Molecular microbial ecology of Polar aerial environments

Lewis Paul Cuthbertson

LPC

PhD

2019

Molecular microbial ecology of Polar aerial environments

Lewis Paul Cuthbertson (BSc. Hons.)

A thesis submitted in partial fulfilment of the requirements of the University of Northumbria at Newcastle for the degree of Doctor of Philosophy

Research undertaken in the School of Health & Life Sciences

October 2019

Abstract

The biodiversity of bacterial communities in the Polar atmosphere is understudied, and as a result, the degree to which these communities influence macroecological patterns of biodiversity is poorly understood. This study aimed to investigate the bacterial biodiversity of the atmosphere by testing the hypothesis that bacteria are ubiquitous and present in polar air as heterogeneous communities. The study found bacterial DNA in all samples collected from both Poles, and whilst a degree of heterogeneity was observed in Arctic bacterial communities, there was an unexpectedly high level of sequence in the Antarctic.

Currently, there is no consensus as to the most appropriate bioaerosol sampling method, and the degree to which sampling methodology impacts the results of bioaerosol studies is still unknown. This variability was assessed by testing the hypothesis that bacterial community profiles in Polar air samples are not influenced by sampling methodology. However, the findings suggest that choice of bioaerosol sampling methodology can have a strong impact on the biodiversity observed.

The high level of sequence diversity in Antarctic air samples led to an investigation of technical variation as a result of their low biomass; and from this, