
- Ogonowski, M., Motiei, A., Ininbergs, K., Hell, E., Gerdes, Z., Udekwu, K. I., ... & Gorokhova, E. (2018). Evidence for selective bacterial community structuring on microplastics. *Environmental microbiology*, 20(8), 2796-2808.
- Quilliam, R. S., Jamieson, J., & Oliver, D. M. (2014). Seaweeds and plastic debris can influence the survival of faecal indicator organisms in beach environments. *Marine pollution bulletin*, 84(1-2), 201-207.
- Reina M. Blair, Susan Waldron, Vernon Phoenix and Caroline Gauchotte-Lindsay (2017). Micro- and Nanoplastic Pollution of Freshwater and Wastewater Treatment Systems. *Springer Science Reviews*, 5, pp. 19–30.
- Roager, L., & Sonnenschein, E. (2019). Bacterial candidates for colonization and degradation of marine plastic debris. *Environmental Science & Technology*, 53, pp. 11636–11643. <https://doi.org/10.1021/acs.est.9b02212>
- Reina M. Blair, Susan Waldron, Vernon R. Phoenix and Caroline Gauchotte-Lindsay (2019). Microscopy and elemental analysis characterisation of microplastics in sediment of a freshwater urban river in Scotland, UK. *Environmental Science and Pollution Research*, 26, pp. 12491–12504.

Sharpton, T. J., Riesenfeld, S. J., Kembel, S. W., Ladau, J., O'Dwyer, J. P., Green, J. L., ... & Pollard, K. S. (2011). PhylOTU: a high-throughput procedure quantifies microbial community diversity and resolves novel taxa from metagenomic data. *PLoS Comput Biol*, 7(1), e1001061.

Sonja Oberbeckmann, Bernd Kreikemeyer and Matthias Labrenz (2018). Environmental Factors Support the Formation of Specific Bacterial Assemblages on Microplastics. *Frontiers in Microbiology*, 8.

Sinclair, L., Ijaz, U., Jensen, L., Coolen, M., Gubry-Rangin, C., Chronakova, A., Oulas, A., Pavloudi, C., Schnetzer, J., Weimann, A., Ijaz, A., Eiler, A., Quince, C., & Pafilis, E. (2016). Seqenv: Linking sequences to environments through text mining. *PeerJ*, 4, e2690. <https://doi.org/10.7717/peerj.2690>

Strungaru, S. A., Jijie, R., Nicoara, M., Plavan, G., & Faggio, C. (2019). Micro-(nano). plastics in freshwater ecosystems: abundance, toxicological impact and quantification methodology. *TrAC Trends in Analytical Chemistry*, 110, 116-128.

Suh, S. S., Park, M., Hwang, J., Kil, E. J., Jung, S. W., Lee, S., & Lee, T. K. (2015). Seasonal dynamics of marine microbial community in the South Sea of Korea. *PLoS One*, 10(6), e0131633.

University of Rhode Island, illumina miseq next generation sequencer (n.d). Retrieved from <https://web.uri.edu/gsc/illumina-miseq-next-generation-sequencer/>

Yang, Y., Liu, W., Zhang, Z., Grossart, H. P., & Gadd, G. M. (2020). Microplastics provide new microbial niches in aquatic environments. *Applied Microbiology and Biotechnology*.

Zimmermann, J., Jahn, R., & Gemeinholzer, B. (2011). Barcoding diatoms: evaluation of the V4 subregion on the 18S rRNA gene, including new primers and protocols. *Organisms Diversity & Evolution*, 11(3), 173.

Zettler, E. R., Mincer, T. J., & Amaral-Zettler, L. A. (2013). Life in the “plastisphere”: microbial communities on plastic marine debris. *Environmental science & technology*, 47(13), 7137-7146.

## *Appendix*

### *Appendix I. Steps for Data collection & Qiime2 analysis with DADA2*

#### 1. Data Collection

Step 1. On Orion cluster, enable both of SRA toolkit and NCBI's e-utils software:

```
export PATH=/home/opt/sratoolkit.2.9.0-centos_linux64/bin:$PATH
export PATH=/home/opt/edirect:$PATH
```

Step 2: Get all the SRR numbers associated with a bioproject PRJNA, download four bioprojects to support meta-analysis:

```
esearch -db sra -query PRJNA***** | efetch --format runinfo | cut -d"," -f 1 >
SRR.numbers
```

Step 3: Retain only the SRR numbers in the files.

```
awk '/SRR/' SRR.numbers > SRR.numbers.filtered
```

Step 4: Use the command in a for loop to retain all SRR numbers into sequences folder:

```
for i in $(cat SRR.numbers.filtered); do echo Processing $i; fastq-dump --split-
files --origfmt --gzip $i ; done
```

#### 2. File Organization

Step 1: Turn the VPN on University of Glasgow and then install CISCO connect. Then use Terminal that internally installed within Mac laptop to connect:

```
ssh MScBioinf@becker.eng.gla.ac.uk
```

Step 2: Move into Caroline folder, and make the project folder and move into the project folder:

```
cd Caroline
mkdir kdl
cd kdl
```

Step 3: Make the sequences folder and move into this folder:

```
mkdir sequences
```

```
cd sequences
```

Step 4: The folder names were extracted from the paired-end files, Place all FASTQ files within corresponding folders. Within each of these folders, dump the raw sequences in a “Raw” folder:

```
for i in $(awk -F"_" '{print $1}'<(ls *.fastq) | sort | uniq); do mkdir $i; mkdir $i/Raw; mv $i*.fastq$i/Raw/.; done
```

### 3. Qiime2 analysis with DADA2

Step 1: Make the folder named “QIIME2\_tutorial” to analyse the data. Then sets path to folder variable.

```
Cd ..  
mkdir qiime2_tutorial;  
d="/home/eng/MScBioinf/umer/sequences/";  
cd qiime2_tutorial;
```

Step 2: Generate fictitious barcode required to import data in Earth Microbiome Project (EMP) format.

Sets a one-line code to view contents of folder

```
t=$(ls $d | wc -l);
```

#Creates fictitious 8 bp barcodes and saves them in a .tsv file.

```
paste <(ls $d) <(perl -le'sub p{my $l=pop @_;unless(@_){return map  
[$_],@$_;}return map { my$ll=$_; map [@$ll,$_],@$_} p(@_);} @a=[A,C,G,T]; print  
join("", @$_)for p(@a,@a,@a,@a,@a,@a,@a,@a);' | awk -v k=$t 'NR<=k{print}') |  
awk 'BEGIN{print "sample-id\tbarcode-sequence\n#q2:types\tcategorical"}1'>  
sample_metadata.tsv
```

Typing `cat sample_metadata.tsv` to list produced sample\_metadata.tsv file:

```
sample-id    barcode-sequence
```

|            |             |
|------------|-------------|
| #q2:types  | categorical |
| SRR1342456 | AAAAAAA     |
| SRR1342458 | AAAAAAC     |
| SRR1342459 | AAAAAAG     |
| SRR1342460 | AAAAAAT     |
| SRR1342461 | AAAAACA     |
| SRR1342462 | AAAAACC     |
| SRR1342463 | AAAAACG     |
| SRR1342464 | AAAAACT     |
| SRR1342465 | AAAAAGA     |
| SRR1342466 | AAAAAGC     |
| SRR1342467 | AAAAAGG     |
| SRR1342468 | AAAAAGT     |
| SRR1342469 | AAAAATA     |
| SRR1342470 | AAAAATC     |
| SRR1342471 | AAAAATG     |
| SRR1342472 | AAAAATT     |
| SRR3669131 | AAAAACA     |
| SRR3669132 | AAAAACAC    |
| SRR3669133 | AAAAACAG    |
| SRR3669134 | AAAAACAT    |
| SRR3669135 | AAAAACCA    |
| SRR3669136 | AAAAACCC    |
| SRR3669137 | AAAAACCG    |
| SRR3669138 | AAAAACCT    |
| SRR3669139 | AAAAACGA    |
| SRR3669140 | AAAAACGC    |
| SRR3669141 | AAAAACGG    |
| SRR3669142 | AAAAACGT    |
| SRR3669143 | AAAAACTA    |
| SRR3669144 | AAAAACTC    |
| SRR3669145 | AAAAACTG    |

SRR3669146 AAAAACTT  
SRR3669147 AAAAAAGAA  
SRR3669148 AAAAAAGAC  
SRR3669149 AAAAAAGAG  
SRR3669150 AAAAAAGAT  
SRR3669151 AAAAAAGCA  
SRR3669152 AAAAAAGCC  
SRR3669153 AAAAAAGCG  
SRR3669154 AAAAAAGCT  
SRR3669155 AAAAAAGGA  
SRR3669156 AAAAAAGGC  
SRR3669157 AAAAAAGGG  
SRR3669158 AAAAAAGGT  
SRR3669159 AAAAAAGTA  
SRR3669160 AAAAAAGTC  
SRR3669161 AAAAAAGTG  
SRR3669162 AAAAAAGTT  
SRR3669163 AAAAAATAA  
SRR3669164 AAAAAATAC  
SRR3669165 AAAAAATAG  
SRR3669166 AAAAAATAT  
SRR3669167 AAAAAATCA  
SRR3669168 AAAAAATCC  
SRR3669169 AAAAAATCG  
SRR3669170 AAAAAATCT  
SRR3669171 AAAAAATGA  
SRR3669172 AAAAAATGC  
SRR3669173 AAAAAATGG  
SRR3669174 AAAAAATGT  
SRR3669175 AAAAAATTA  
SRR3669176 AAAAAATTC  
SRR3669177 AAAAAATTG

SRR3669178 AAAAATTT  
SRR3669179 AAAACAAA  
SRR3669180 AAAACAAC  
SRR3669181 AAAACAAG  
SRR3669182 AAAACAAT  
SRR3669183 AAAACACA  
SRR3669184 AAAACACC  
SRR3669185 AAAACACG  
SRR3669186 AAAAACTT  
SRR3669187 AAAACAGA  
SRR3669188 AAAACAGC  
SRR3669189 AAAACAGG  
SRR3669190 AAAACAGT  
SRR3669191 AAAACATA  
SRR3669192 AAAACATC  
SRR3669193 AAAACATG  
SRR3669194 AAAACATT  
SRR3669195 AAAACCAA  
SRR3669196 AAAACCAC  
SRR3669197 AAAACCAG  
SRR3669198 AAAACCAT  
SRR3669199 AAAACCCA  
SRR3669200 AAAACCCC  
SRR3669201 AAAACCCG  
SRR3669202 AAAAC CCT  
SRR3669203 AAAACCGA  
SRR3669205 AAAACCGC  
SRR3669206 AAAACCGG  
SRR3669207 AAAACCGT  
SRR3669208 AAAACCTA  
SRR3669209 AAAACCTC  
SRR3669210 AAAACCTG



SRR3669211 AAAACCTT  
SRR3669212 AAAACGAA  
SRR3669213 AAAACGAC  
SRR3669214 AAAACGAG  
SRR3669215 AAAACGAT  
SRR3669216 AAAACGCA  
SRR3669217 AAAACGCC  
SRR3669218 AAAACGCG  
SRR3669219 AAAACGCT  
SRR3669220 AAAACGGA  
SRR3669221 AAAACGGC  
SRR3669222 AAAACGGG  
SRR3669223 AAAACGGT  
SRR3669224 AAAACGTA  
SRR3669226 AAAACGTC  
SRR3669227 AAAACGTG  
SRR3669228 AAAACGTT  
SRR3669229 AAAACTAA  
SRR3669230 AAAACTAC  
SRR3669231 AAAACTAG  
SRR3669232 AAAACTAT  
SRR3669233 AAAACTCA  
SRR3669234 AAAACTCC  
SRR3669235 AAAACTCG  
SRR3669236 AAAACTCT  
SRR3669237 AAAACTGA  
SRR3669238 AAAACTGC  
SRR3669239 AAAACTGG  
SRR3669240 AAAACTGT  
SRR3669241 AAAACTTA  
SRR3669242 AAAACTTC  
SRR3669243 AAAACTTG

SRR3669244 AAAACTTT  
SRR3669245 AAAAGAAA  
SRR3669246 AAAAGAAC  
SRR3669247 AAAAGAAG  
SRR3669248 AAAAGAAT  
SRR3669249 AAAAGACA  
SRR3669250 AAAAGACC  
SRR3669251 AAAAGACG  
SRR3669252 AAAAGACT  
SRR3669253 AAAAGAGA  
SRR3669254 AAAAGAGC  
SRR3669255 AAAAGAGG  
SRR3669256 AAAAGAGT  
SRR3669257 AAAAGATA  
SRR3669258 AAAAGATC  
SRR3669259 AAAAGATG  
SRR3669260 AAAAGATT  
SRR3669261 AAAAGCAA  
SRR3669263 AAAAGCAC  
SRR3669264 AAAAGCAG  
SRR3669265 AAAAGCAT  
SRR3669266 AAAAGCCA  
SRR3669267 AAAAGCCC  
SRR3669268 AAAAGCCG  
SRR3669269 AAAAGCCT  
SRR3669270 AAAAGCGA  
SRR3669271 AAAAGCGC  
SRR3669272 AAAAGCGG  
SRR3669273 AAAAGCGT  
SRR3669274 AAAAGCTA  
SRR3669276 AAAAGCTC  
SRR3669277 AAAAGCTG

SRR3669278 AAAAGCTT  
SRR3669279 AAAAGGAA  
SRR3669280 AAAAGGAC  
SRR3669281 AAAAGGAG  
SRR3669282 AAAAGGAT  
SRR3669283 AAAAGGCA  
SRR3669284 AAAAGGCC  
SRR3669285 AAAAGGCG  
SRR3669286 AAAAGGCT  
SRR3669287 AAAAGGGA  
SRR3669288 AAAAGGGC  
SRR3669289 AAAAGGGG  
SRR3669290 AAAAGGGT  
SRR3669291 AAAAGGTA  
SRR3669292 AAAAGGTC  
SRR3669293 AAAAGGTG  
SRR3669294 AAAAGGTT  
SRR3669295 AAAAGTAA  
SRR3669296 AAAAGTAC  
SRR3669297 AAAAGTAG  
SRR3669298 AAAAGTAT  
SRR3669299 AAAAGTCA  
SRR3669300 AAAAGTCC  
SRR3669301 AAAAGTCG  
SRR3669302 AAAAGTCT  
SRR3669303 AAAAGTGA  
SRR3669304 AAAAGTGC  
SRR3669305 AAAAGTGG  
SRR3669306 AAAAGTGT  
SRR3669307 AAAAGTTA  
SRR3669308 AAAAGTTC  
SRR3669309 AAAAGTTG

SRR3669310 AAAAGTTT  
SRR3669311 AAAATAAA  
SRR3669312 AAAATAAC  
SRR3669313 AAAATAAG  
SRR3669314 AAAATAAT  
SRR3669315 AAAATACA  
SRR3669316 AAAATACC  
SRR3669317 AAAATACG  
SRR3669318 AAAATACT  
SRR3669320 AAAATAGA  
SRR3669321 AAAATAGC  
SRR3669322 AAAATAGG  
SRR3669323 AAAATAGT  
SRR3669324 AAAATATA  
SRR3669325 AAAATATC  
SRR4039549 AAAATATG  
SRR4039550 AAAATATT  
SRR4039551 AAAATCAA  
SRR4039552 AAAATCAC  
SRR4039553 AAAATCAG  
SRR4039554 AAAATCAT  
SRR4039555 AAAATCCA  
SRR4039556 AAAATCCC  
SRR4039557 AAAATCCG  
SRR4039558 AAAATCCT  
SRR4039559 AAAATCGA  
SRR4039560 AAAATCGC  
SRR4039561 AAAATCGG  
SRR4039562 AAAATCGT  
SRR4039563 AAAATCTA  
SRR4039564 AAAATCTC  
SRR4039565 AAAATCTG

SRR4039566 AAAATCTT  
SRR4039567 AAAATGAA  
SRR4039568 AAAATGAC  
SRR4039569 AAAATGAG  
SRR4039570 AAAATGAT  
SRR4039571 AAAATGCA  
SRR4039572 AAAATGCC  
SRR4039573 AAAATGCG  
SRR4039574 AAAATGCT  
SRR4039575 AAAATGGA  
SRR4039576 AAAATGGC  
SRR4039577 AAAATGGG  
SRR4039578 AAAATGGT  
SRR4039579 AAAATGTA  
SRR4039580 AAAATGTC  
SRR4039581 AAAATGTG  
SRR4039582 AAAATGTT  
SRR4039583 AAAATTAA  
SRR4039584 AAAATTAC  
SRR4039585 AAAATTAG  
SRR4039586 AAAATTAT  
SRR4039587 AAAATTCA  
SRR4039588 AAAATTCC  
SRR4039589 AAAATTCG  
SRR4039590 AAAATTCT  
SRR4039591 AAAATTGA  
SRR4039592 AAAATTGC  
SRR4039593 AAAATTGG  
SRR4039594 AAAATTGT  
SRR4039595 AAAATTTA  
SRR4039596 AAAATTTC  
SRR4039597 AAAATTTG

SRR4039598 AAAATTTT  
SRR4039599 AAACAAAA  
SRR4039600 AAACAAAC  
SRR4039601 AAACAAAG  
SRR4039602 AAACAAAT  
SRR4039603 AAACAACA  
SRR4039604 AAACAACC  
SRR4039605 AAACAACG  
SRR4039606 AAACAACT  
SRR4039607 AAACAAGA  
SRR4039608 AAACAAGC  
SRR4039609 AAACAAGG  
SRR4039610 AAACAAGT  
SRR4039611 AAACAATA  
SRR4039612 AAACAATC  
SRR4039613 AAACAATG  
SRR4039614 AAACAATT  
SRR4039615 AAACACAA  
SRR4039616 AAACACAC  
SRR4039617 AAACACAG  
SRR4039618 AAACACAT  
SRR4039619 AAACACCA  
SRR4039620 AAACACCC  
SRR4039621 AAACACCG  
SRR4039622 AAACACCT  
SRR4039623 AAACACGA  
SRR4039624 AAACACGC  
SRR4039625 AAACACGG  
SRR4039626 AAACACGT  
SRR4039627 AAACACTA  
SRR4039628 AAACACTC  
SRR4039629 AAACACTG

SRR4039630 AAACACTT  
SRR4039631 AAACAGAA  
SRR4039632 AAACAGAC  
SRR4039633 AAACAGAG  
SRR4039634 AAACAGAT  
SRR4039635 AAACAGCA  
SRR4039636 AAACAGCC  
SRR4039637 AAACAGCG  
SRR4039638 AAACAGCT  
SRR4039639 AAACAGGA  
SRR4039640 AAACAGGC  
SRR4039641 AAACAGGG  
SRR4039642 AAACAGGT  
SRR4039643 AAACAGTA  
SRR4039644 AAACAGTC  
SRR4039645 AAACAGTG  
SRR4039646 AAACAGTT  
SRR4039647 AAACATAA  
SRR4039648 AAACATAC  
SRR4039649 AAACATAG  
SRR4039650 AAACATAT  
SRR4039651 AAACATCA  
SRR4039652 AAACATCC  
SRR4039653 AAACATCG  
SRR4039654 AAACATCT  
SRR4039655 AAACATGA  
SRR4039656 AAACATGC  
SRR4039657 AAACATGG  
SRR4039658 AAACATGT  
SRR4039659 AAACATTA  
SRR4039660 AAACATTC  
SRR4039661 AAACATTG

SRR4039662 AAACATTT  
SRR4039663 AAACCAA  
SRR4039664 AAACCAAC  
SRR4039665 AAACCAAG  
SRR4039666 AAACCAAT  
SRR4039667 AAACCACA  
SRR4039668 AAACCACC  
SRR4039669 AAACCACG  
SRR4039670 AAACCACT  
SRR4039671 AAACCAGA  
SRR4039672 AAACCAGC  
SRR4039673 AAACCAGG  
SRR4039674 AAACCAGT  
SRR4039675 AAACCATA  
SRR4039676 AAACCATC  
SRR4039677 AAACCATG  
SRR4039678 AAACCATT  
SRR4039679 AAACCCAA  
SRR4039680 AAACCCAC  
SRR4039681 AAACCCAG  
SRR4039682 AAACCCAT  
SRR4039683 AAACCCCA  
SRR4039684 AAACCCCC  
SRR4039685 AAACCCCG  
SRR4039686 AAACCCCT  
SRR4039687 AAACCCGA  
SRR4039688 AAACCCGC  
SRR4039689 AAACCCGG  
SRR4039690 AAACCCGT  
SRR4039691 AAACCCTA  
SRR4039692 AAACCCTC  
SRR4039693 AAACCCTG



SRR4039694 AAACCTT  
SRR4039695 AAACCGAA  
SRR4039696 AAACCGAC  
SRR4039697 AAACCGAG  
SRR4039698 AAACCGAT  
SRR5512254 AAACCGCA  
SRR5512255 AAACCGCC  
SRR5626805 AAACCGCG  
SRR5626806 AAACCGCT  
SRR8248029 AAACCGGA  
SRR8248030 AAACCGGC  
SRR8248031 AAACCGGG  
SRR8248032 AAACCGGT  
SRR8248033 AAACCGTA  
SRR8248034 AAACCGTC  
SRR8248035 AAACCGTG  
SRR8248036 AAACCGTT  
SRR8248037 AAACCTAA  
SRR8248038 AAACCTAC  
SRR8248039 AAACCTAG  
SRR8248040 AAACCTAT  
SRR8248041 AAACCTCA  
SRR8248042 AAACCTCC  
SRR8248043 AAACCTCG  
SRR8248044 AAACCTCT  
SRR8248045 AAACCTGA  
SRR8248046 AAACCTGC  
SRR8248047 AAACCTGG  
SRR8248048 AAACCTGT  
SRR8248049 AAACCTTA  
SRR8248050 AAACCTTC  
SRR8248051 AAACCTTG

SRR8248052 AAACCTTT  
SRR8248053 AAACGAAA  
SRR8248054 AAACGAAC  
SRR8248055 AAACGAAG  
SRR8248056 AAACGAAT  
SRR8248057 AAACGACA  
SRR8248058 AAACGACC  
SRR8248059 AAACGACG  
SRR8248060 AAACGACT  
SRR8248061 AAACGAGA  
SRR8248062 AAACGAGC  
SRR8248063 AAACGAGG  
SRR8248064 AAACGAGT  
SRR8248065 AAACGATA  
SRR8248066 AAACGATC  
SRR8248067 AAACGATG  
SRR8248068 AAACGATT  
SRR8248069 AAACGCAA  
SRR8248070 AAACGCAC  
SRR8248071 AAACGCAG  
SRR8248072 AAACGCAT  
SRR8248073 AAACGCCA  
SRR8248074 AAACGCCC  
SRR8248075 AAACGCCG  
SRR8248076 AAACGCCT  
SRR8248077 AAACGCGA  
SRR8248078 AAACGCGC  
SRR8248079 AAACGCGG  
SRR8248080 AAACGCGT  
SRR8248081 AAACGCTA  
SRR8248082 AAACGCTC  
SRR8248083 AAACGCTG

SRR8248084 AAACGCTT  
SRR8248085 AAACGGAA  
SRR8248086 AAACGGAC  
SRR8248087 AAACGGAG  
SRR8248088 AAACGGAT  
SRR8248089 AAACGGCA  
SRR8248090 AAACGGCC  
SRR8248091 AAACGGCG  
SRR8248092 AAACGGCT  
SRR8248093 AAACGGGA  
SRR8248094 AAACGGGC  
SRR8248095 AAACGGGG  
SRR8248096 AAACGGGT  
SRR8248097 AAACGGTA  
SRR8248098 AAACGGTC  
SRR8248099 AAACGGTG  
SRR8248100 AAACGGTT  
SRR8248101 AAACGTAA  
SRR8248102 AAACGTAC  
SRR8248103 AAACGTAG  
SRR8248104 AAACGTAT  
SRR8248105 AAACGTCA  
SRR8248106 AAACGTCC  
SRR8248107 AAACGTCG  
SRR8248108 AAACGTCT  
SRR8248109 AAACGTGA  
SRR8248110 AAACGTGC  
SRR8248111 AAACGTGG  
SRR8248112 AAACGTGT  
SRR8248113 AAACGTTA  
SRR8248114 AAACGTTC  
SRR8248115 AAACGTTG

SRR8248116 AAACGTTT  
SRR8248117 AAATAAA  
SRR8248118 AAATAAC  
SRR8248119 AAATAAG  
SRR8248120 AAATAAT  
SRR8248121 AAATAACA  
SRR8248122 AAATAACC  
SRR8248123 AAATAACG  
SRR8248124 AAATAACT  
SRR8248125 AAATAGA  
SRR8248126 AAATAGC  
SRR8248127 AAATAGG  
SRR8248128 AAATAGT  
SRR8248129 AAATATA  
SRR8248130 AAATATC  
SRR8248131 AAATATG  
SRR8248132 AAATATT  
SRR8248133 AAATCAA  
SRR8248134 AAATCAC  
SRR8248135 AAATCAG  
SRR8248136 AAATCAT  
SRR8248137 AAATCCA  
SRR8248138 AAATCCC  
SRR8248139 AAATCCG  
SRR8248140 AAATCCT  
SRR8248141 AAATCGA  
SRR8248142 AAATCGC  
SRR8248143 AAATCGG  
SRR8248144 AAATCGT  
SRR8248145 AAATCTA  
SRR8248146 AAATCTC  
SRR8248147 AAATCTG

SRR8248148 AACTCTT  
SRR8248149 AACTGAA  
SRR8248150 AACTGAC  
SRR8248151 AACTGAG  
SRR8248152 AACTGAT  
SRR8248153 AACTGCA  
SRR8248154 AACTGCC  
SRR8248155 AACTGCG  
SRR8248156 AACTGCT  
SRR8248157 AACTGGA  
SRR8248158 AACTGGC  
SRR8248159 AACTGGG  
SRR8248160 AACTGGT  
SRR8248161 AACTGTA  
SRR8248162 AACTGTC  
SRR8248163 AACTGTG  
SRR8248164 AACTGTT  
SRR8248165 AACTTAA  
SRR8248166 AACTTAC  
SRR8248167 AACTTAG  
SRR8248168 AACTTAT  
SRR8248169 AACTTCA  
SRR8248170 AACTTCC  
SRR8248171 AACTTCG  
SRR8248172 AACTTCT  
SRR8248173 AACTTGA  
SRR8248174 AACTTGC  
SRR8248175 AACTTGG  
SRR8248176 AACTTGT  
SRR8248177 AACTTTA  
SRR8248178 AACTTTC  
SRR8248179 AACTTTG

SRR8248180 AAAC TTTT  
SRR8248181 AAAG AAAA  
SRR8248182 AAAG AAAC  
SRR8248183 AAAG AAAG  
SRR8248184 AAAG AAAT  
SRR8248185 AAAG AACA  
SRR8248186 AAAG AACC  
SRR8248187 AAAG AACG  
SRR8248188 AAAG AACT  
SRR8248189 AAAG AAGA  
SRR8248190 AAAG AAGC  
SRR8248191 AAAG AAGG  
SRR8248192 AAAG AAGT  
SRR8248193 AAAG AATA  
SRR8248194 AAAG AATC  
SRR8248195 AAAG AATG  
SRR8248196 AAAG AATT  
SRR8248197 AAAG ACAA  
SRR8248198 AAAG ACAC  
SRR8248199 AAAG ACAG  
SRR8248200 AAAG ACAT  
SRR8248201 AAAG ACCA  
SRR8248202 AAAG ACCC  
SRR8248203 AAAG ACCG  
SRR8248204 AAAG ACCT  
SRR8248205 AAAG ACGA  
SRR8248206 AAAG ACGC  
SRR8248207 AAAG ACGG  
SRR8248208 AAAG ACGT  
SRR8248209 AAAG ACTA  
SRR8248210 AAAG ACTC  
SRR8248211 AAAG ACTG

SRR8248212 AAAGACTT  
SRR8248213 AAAGAGAA  
SRR8248214 AAAGAGAC  
SRR8248215 AAAGAGAG  
SRR8248216 AAAGAGAT  
SRR8248217 AAAGAGCA  
SRR8248218 AAAGAGCC  
SRR8248219 AAAGAGCG  
SRR8248220 AAAGAGCT  
SRR8248221 AAAGAGGA  
SRR8248222 AAAGAGGC  
SRR8248223 AAAGAGGG  
SRR8248224 AAAGAGGT  
SRR8248225 AAAGAGTA  
SRR8248226 AAAGAGTC  
SRR8248227 AAAGAGTG  
SRR8248228 AAAGAGTT  
SRR8248229 AAAGATAA  
SRR8248230 AAAGATAC  
SRR8248231 AAAGATAG  
SRR8248232 AAAGATAT  
SRR8248233 AAAGATCA  
SRR8248234 AAAGATCC  
SRR8248235 AAAGATCG  
SRR8248236 AAAGATCT  
SRR8248237 AAAGATGA  
SRR8248238 AAAGATGC  
SRR8248239 AAAGATGG  
SRR8248240 AAAGATGT  
SRR8248241 AAAGATTA  
SRR8248242 AAAGATTC  
SRR8248243 AAAGATTG

SRR8248244 AAAGATTT  
SRR8248245 AAAGCAA  
SRR8248246 AAAGCAAC  
SRR8248247 AAAGCAAG  
SRR8248248 AAAGCAAT  
SRR8248249 AAAGCACA  
SRR8248250 AAAGCACC  
SRR8248251 AAAGCACG  
SRR8248252 AAAGCACT  
SRR8248253 AAAGCAGA  
SRR8248254 AAAGCAGC  
SRR8248255 AAAGCAGG  
SRR8248256 AAAGCAGT  
SRR8248257 AAAGCATA  
SRR8248258 AAAGCATC  
SRR8248259 AAAGCATG  
SRR8248260 AAAGCATT  
SRR8248261 AAAGCCAA  
SRR8248262 AAAGCCAC  
SRR8248263 AAAGCCAG  
SRR8248264 AAAGCCAT  
SRR8248265 AAAGCCCA  
SRR8248266 AAAGCCCC  
SRR8248267 AAAGCCCG  
SRR8248268 AAAGCCCT  
SRR8248269 AAAGCCGA  
SRR8248270 AAAGCCGC  
SRR8248271 AAAGCCGG  
SRR8248272 AAAGCCGT  
SRR8248273 AAAGCCTA  
SRR8248274 AAAGCCTC  
SRR8248275 AAAGCCTG



SRR8248276 AAAGCCTT  
SRR8248277 AAAGCGAA  
SRR8248278 AAAGCGAC  
SRR8248279 AAAGCGAG  
SRR8248280 AAAGCGAT  
SRR8248281 AAAGCGCA  
SRR8248282 AAAGCGCC  
SRR8248283 AAAGCGCG  
SRR8248284 AAAGCGCT  
SRR8248285 AAAGCGGA  
SRR8248286 AAAGCGGC  
SRR8248287 AAAGCGGG  
SRR8248288 AAAGCGGT  
SRR8248289 AAAGCGTA  
SRR8248290 AAAGCGTC  
SRR8248291 AAAGCGTG  
SRR8248292 AAAGCGTT  
SRR8248293 AAAGCTAA  
SRR8248294 AAAGCTAC  
SRR8248295 AAAGCTAG  
SRR8248296 AAAGCTAT  
SRR8248297 AAAGCTCA  
SRR8248298 AAAGCTCC  
SRR8248299 AAAGCTCG  
SRR8248300 AAAGCTCT  
SRR8248301 AAAGCTGA  
SRR8248302 AAAGCTGC  
SRR8248303 AAAGCTGG  
SRR8248304 AAAGCTGT  
SRR8248305 AAAGCTTA  
SRR8248306 AAAGCTTC  
SRR8248307 AAAGCTTG

```

SRR8248308 AAAGCTTT
SRR8248309 AAAGGAAA
SRR8248310 AAAGGAAC
SRR8248311 AAAGGAAG
SRR8248312 AAAGGAAT
SRR8248313 AAAGGACA
SRR8248314 AAAGGACC
SRR8248315 AAAGGACG
SRR8248316 AAAGGACT
SRR8248317 AAAGGAGA
SRR8248318 AAAGGAGC
SRR8248319 AAAGGAGG
SRR8248320 AAAGGAGT
SRR8248321 AAAGGATA
SRR8248322 AAAGGATC
SRR8248323 AAAGGATG
SRR8248324 AAAGGATT
SRR8248325 AAAGGCAA
SRR8248326 AAAGGCAC
SRR8248327 AAAGGCAG
SRR8248328 AAAGGCAT
SRR8248329 AAAGGCCA
SRR.numbers.filtered AAAGGCCG
[MScBioinf@becker ~/Caroline/kdl/qiime2_tutorial]$

```

Step 3: Generate barcodes for each read.

```

(for i in $(ls $d); do bc=$(awk -v k=$i ' $1==k{print $2}' sample_metadata.tsv);
bioawk-cfastx -v k=$bc '{print "@ "$1 "
"$4"\n"k"\n+";for(i=0;i<length(k);i++){printf "#";printf "\n"}'
$d/$i/Raw/*_R1_*.fastq ;done) > barcodes.fastq

```

Extracting the read headers from all the forward FASTQ files, and we assign the barcodes generated from sample\_metadata.tsv file to those headers, the head barcode and tail barcode output as shown:

```
[MScBioinf@becker ~/Caroline/kdl/qiime2_tutorial]$ head barcodes.fastq
@1
AAAAAAA
+
#####
@2
AAAAAAA
+
#####
@3
AAAAAAA
[MScBioinf@becker ~/Caroline/kdl/qiime2_tutorial]$ tail barcodes.fastq
+
#####
@75018
AAAGGCCA
+
#####
@75019
AAAGGCCA
+
#####
```

Step 4: Assemble all the forward reads from all the folders together in a single forward.fastq file:

```
(for i in $(ls $d); do cat$d/$i/Raw/*_R1_*.fastq ; done) > forward.fastq
```

Step 5: Assemble all the reverse reads from all the folders together in a single reverse.fastq file

```
(for i in $(ls $d); do cat$d/$i/Raw/*_R2_*.fastq ; done) > reverse.fastq
```

See if the size of forward, reverse and barcodes FASTQ files could match:

```
bioawk -cfastx 'END{printNR}' forward.fastq
bioawk -cfastx 'END{printNR}' reverse.fastq
bioawk -cfastx 'END{printNR}' barcodes.fastq
```

```
[MScBioinf@becker ~/Caroline/kdl/qiime2_tutorial]$ (for i in $(ls $d); do cat $d/$i/Raw/*_1.fastq ; done) > forward.fastq
cat: /home/eng/MScBioinf/Caroline/kdl/sequences//SRR.numbers/Raw/*_1.fastq: Not a directory
cat: /home/eng/MScBioinf/Caroline/kdl/sequences//SRR.numbers.filtered/Raw/*_1.fastq: Not a directory
[MScBioinf@becker ~/Caroline/kdl/qiime2_tutorial]$ (for i in $(ls $d); do cat $d/$i/Raw/*_2.fastq ; done) > reverse.fastq
cat: /home/eng/MScBioinf/Caroline/kdl/sequences//SRR.numbers/Raw/*_2.fastq: Not a directory
cat: /home/eng/MScBioinf/Caroline/kdl/sequences//SRR.numbers.filtered/Raw/*_2.fastq: Not a directory
[MScBioinf@becker ~/Caroline/kdl/qiime2_tutorial]$ bioawk -cfastx 'END{print NR}' forward.fastq
74021753
[MScBioinf@becker ~/Caroline/kdl/qiime2_tutorial]$ bioawk -cfastx 'END{print NR}' reverse.fastq
74021753
[MScBioinf@becker ~/Caroline/kdl/qiime2_tutorial]$ bioawk -cfastx 'END{print NR}' barcodes.fastq
74021753
```

Step 6: Make the folder named “emp-paired-end-sequences”, then zip all the FASTQ files and move them to the folder.

```
gzip *.fastq
```

```
mkdir emp-paired-end-sequences; mv *.gz emp-paired-endsequences/.
```

```
[MScBioinf@becker ~/Caroline/kdl/qiime2_tutorial]$ cd emp-paired-end-sequences
[MScBioinf@becker ~/Caroline/kdl/qiime2_tutorial/emp-paired-end-sequences]$ ls
barcodes.fastq.gz forward.fastq.gz reverse.fastq.gz
```

Next, activate Qiime2 (Version 2019.7) on the Orion cluster:

```
export PATH=/home/opt/miniconda2/bin:$PATH;
```

```
source activate qiime2-2019.7;
```

Step 7. Import the zipped sequences files in “emp-paired-end-sequences” folder to Qiime2 platform.

```
qiime tools import --type EMPPairedEndSequences --input-path emp-paired-end-sequences --output-path emp-paired-end-sequences.qza
```

Step 8: Demultiplex the sequences in Qiime2:

```
qiimedemux emp-paired --p-no-golay-error-correction --i-seqsemp-paired-end-sequences.qza --m-barcodes-file sample_metadata.tsv--m-barcodes-column barcode-sequence --o-per-sample-sequencesdemux.qza --o-error-correction-details demux-details.qza
```

Step 9: In order to tune Dada2 algorithm by specifying the thresholds, (1) the file format of Demultiplexed sequences then were converted from .qza file into .qzv file. The new format is visualizable, (2) export and (3) download it to local computer, drag and drop the file to Qiime2 viewer website (<https://view.qiime2.org/>) and manually figure out the thresholds.

```
qiimedemux summarize --i-data ./demux.qza --o-visualization ./demux.qzv
```

```
qiime tools export --input-path demux.qzv --output-path output
```

Scp [MScBioinf@becker.eng.gla.ac.uk:~/umer/qiime2 tutorial/demux.qzv](mailto:MScBioinf@becker.eng.gla.ac.uk:~/umer/qiime2_tutorial/demux.qzv)

Figure 1.(B) (C), The y axis for is the quality score and x-axis is the number of base pairs - trimming the ends that have a quality score of less than 30-25. In this case, trimmed the forward at 200, everything after 200 is discarded and trimmed the reverse at 180, everything after 180 is discarded.

Output Figure 1. (A),(B),(C) (D) from Qiime2 viewer:

|          |                    |
|----------|--------------------|
| Minimum: | 22                 |
| Median:  | 102442.0           |
| Mean:    | 111984.49773071104 |
| Maximum: | 399235             |
| Total:   | 74021753           |

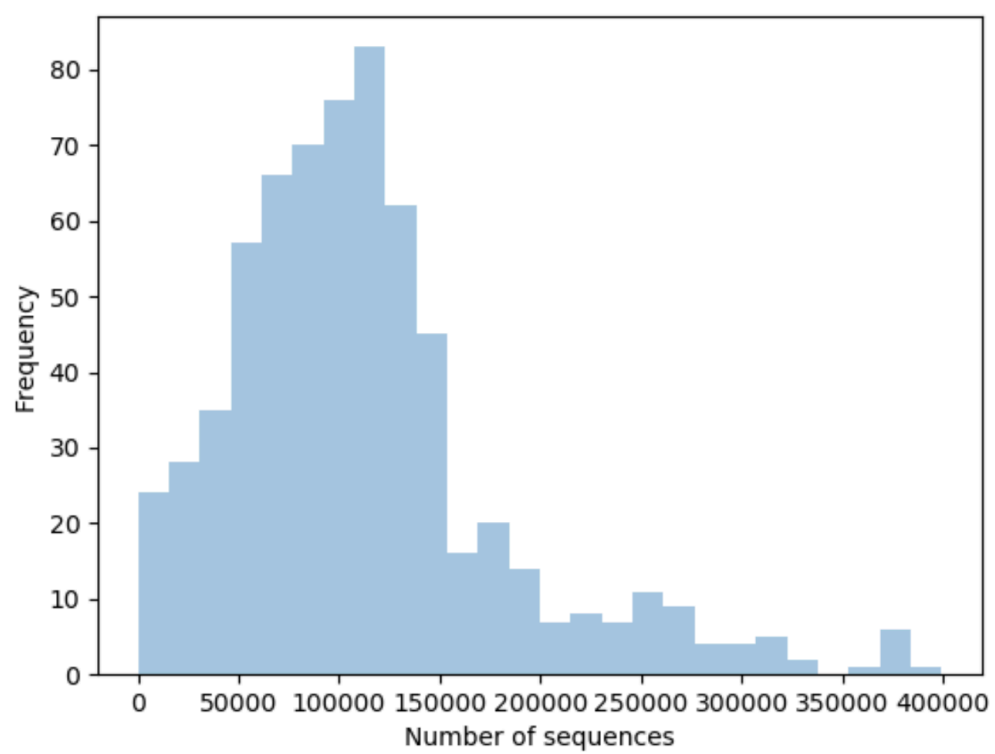
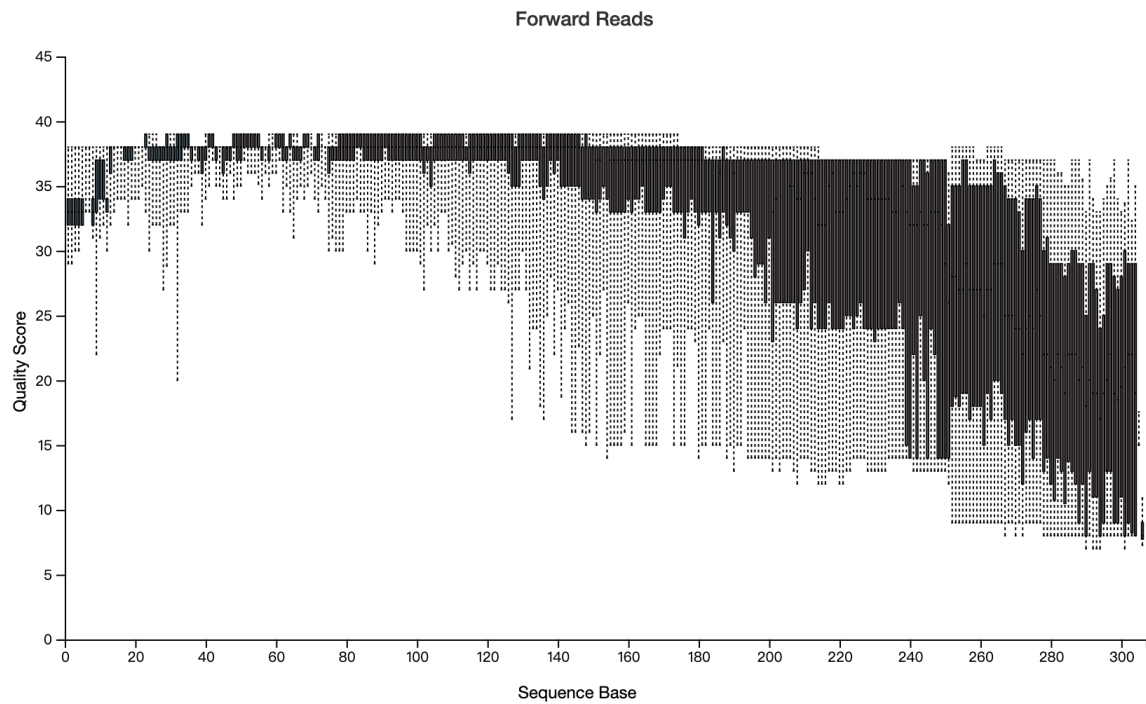


Figure 1. (A) Demultiplexed sequence counts summary



|         |        |         |        |         |        |         |        |         |        |         |        |         |        |
|---------|--------|---------|--------|---------|--------|---------|--------|---------|--------|---------|--------|---------|--------|
| 039657  | 399235 | SRR4039 | 228551 | SRR4039 | 168621 | SRR3669 | 139814 | SRR4039 | 127662 | SRR8248 | 117914 | SRR8248 | 108892 |
| 039680  | 379142 | SRR4039 | 226857 | SRR8248 | 167161 | SRR8248 | 139648 | SRR8248 | 126881 | SRR8248 | 117901 | SRR8248 | 108760 |
| 039671  | 378905 | SRR3669 | 226189 | SRR1342 | 166231 | SRR8248 | 139575 | SRR4039 | 126728 | SRR4039 | 117714 | SRR8248 | 108428 |
| 039675  | 377322 | SRR4039 | 223108 | SRR1342 | 165407 | SRR8248 | 139407 | SRR4039 | 126726 | SRR8248 | 117532 | SRR4039 | 107590 |
| 039682  | 375253 | SRR3669 | 221861 | SRR8248 | 162936 | SRR8248 | 139254 | SRR8248 | 126719 | SRR8248 | 117518 | SRR8248 | 107577 |
| 039674  | 373303 | SRR4039 | 221194 | SRR8248 | 162542 | SRR8248 | 139231 | SRR8248 | 126287 | SRR8248 | 116930 | SRR8248 | 107331 |
| 039685  | 369469 | SRR4039 | 218460 | SRR1342 | 161955 | SRR8248 | 139091 | SRR8248 | 126212 | SRR8248 | 116823 | SRR8248 | 107014 |
| 039672  | 364204 | SRR5512 | 216530 | SRR8248 | 159881 | SRR3669 | 138645 | SRR4039 | 126125 | SRR4039 | 116657 | SRR8248 | 106747 |
| 039660  | 325186 | SRR4039 | 213461 | SRR4039 | 158536 | SRR8248 | 138478 | SRR8248 | 126055 | SRR4039 | 116541 | SRR3669 | 106268 |
| 039621  | 322840 | SRR1342 | 212113 | SRR3669 | 158413 | SRR8248 | 138424 | SRR4039 | 126001 | SRR8248 | 116528 | SRR8248 | 106260 |
| 039681  | 320882 | SRR5512 | 211359 | SRR8248 | 157760 | SRR4039 | 138019 | SRR3669 | 125727 | SRR8248 | 116515 | SRR8248 | 106123 |
| 039659  | 319312 | SRR4039 | 211177 | SRR4039 | 156369 | SRR8248 | 137820 | SRR8248 | 125545 | SRR8248 | 116346 | SRR4039 | 106090 |
| 039698  | 316166 | SRR4039 | 211016 | SRR3669 | 156104 | SRR8248 | 137694 | SRR4039 | 125376 | SRR4039 | 115735 | SRR8248 | 106061 |
| 039632  | 312174 | SRR1342 | 207709 | SRR8248 | 155140 | SRR8248 | 137289 | SRR8248 | 125369 | SRR8248 | 115481 | SRR3669 | 105669 |
| 039665  | 307242 | SRR4039 | 207376 | SRR8248 | 154209 | SRR4039 | 136278 | SRR4039 | 124903 | SRR8248 | 115421 | SRR3669 | 105660 |
| 039554  | 306431 | SRR1342 | 198582 | SRR8248 | 151884 | SRR8248 | 136263 | SRR8248 | 124670 | SRR4039 | 115404 | SRR8248 | 105652 |
| 669316  | 303665 | SRR4039 | 197199 | SRR3669 | 151860 | SRR8248 | 136228 | SRR4039 | 124636 | SRR3669 | 115339 | SRR4039 | 105622 |
| 039552  | 296986 | SRR1342 | 194742 | SRR8248 | 151843 | SRR3669 | 136075 | SRR8248 | 124148 | SRR3669 | 115180 | SRR3669 | 105532 |
| 039676  | 295909 | SRR4039 | 194184 | SRR8248 | 150786 | SRR8248 | 135922 | SRR8248 | 123975 | SRR8248 | 114188 | SRR8248 | 105351 |
| 039694  | 289041 | SRR1342 | 193084 | SRR8248 | 149958 | SRR4039 | 135691 | SRR3669 | 123801 | SRR4039 | 114144 | SRR8248 | 105282 |
| 039619  | 288674 | SRR3669 | 192954 | SRR8248 | 149687 | SRR4039 | 135072 | SRR8248 | 123714 | SRR8248 | 113926 | SRR3669 | 104646 |
| 039679  | 285588 | SRR1342 | 192309 | SRR3669 | 149549 | SRR8248 | 134697 | SRR8248 | 123571 | SRR3669 | 113638 | SRR3669 | 104517 |
| 039641  | 282157 | SRR4039 | 191629 | SRR8248 | 149509 | SRR8248 | 134461 | SRR8248 | 122732 | SRR8248 | 113561 | SRR8248 | 104399 |
| 039550  | 275265 | SRR4039 | 190560 | SRR8248 | 148997 | SRR3669 | 134300 | SRR3669 | 122693 | SRR8248 | 113184 | SRR8248 | 103745 |
| 039658  | 274564 | SRR1342 | 190017 | SRR8248 | 148854 | SRR8248 | 133973 | SRR8248 | 122639 | SRR8248 | 113139 | SRR8248 | 103349 |
| 039646  | 273255 | SRR4039 | 189369 | SRR3669 | 148620 | SRR3669 | 133709 | SRR4039 | 122493 | SRR8248 | 113120 | SRR8248 | 103301 |
| 039651  | 272736 | SRR1342 | 189328 | SRR4039 | 147888 | SRR8248 | 133703 | SRR8248 | 122294 | SRR4039 | 112991 | SRR4039 | 103245 |
| 039667  | 272546 | SRR4039 | 185968 | SRR8248 | 147460 | SRR4039 | 133517 | SRR8248 | 121836 | SRR8248 | 112984 | SRR3669 | 103123 |
| 039647  | 271616 | SRR3669 | 184845 | SRR8248 | 147218 | SRR4039 | 133274 | SRR8248 | 121632 | SRR8248 | 112734 | SRR8248 | 102961 |
| 039653  | 269687 | SRR1342 | 183048 | SRR8248 | 146089 | SRR8248 | 133160 | SRR8248 | 121074 | SRR8248 | 112194 | SRR8248 | 102450 |
| 039639  | 265712 | SRR4039 | 182761 | SRR8248 | 145415 | SRR8248 | 132870 | SRR8248 | 120818 | SRR8248 | 112101 | SRR8248 | 102442 |
| 039642  | 262958 | SRR8248 | 182647 | SRR8248 | 144930 | SRR8248 | 132857 | SRR8248 | 120487 | SRR8248 | 112064 | SRR4039 | 101639 |
| 039649  | 260191 | SRR4039 | 181785 | SRR8248 | 144843 | SRR4039 | 132816 | SRR8248 | 120295 | SRR8248 | 112060 | SRR8248 | 101578 |
| 039656  | 259750 | SRR1342 | 179873 | SRR8248 | 144715 | SRR8248 | 132330 | SRR8248 | 120196 | SRR3669 | 111769 | SRR3669 | 101551 |
| 248076  | 258908 | SRR4039 | 179169 | SRR8248 | 144458 | SRR3669 | 132290 | SRR8248 | 120093 | SRR8248 | 111726 | SRR8248 | 101512 |
| 039620  | 258655 | SRR4039 | 178749 | SRR8248 | 144004 | SRR8248 | 132228 | SRR4039 | 120048 | SRR4039 | 111582 | SRR8248 | 101072 |
| 039662  | 253904 | SRR1342 | 178722 | SRR8248 | 143522 | SRR8248 | 132010 | SRR8248 | 120038 | SRR4039 | 111373 | SRR8248 | 100897 |
| 039633  | 253482 | SRR3669 | 178650 | SRR8248 | 143313 | SRR8248 | 131627 | SRR8248 | 119849 | SRR8248 | 111341 | SRR4039 | 100763 |
| 039645  | 253023 | SRR4039 | 178323 | SRR3669 | 142773 | SRR8248 | 130746 | SRR8248 | 119584 | SRR8248 | 111218 | SRR8248 | 100606 |
| 039625  | 250365 | SRR1342 | 178223 | SRR8248 | 142361 | SRR8248 | 130700 | SRR8248 | 119541 | SRR8248 | 111090 | SRR8248 | 100421 |
| 039644  | 247345 | SRR8248 | 178075 | SRR8248 | 142308 | SRR3669 | 130366 | SRR8248 | 119446 | SRR3669 | 111013 | SRR4039 | 100078 |
| 039669  | 247225 | SRR4039 | 176713 | SRR8248 | 142116 | SRR8248 | 130234 | SRR8248 | 119383 | SRR3669 | 110962 | SRR4039 | 99884  |
| 039652  | 246215 | SRR8248 | 174213 | SRR8248 | 142084 | SRR8248 | 130180 | SRR8248 | 119378 | SRR4039 | 110857 | SRR3669 | 99289  |
| 039677  | 245450 | SRR1342 | 171405 | SRR3669 | 142013 | SRR3669 | 129810 | SRR8248 | 119340 | SRR4039 | 110668 | SRR4039 | 99960  |
| 039689  | 243022 | SRR4039 | 171214 | SRR8248 | 142001 | SRR3669 | 129764 | SRR3669 | 119235 | SRR8248 | 110406 | SRR3669 | 98865  |
| 039661  | 238447 | SRR3669 | 170768 | SRR3669 | 141796 | SRR4039 | 129282 | SRR4039 | 119078 | SRR3669 | 110339 | SRR3669 | 98838  |
| 039640  | 238242 | SRR3669 | 169515 | SRR8248 | 140615 | SRR4039 | 128833 | SRR8248 | 118812 | SRR4039 | 110148 | SRR8248 | 98610  |
| 039627  | 238025 | SRR4039 | 169451 | SRR8248 | 140149 | SRR8248 | 128661 | SRR3669 | 118436 | SRR8248 | 109450 | SRR8248 | 98356  |
| 039638  | 237404 | SRR4039 | 169325 | SRR8248 | 139954 | SRR8248 | 128604 | SRR8248 | 118254 | SRR8248 | 109446 | SRR8248 | 98085  |
| 248154  | 231104 | SRR4039 | 168696 | SRR8248 | 139930 | SRR4039 | 128303 | SRR8248 | 118010 | SRR8248 | 109339 | SRR3669 | 97636  |
| SRR4039 | 97511  | SRR8248 | 87394  | SRR4039 | 76952  | SRR8248 | 66638  | SRR8248 | 54752  | SRR3669 | 33631  | SRR3669 | 3937   |
| SRR8248 | 97483  | SRR8248 | 87358  | SRR3669 | 76530  | SRR8248 | 66279  | SRR8248 | 54640  | SRR3669 | 33501  | SRR3669 | 3242   |
| SRR3669 | 97343  | SRR8248 | 87030  | SRR4039 | 76353  | SRR8248 | 66223  | SRR8248 | 54608  | SRR3669 | 33140  | SRR3669 | 3134   |
| SRR3669 | 97248  | SRR4039 | 86538  | SRR8248 | 76291  | SRR8248 | 66111  | SRR3669 | 54499  | SRR3669 | 32729  | SRR3669 | 2377   |
| SRR3669 | 96883  | SRR8248 | 86527  | SRR3669 | 76172  | SRR8248 | 65955  | SRR4039 | 54296  | SRR3669 | 32249  | SRR3669 | 2133   |
| SRR3669 | 96755  | SRR8248 | 85327  | SRR8248 | 75897  | SRR8248 | 65809  | SRR8248 | 54265  | SRR3669 | 31964  | SRR5626 | 1928   |
| SRR8248 | 96750  | SRR3669 | 85248  | SRR8248 | 75476  | SRR8248 | 65488  | SRR8248 | 54167  | SRR3669 | 31784  | SRR3669 | 1772   |
| SRR8248 | 96486  | SRR8248 | 84883  | SRR8248 | 75471  | SRR8248 | 65006  | SRR3669 | 54155  | SRR3669 | 31723  | SRR3669 | 1540   |
| SRR8248 | 96401  | SRR8248 | 84238  | SRR8248 | 75444  | SRR3669 | 64727  | SRR3669 | 53580  | SRR3669 | 30880  | SRR8248 | 1443   |
| SRR8248 | 95982  | SRR3669 | 84180  | SRR8248 | 75019  | SRR8248 | 64365  | SRR3669 | 52595  | SRR3669 | 30551  | SRR5626 | 418    |
| SRR8248 | 95777  | SRR8248 | 83488  | SRR4039 | 74722  | SRR8248 | 64123  | SRR8248 | 52208  | SRR3669 | 30373  | SRR3669 | 22     |
| SRR8248 | 95768  | SRR4039 | 83248  | SRR8248 | 74267  | SRR8248 | 63431  | SRR3669 | 51558  | SRR4039 | 29719  |         |        |
| SRR3669 | 95728  | SRR4039 | 83184  | SRR8248 | 74202  | SRR8248 | 63260  | SRR3669 | 50928  | SRR3669 | 29269  |         |        |
| SRR3669 | 95531  | SRR8248 | 82586  | SRR3669 | 74125  | SRR8248 | 62842  | SRR8248 | 50517  | SRR3669 | 28775  |         |        |
| SRR3669 | 95110  | SRR8248 | 82391  | SRR3669 | 73955  | SRR8248 | 62814  | SRR4039 | 50205  | SRR3669 | 28494  |         |        |
| SRR8248 | 94935  | SRR4039 | 82311  | SRR3669 | 73761  | SRR8248 | 62667  | SRR3669 | 49695  | SRR3669 | 28291  |         |        |
| SRR4039 | 94525  | SRR8248 | 82310  | SRR8248 | 73315  | SRR4039 | 61623  | SRR3669 | 49153  | SRR8248 | 26839  |         |        |
| SRR3669 | 94523  | SRR8248 | 82222  | SRR4039 | 73057  | SRR3669 | 60867  | SRR8248 | 48899  | SRR3669 | 26305  |         |        |
| SRR8248 | 94219  | SRR4039 | 82092  | SRR3669 | 72965  | SRR3669 | 60763  | SRR8248 | 48565  | SRR3669 | 26296  |         |        |
| SRR8248 | 94117  | SRR3669 | 82066  | SRR8248 | 72736  | SRR4039 | 60558  | SRR3669 | 48517  | SRR3669 | 26012  |         |        |
| SRR3669 | 94072  | SRR3669 | 82012  | SRR3669 | 72710  | SRR3669 | 60509  | SRR3669 | 47887  | SRR3669 | 25236  |         |        |
| SRR4039 | 94052  | SRR8248 | 81567  | SRR3669 | 72701  | SRR3669 | 60375  | SRR3669 | 47544  | SRR3669 | 24576  |         |        |
| SRR8248 | 94050  | SRR8248 | 80897  | SRR3669 | 72681  | SRR3669 | 60152  | SRR3669 | 46754  | SRR3669 | 23525  |         |        |
| SRR8248 | 93928  | SRR3669 | 80829  | SRR8248 | 72598  | SRR3669 | 59892  | SRR8248 | 46167  | SRR3669 | 23378  |         |        |
| SRR4039 | 93450  | SRR8248 | 80748  | SRR8248 | 72445  | SRR8248 | 59598  | SRR8248 | 45963  | SRR3669 | 22664  |         |        |
| SRR8248 | 93434  | SRR3669 | 80696  | SRR8248 | 72377  | SRR8248 | 59460  | SRR4039 | 45259  | SRR3669 | 21689  |         |        |
| SRR8248 | 93201  | SRR8248 | 80580  | SRR4039 | 72108  | SRR8248 | 59007  | SRR3669 | 44699  | SRR3669 | 20674  |         |        |
| SRR8248 | 93142  | SRR8248 | 80455  | SRR3669 | 71532  | SRR3669 | 58940  | SRR3669 | 44110  | SRR3669 | 19603  |         |        |
| SRR8248 | 93005  | SRR4039 | 80366  | SRR8248 | 71172  | SRR3669 | 58771  | SRR3669 | 44077  | SRR3669 | 18848  |         |        |
| SRR3669 | 92578  | SRR4039 | 80193  | SRR8248 | 71172  | SRR4039 | 58560  | SRR8248 | 44066  | SRR3669 | 18802  |         |        |
| SRR4039 | 92229  | SRR8248 | 80057  | SRR8248 | 70935  | SRR3669 | 58547  | SRR3669 | 43124  | SRR3669 | 18109  |         |        |
| SRR3669 | 91760  | SRR8248 | 79957  | SRR8248 | 70785  | SRR8248 | 58446  | SRR4039 | 42704  | SRR8248 | 18104  |         |        |
| SRR4039 | 91444  | SRR8248 | 79913  | SRR8248 | 70533  | SRR3669 | 58278  | SRR3669 | 42293  | SRR3669 | 17626  |         |        |
| SRR3669 | 91318  | SRR8248 | 79578  | SRR3669 | 70511  | SRR8248 | 58005  | SRR3669 | 42030  | SRR8248 | 17474  |         |        |
| SRR8248 | 91125  | SRR8248 | 79565  | SRR8248 | 70324  | SRR3669 |        |         |        |         |        |         |        |

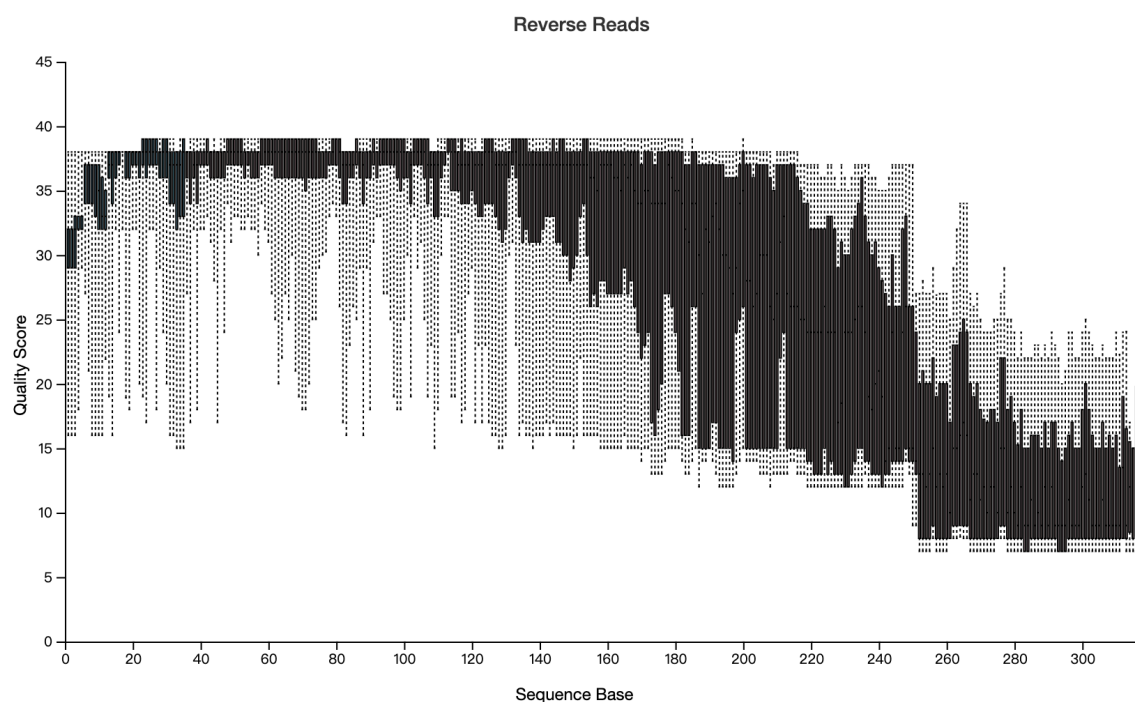


The plot at position 304 was generated using a random sampling of 9 out of 74021753 sequences without replacement. This position (304) is greater than the minimum sequence length observed during subsampling (35 bases). As a result, the plot at this position is not based on data from all of the sequences, so it should be interpreted with caution when compared to plots for other positions. Outlier quality scores are not shown in box plots for clarity.

| Parametric seven-number summary for <b>position 304</b> |               |               |
|---------------------------------------------------------|---------------|---------------|
| Box plot feature                                        | Percentile    | Quality score |
| (Not shown in box plot)                                 | 2nd           | 8             |
| Lower Whisker                                           | 9th           | 8             |
| Bottom of Box                                           | 25th          | 8             |
| Middle of Box                                           | 50th (Median) | 19            |
| Top of Box                                              | 75th          | 29            |
| Upper Whisker                                           | 91st          | 35            |
| (Not shown in box plot)                                 | 98th          | 35            |

Figure 1. (C) Quality plot of forward paired-end sequencing reads





These plots were generated using a random sampling of 10000 out of 74021753 sequences without replacement. The minimum sequence length identified during subsampling was 35 bases. Outlier quality scores are not shown in box plots for clarity.

| Parametric seven-number summary |               |               |
|---------------------------------|---------------|---------------|
| Box plot feature                | Percentile    | Quality score |
| (Not shown in box plot)         | 2nd           | ...           |
| Lower Whisker                   | 9th           | ...           |
| Bottom of Box                   | 25th          | ...           |
| Middle of Box                   | 50th (Median) | ...           |
| Top of Box                      | 75th          | ...           |
| Upper Whisker                   | 91st          | ...           |
| (Not shown in box plot)         | 98th          | ...           |

Figure 1. (D) Quality plot of reverse paired-end sequencing reads

Step 10: Run Dada2 algorithm to make the quality control step.

```
qiimedada2 denoise-paired --i-demultiplexed-seqs demux.qza --p-trim-left-f0--p-trim-left-r 0--p-trunc-len-f 200--p-trunc-len-r 180--p-n-threads 0 --o-table table.qza --o-representative-sequencesrep-seqs.qza --o-denoising-stats denoising-stats.qza --verbose
```

## Step 11: Generate the phylogenetic tree for the ASVs:

Unset MAFFT\_BINARIES;

```
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 --p-n-threads 0 --o-tree unrooted-tree.qza --o-rooted-tree
rooted-tree.qza ;
```

## Step 12: Create a taxonomy for the ASVs:

```
qiime feature-classifier classify-sklearn -i classifier/software
/qiime2_databases/silva-132-99-nb-classifier.qza --i-reads rep-
seqs.qza --o-classification taxonomy.qza
```

The created table shows for each ASV that correspond to taxonomy: (Figure 2)

| Feature ID                        | taxonomy                                                                                                                                     |
|-----------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------|
| b744eae1244325f606575483d0de67e   | d__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__SAR11_clade; f__Clade_III; g__Clade_III; s__Candidatus_Fonsibacter                |
| 08bd45c0807b6fb44faa54fb53e4213   | d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Burkholderiales; f__Rhodocyclaceae; g__Zoogloea                                   |
| 38c27ceaed634984c1225a82848cf571  | d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pseudomonadales; f__Pseudomonadaceae; g__Pseudomonas                              |
| 9c7e018344e95c9245cbe67238f23c8   | d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pseudomonadales; f__Pseudomonadaceae; g__Pseudomonas                              |
| 9539a6a644963195a48b457512f36720  | d__Bacteria; p__Cyanobacteria; c__Cyanobacteria; o__Cyanobacteriales; f__Microcystaceae; g__Synechocystis_PCC-6803; s__uncultured_bacterium  |
| 39853883f6e116ee7f781c5c2260a0049 | d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Thiotrichales; f__Thiotrichaceae; g__Thiothrix                                    |
| 3f456a18968cea5c5e549e281f0f5808  | d__Bacteria; p__Actinobacteriota; c__Actinobacteria; o__Frankiales; f__Sporichthyaceae; g__hgcl_clade                                        |
| 41d7f4693dc3ab6ff428a8acb1c55a61  | d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pseudomonadales; f__Pseudomonadaceae; g__Pseudomonas                              |
| 7efb84e235dedd3f438ec486ebdb10    | d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pseudomonadales; f__Pseudomonadaceae; g__Pseudomonas                              |
| 11d7990b97e60b84cc0c21555c399c8e  | d__Bacteria; p__Verrucomicrobiota; c__Verrucomicrobiae; o__Methylacidiphilales; f__Methylacidiphilaceae; g__uncultured; s__marine_metagenome |
| 41d62d14db1d9bacab8293bd76a612732 | d__Bacteria; p__Verrucomicrobiota; c__Verrucomicrobiae; o__Methylacidiphilales; f__Methylacidiphilaceae; g__uncultured; s__marine_metagenome |
| e84af897d777bb37d4da8511f4257206  | d__Bacteria; p__Bacteroidota; c__Bacteroidia; o__Flavobacteriales; f__Flavobacteriaceae                                                      |
| d36ef3f7d940c1eb072258422ea02cd   | d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Vibrionales; f__Vibrionaceae; g__Vibrio                                           |
| 14e2ebe19f95e1238490eeb06aa51912  | d__Bacteria; p__Planctomycetota; c__Planctomycetes; o__Planctomycetales; f__Rubinisphaeraceae                                                |
| 08d75b1c857d493c2cddf356e4e934bb  | d__Bacteria; p__Verrucomicrobiota; c__Verrucomicrobiae; o__Methylacidiphilales; f__Methylacidiphilaceae; g__uncultured; s__marine_metagenome |
| a8c344ef603f33ec5b3ce10c0245f6c   | d__Bacteria; p__Verrucomicrobiota; c__Verrucomicrobiae; o__Chthoniobacteriales; f__Chthoniobacteraceae; g__LD29; s__unidentified_eubacterium |
| 6832e51787e573fa91cf377cf0ab0f8   | d__Bacteria; p__Verrucomicrobiota; c__Verrucomicrobiae; o__Chthoniobacteriales; f__Chthoniobacteraceae; g__LD29; s__unidentified_eubacterium |
| 9f7ec1db5795178ae94f0b5bfc87cdcf  | d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Thiotrichales; f__Thiotrichaceae; g__Thiothrix                                    |
| 3972e7d5970e45ae95dbf3ea806e1c5   | d__Bacteria; p__Bacteroidota; c__Bacteroidia; o__Flavobacteriales; f__Flavobacteriaceae                                                      |
| 7ba545dcbe01429d8d3cc132030b365e  | d__Bacteria; p__Cyanobacteria; c__Cyanobacteria; o__Chloroplast; f__Chloroplast; g__Chloroplast                                              |
| 85554e13a8cf6ef149e90ce6c6692c    | d__Bacteria; p__Cyanobacteria; c__Cyanobacteria; o__Cyanobacteriales; f__Microcystaceae; g__Snowella_OTU37504                                |
| 107273e73071282e4d0f4a0b45548da5  | d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Aeromonadales; f__Aeromonadaceae; g__Aeromonas                                    |
| 7c59d7d091dc4b4c4f6ec8917812cde   | d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Vibrionales; f__Vibrionaceae; g__Vibrio                                           |
| 8488b9a2dabec53087c467aee4e125c7  | d__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rhodobacterales; f__Rhodobacteraceae                                              |
| 2a2902e25fc39e421059b28def5d4a9   | d__Bacteria; p__Bacteroidota; c__Bacteroidia; o__Flavobacteriales; f__Flavobacteriaceae                                                      |
| 19c64c698117fe1c46fa4667b9feda5e  | d__Bacteria; p__Planctomycetota; c__Planctomycetes; o__Planctomycetales; f__Rubinisphaeraceae                                                |
| 9f4914b5f5ec018d4aa91914fb38079   | d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Vibrionales; f__Vibrionaceae; g__Vibrio                                           |
| e908557b17933be818126141d05a11b8  | d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pseudomonadales; f__Pseudomonadaceae; g__Pseudomonas                              |
| cf4d4ae9306f7c072495169d79130624  | d__Bacteria; p__Cyanobacteria; c__Cyanobacteria; o__Chloroplast; f__Chloroplast; g__Chloroplast                                              |
| 82c4219b80760ea2932b4a3abb1c3638  | d__Bacteria; p__Cyanobacteria; c__Cyanobacteria; o__Cyanobacteriales; f__Microcystaceae; g__Snowella_OTU37504                                |
| 69179c4a6011d0c9d30036cf18626520  | d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pseudomonadales; f__Moraxellaceae; g__Acinetobacter                               |
| d1fba131f7da0b97dff601df31adffc   | d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pseudomonadales; f__Pseudomonadaceae; g__Pseudomonas                              |
| 1cb5a06599fa021fd24666771272d1a   | d__Bacteria; p__Verrucomicrobiota; c__Verrucomicrobiae; o__Chthoniobacteriales; f__Chthoniobacteraceae; g__LD29; s__unidentified_eubacterium |
| 9a20dfcc47dadef8624914988bc4484   | d__Bacteria; p__Actinobacteriota; c__Actinobacteria; o__Frankiales; f__Sporichthyaceae; g__hgcl_clade                                        |
| 0040d84e34647a5baa9c9919792e86    | d__Bacteria; p__Myxococcota; c__Polyangia; o__Nannocystales; f__Nannocystaceae; g__uncultured; s__metagenome                                 |

Figure 2: Taxonomy representation of each samples.

Step 13: Export all the files (4 files) that Qiime2 generated, which is compatible in R with the

package Phyloseq.

1. table.qza : Created the feature-table.biom file to represent ASVs table.
2. rep-seqs.qza: Created the dna-sequences.fasta file to represent ASV sequences.
3. rooted-tree.qza: Created the tree.nwk file to represent tree file.
4. taxonomy.qza: Created the taxonomy.tsv file to represents the taxonomy as step12 shown.

```
(qiime2-2019.7) [MScBioinf@becker~/Caroline/kdl/qiime2_tutorial]$ qiime
tools export --input-path table.qza --output-path output ;
```

```
(qiime2-2019.7) [MScBioinf@becker ~/Caroline/kdl/
qiime2_tutorial]$ qiime tools export --input-path rep-seqs.qza --output-
path output;
```

```
(qiime2-2019.7) [MScBioinf@becker ~/Caroline/kdl/
qiime2_tutorial]$ qiime tools export --input-path rooted-tree.qza --
output-path output;
```

```
(qiime2-2019.7) [MScBioinf@becker ~/Caroline/kdl/
qiime2_tutorial]$ qiime tools export --input-path taxonomy.qza --output-
path output;
```

```
[MScBioinf@becker ~/Caroline/kdl/qiime2_tutorial/output]$ ls
data.jsonp          dna-sequences.fasta  forward-seven-number-summaries.csv  q2templateassets  tree.nwk
demultiplex-summary.pdf  feature-table.biom  index.html  quality-plot.html
demultiplex-summary.png  feature-table.tsv  overview.html  reverse-seven-number-summaries.csv
dist                 feature_w_tax.biom  per-sample-fastq-counts.csv  taxonomy.tsv
[MScBioinf@becker ~/Caroline/kdl/qiime2_tutorial/output]$
```

Step 14: Create biom file that compatible with R and Phyloseq. It will work in the statistical analysis. Initially, converted .biom format to .tsv format file.

```
(qiime2-2019.7) [MScBioinf@becker~/ Caroline /kdl/ qiime2_tutorial
```

```
/output]$ biom convert -i feature-table.biom -o feature-table.tsv -to-  
tsv;
```

```
(qiime2-2019.7) [MScBioinf@becker ~/Caroline/kdl/qiime2_tutorial/  
output]$ head feature-table.tsv;
```

```
# Constructed from biom file
```

```
#OTU ID      SRR1342456 SRR1342458 SRR1342459 SRR1342460 SRR1342461 SRR1342462 SRR1342463  
SRR1342464 SRR1342465 SRR1342466 SRR1342467 SRR1342468 SRR1342469 SRR1342470  
SRR1342471 SRR1342472 SRR3669131 SRR3669132 SRR3669133 SRR3669134 SRR3669135  
SRR3669136 SRR3669137 SRR3669138 SRR3669139 SRR3669140 SRR3669141 SRR3669142  
SRR3669143 SRR3669144 SRR3669145 SRR3669146 SRR3669147 SRR3669148 SRR3669149  
SRR3669150 SRR3669151 SRR3669152 SRR3669153 SRR3669154 SRR3669155 SRR3669156  
SRR3669157 SRR3669158 SRR3669159 SRR3669160 SRR3669161 SRR3669162 SRR3669163  
SRR3669164 SRR3669165 SRR3669166 SRR3669167 SRR3669168 SRR3669169 SRR3669170  
SRR3669171 SRR3669172 SRR3669173 SRR3669174 SRR3669175 SRR3669176 SRR3669177  
SRR3669178 SRR3669179 SRR3669180 SRR3669181 SRR3669182 SRR3669183 SRR3669184  
SRR3669185 SRR3669186 SRR3669187 SRR3669188 SRR3669189 SRR3669190 SRR3669191  
SRR3669192 SRR3669193 SRR3669194 SRR3669195 SRR3669196 SRR3669197 SRR3669198  
SRR3669199 SRR3669200 SRR3669201 SRR3669202 SRR3669203 SRR3669205 SRR3669206  
SRR3669207 SRR3669208 SRR3669209 SRR3669210 SRR3669211 SRR3669212 SRR3669213  
SRR3669214 SRR3669215 SRR3669216 SRR3669217 SRR3669218 SRR3669219 SRR3669220  
SRR3669221 SRR3669222 SRR3669223 SRR3669224 SRR3669226 SRR3669227 SRR3669228  
SRR3669229 SRR3669230 SRR3669231 SRR3669232 SRR3669233 SRR3669234 SRR3669235  
SRR3669236 SRR3669237 SRR3669238 SRR3669239 SRR3669240 SRR3669241 SRR3669242  
SRR3669243 SRR3669244 SRR3669245 SRR3669246 SRR3669247 SRR3669248 SRR3669249  
SRR3669250 SRR3669251 SRR3669252 SRR3669253 SRR3669254 SRR3669255 SRR3669256  
SRR3669257 SRR3669258 SRR3669259 SRR3669260 SRR3669261 SRR3669263 SRR3669264  
SRR3669265 SRR3669266 SRR3669267 SRR3669268 SRR3669269 SRR3669270 SRR3669271  
SRR3669272 SRR3669273 SRR3669274 SRR3669276 SRR3669277 SRR3669278 SRR3669279  
SRR3669280 SRR3669281 SRR3669282 SRR3669283 SRR3669284 SRR3669285 SRR3669286  
SRR3669287 SRR3669288 SRR3669289 SRR3669290 SRR3669291 SRR3669292 SRR3669293  
SRR3669294 SRR3669295 SRR3669296 SRR3669297 SRR3669298 SRR3669299 SRR3669300  
SRR3669301 SRR3669302 SRR3669303 SRR3669304 SRR3669305 SRR3669306 SRR3669307  
SRR3669308 SRR3669309 SRR3669310 SRR3669311 SRR3669312 SRR3669313 SRR3669314  
SRR3669315 SRR3669316 SRR3669317 SRR3669318 SRR3669320 SRR3669321 SRR3669322  
SRR3669323 SRR3669324 SRR3669325 SRR4039549 SRR4039550 SRR4039551 SRR4039552  
SRR4039553 SRR4039554 SRR4039555 SRR4039556 SRR4039557 SRR4039558 SRR4039559  
SRR4039560 SRR4039561 SRR4039562 SRR4039563 SRR4039564 SRR4039565 SRR4039566  
SRR4039567 SRR4039568 SRR4039569 SRR4039570 SRR4039571 SRR4039572 SRR4039573  
SRR4039574 SRR4039575 SRR4039576 SRR4039577 SRR4039578 SRR4039579 SRR4039580  
SRR4039581 SRR4039582 SRR4039583 SRR4039584 SRR4039585 SRR4039586 SRR4039587  
SRR4039588 SRR4039589 SRR4039590 SRR4039591 SRR4039592 SRR4039593 SRR4039594  
SRR4039595 SRR4039596 SRR4039597 SRR4039598 SRR4039599 SRR4039600 SRR4039601  
SRR4039602 SRR4039603 SRR4039604 SRR4039605 SRR4039606 SRR4039607 SRR4039608  
SRR4039609 SRR4039610 SRR4039611 SRR4039612 SRR4039613 SRR4039614 SRR4039615  
SRR4039616 SRR4039617 SRR4039618 SRR4039619 SRR4039620 SRR4039621 SRR4039622  
SRR4039623 SRR4039624 SRR4039625 SRR4039626 SRR4039627 SRR4039628 SRR4039629
```



[illegible]



[illegible][illegible]









[illegible]

```

0.0 0.0 120.0 1520.0 136.0 65.0 0.0 0.0 0.0 0.0 0.0 606.0 72.0 287.0
411.0 0.0 0.0 705.0 639.121.0 128.0 1086.0 1066.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 72.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 227.0 0.0 0.0 51.0 117.0 425.0 419.0 0.0
0.0 150.0 0.0 0.0 0.0 83.0 0.0 0.0 905.0 0.0 0.0 0.0 1249.0 1039.0 0.0
0.0 0.0 0.0 0.0 0.0 125.0 90.0 60.0210.0 0.0 0.0 72.0 182.0 483.0
0.0 0.0 0.0 1342.0 12.0 0.0 0.0 90.0 0.0 1555.0 80.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 3394.0 2852.0 0.0 0.0 0.0
0.0 48.0 0.0 519.0 80.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 952.0 0.0 75.0 50.00.0 0.0
88.0 121.0 0.0 90.0 62.0 129.0 469.0 0.0 82.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 97.0 0.0 337.0 0.0 0.0

```

```
[MScBioinf@becker ~/Caroline/kdl/qiime2_tutorial/output]$
```

Second, the taxonomy.tsv column names were modified:

```
[MScBioinf@becker ~/Caroline/kdl/qiime2_tutorial/output]$ sed -i
s/Taxon/taxonomy/ taxonomy.tsv | sed -i s/Feature\
ID/FeatureID/ taxonomy.tsv
```

```
[MScBioinf@becker ~/Caroline/kdl/qiime2_tutorial/output]$ head
taxonomy.tsv
```

```
[MScBioinf@becker ~/Caroline/kdl/qiime2_tutorial/output]$ head taxonomy.tsv
Feature ID taxonomy Confidence
b744eae1244325f606575483df0de67e d__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__SAR11_clade; f__Clade_III; g__Clade_III; s__Candidatus_Fonsibacter 0.929547650414144
08bd45c0807b6fb44faaa54fb53e4213 d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Burkholderiales; f__Rhodocyclaceae; g__Zoogloea 0.9969641808043913
38c27ceaed634984c1225a82648cf571 d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pseudomonadales; f__Pseudomonadaceae; g__Pseudomonas 0.9963280629074556
9c7e018344e9f5c9245cbe67238f23c8 d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pseudomonadales; f__Pseudomonadaceae; g__Pseudomonas 0.996703976360367
9539a8a644963195a48b457512f36720 d__Bacteria; p__Cyanobacteria; c__Cyanobacteria; o__Cyanobacteriales; f__Microcystaceae; g__Synechocystis_PCC-6803; s__uncultured_bacterium 0.983126138857414
39853883f6e11ee7f81c5c2260a0049 d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Thiotrichales; f__Thiotrichaceae; g__Thiothrix 0.9980379110033479
3f456a18968cea5c5e549e281f8f5808 d__Bacteria; p__Actinobacteriota; c__Actinobacteria; o__Frankiales; f__Sporichthyaceae; g__hgcI_clade 0.9849583834876566
41d7f4693dc3ab6ff428a8ac1c55a61 d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pseudomonadales; f__Pseudomonadaceae; g__Pseudomonas 0.9953651245358917
7efb84aef235dedd3f438ec486ebbd10 d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pseudomonadales; f__Pseudomonadaceae; g__Pseudomonas 0.9951378624553082
[MScBioinf@becker ~/Caroline/kdl/qiime2_tutorial/output]$
```

Finally, the feature-table.tsv was merged with taxonomy.tsv:

```
(qiime2-2019.7) [MScBioinf@becker ~/Caroline/kdl/ qiime2_tutorial]$ biom
add-metadata -i feature-table.tsv -o feature_w_tax.biom --observation-
metadata-fp taxonomy.tsv --observation-header
FeatureID,taxonomy,Confidence --sc-separated taxonomy --float-fields
Confidence
```

### Step 15: Use Picrust2 to do the functional analysis:

```
(qiime2-2019.7) [MScBioinf@becker ~/Caroline/kdl/qiime2_tutorial]$ qiime picrust2 full-pipeline --i-table table.qza --i-seq rep-seqs.qza --output-dir q2-picrust2_output --p-threads 5 --p-hsp-method pic --p-max-nsti 2 --verbose
```

### Step 16 : Export **q2\_picrust2\_output** files as biom files and then as TSV files:

```
[MScBioinf@becker ~/Caroline/kdl/qiime2_tutorial]$ cd q2-picrust2_output/
```

```
(qiime2-2019.7) [MScBioinf@becker ~/Caroline/kdl/qiime2_tutorial/q2-picrust2_output]$ qiime tools export --input-path ko_metagenome.qza --output-path output; mv output/feature-table.biom output/ko_metagenome.biom
```

```
(qiime2-2019.7) [MScBioinf@becker ~/Caroline/kdl/qiime2_tutorial/q2-picrust2_output]$ qiime tools export --input-path ec_metagenome.qza --output-path output; mv output/feature-table.biom output/ec_metagenome.biom
```

```
(qiime2-2019.7) [MScBioinf@becker ~/Caroline/kdl/qiime2_tutorial/q2-picrust2_output]$ qiime tools export --input-path pathway_abundance.qza --output-path output; mv output/feature-table.biom output/pathway_abundance.biom
```

```
(qiime2-2019.7) [MScBioinf@becker ~/Caroline/kdl/qiime2_tutorial/q2-picrust2_output]$ cd output
```

```
(qiime2-2019.7) [MScBioinf@becker ~/Caroline/kdl/qiime2_tutorial/q2-
```

```
picrust2_output/output]$ biom convert -i ko_metagenome.biom -o  
ko_metagenome.tsv --to-tsv
```

```
(qiime2-2019.7) [MScBioinf@becker ~/Caroline/kdl/qiime2_tutorial/q2-  
picrust2_output/output]$ biom convert -i ec_metagenome.biom -o  
ec_metagenome.tsv --to-tsv
```

```
(qiime2-2019.7) [MScBioinf@becker ~/Caroline/kdl/qiime2_tutorial/q2-  
picrust2_output/output]$ biom convert -i pathway_abundance.biom -o  
pathway_abundance.tsv --to-tsv
```

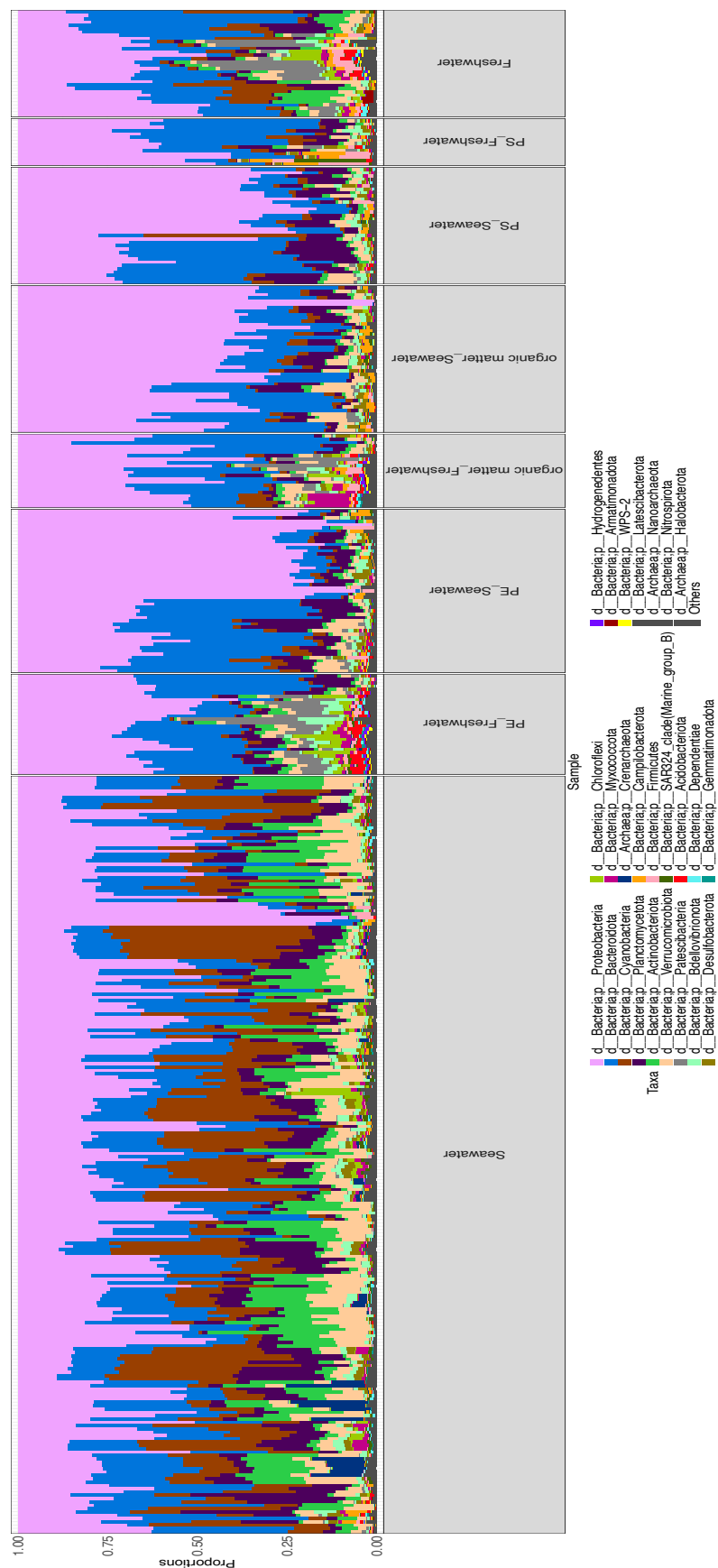


Figure 8. Taxa analysis of all the samples at Phylum level, the groups were classified as sample type.





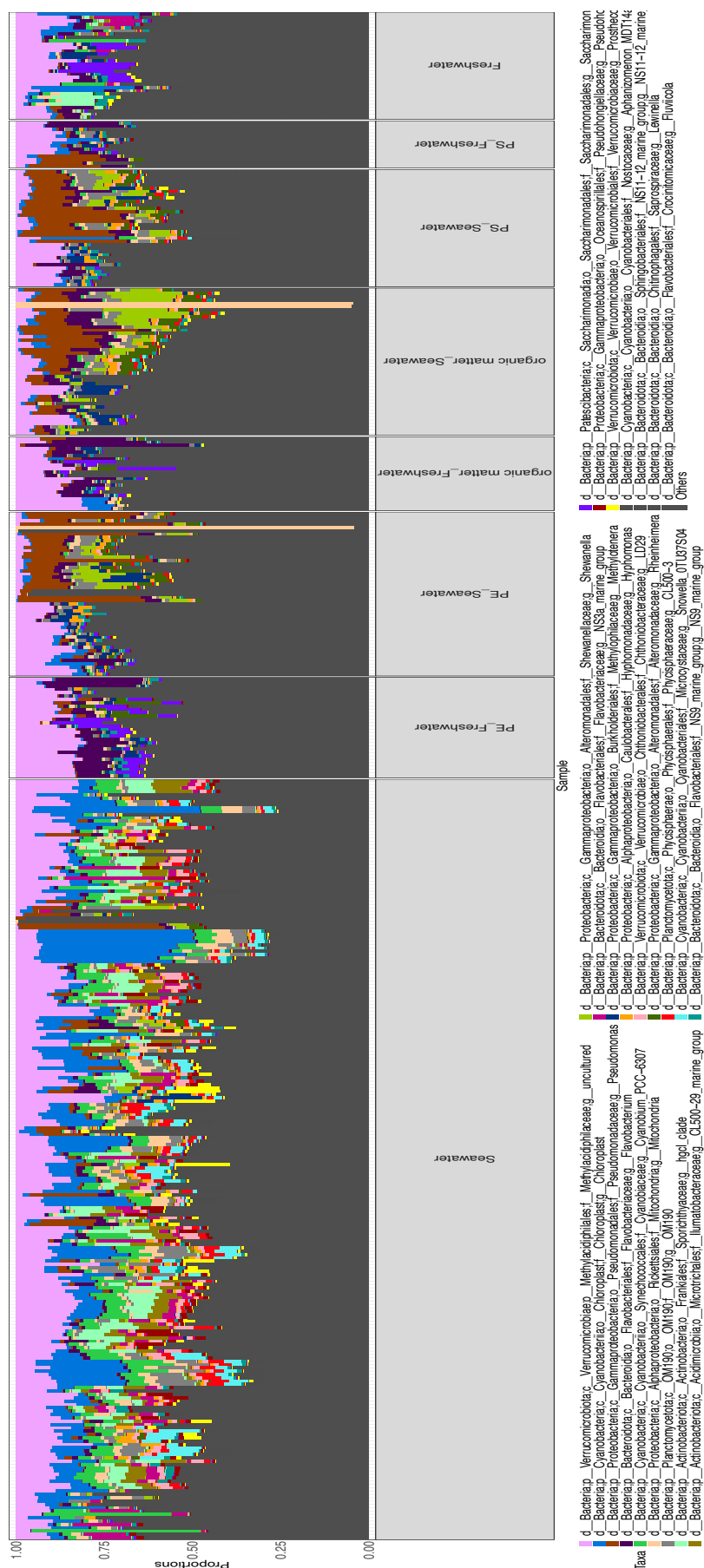


Figure 10. Taxa analysis of all the samples at Genus level, the groups were classified as sample type

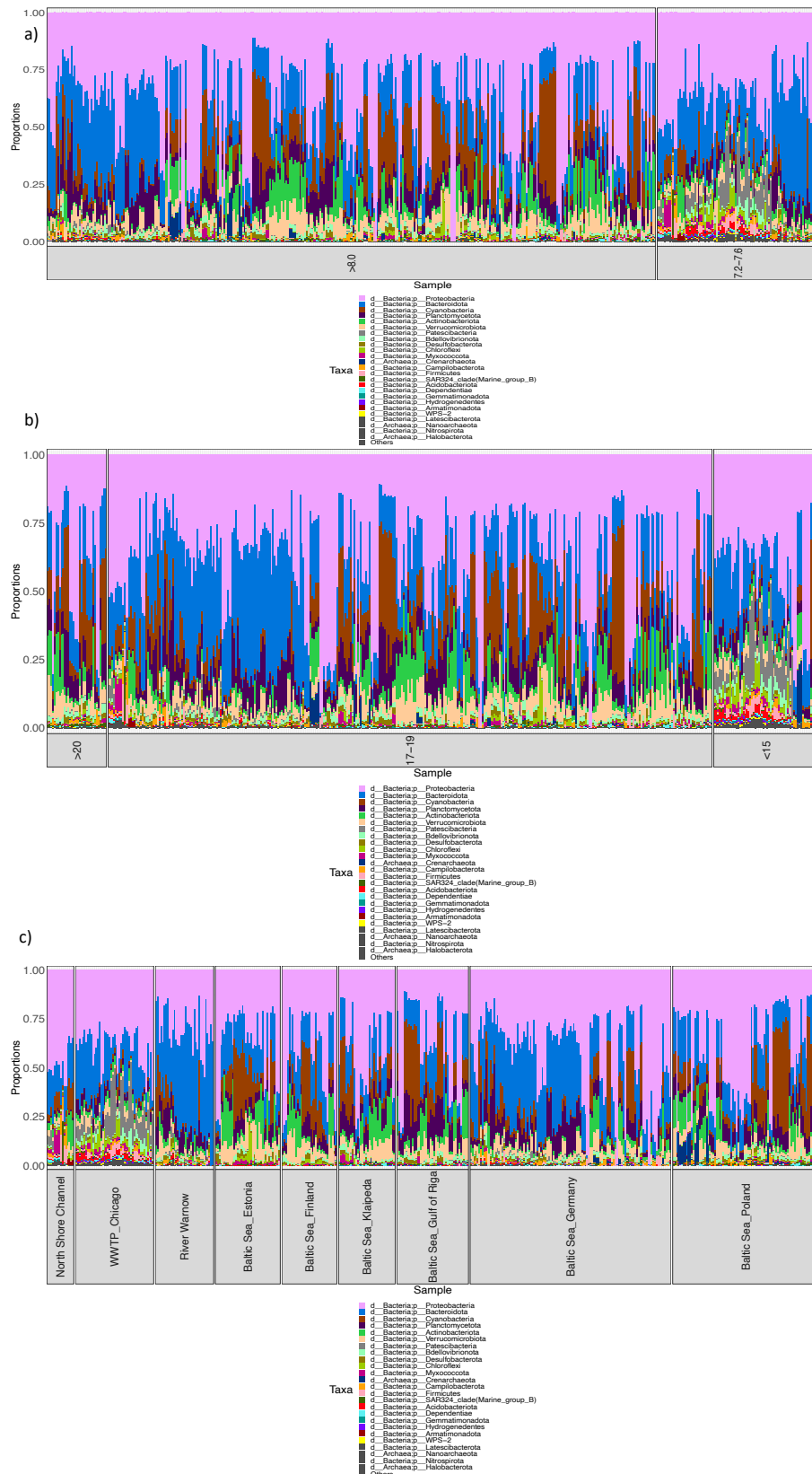
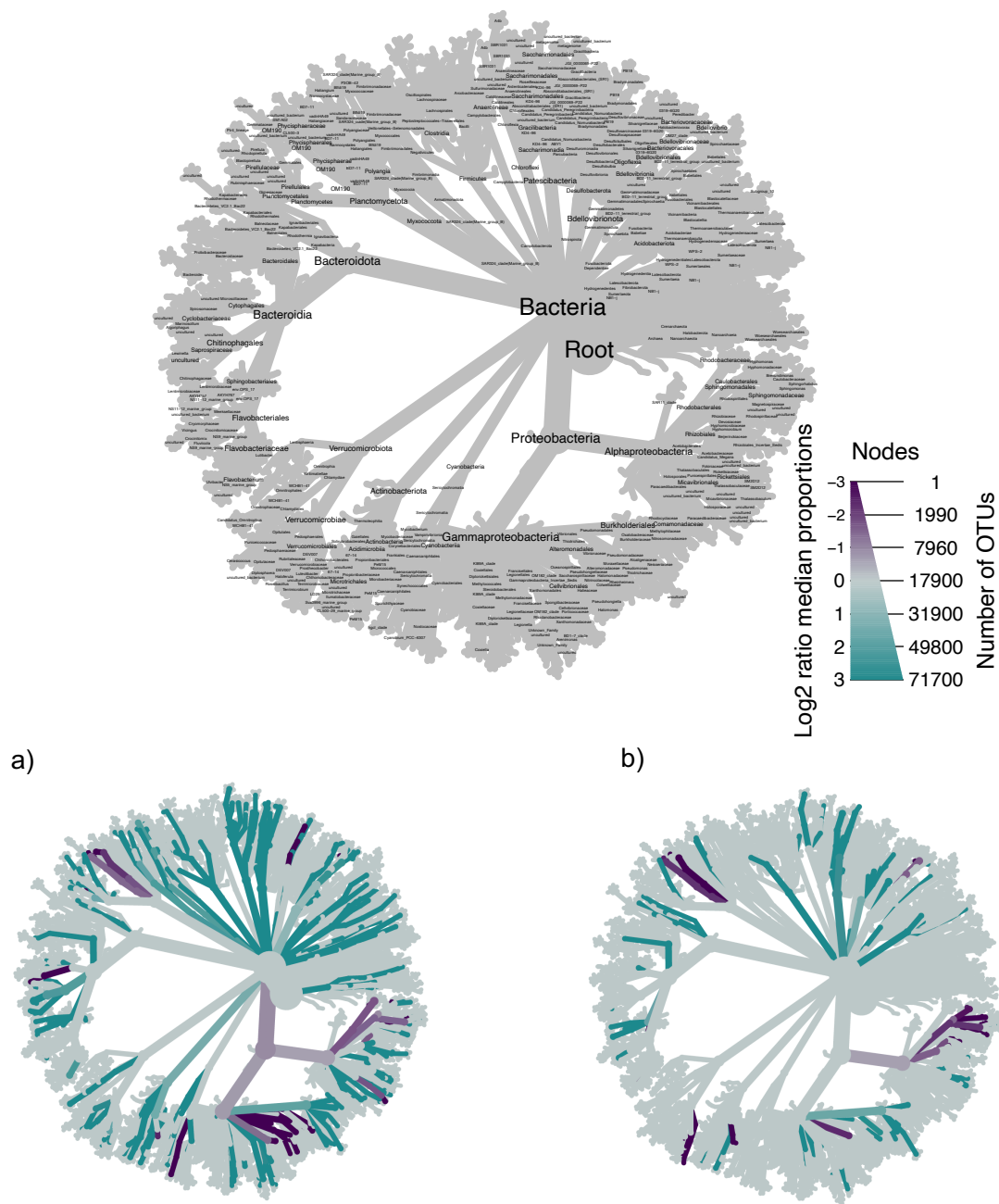


Figure 11. Taxa analysis of the samples at Phylum level, the groups were classified as (a) pH Range (b) Temperature Range (c) Sample Location.



**Fig. 12. Differential Heat tree plot, taxonomy genera given in the middle, the blue branch indicate the categories in the row, the purple branch indicate the categories in the column, the emphasis of comparison points on a) PE\_Freshwater VS PE\_Seawater. b) PS\_Freshwater VS PS\_Seawater.**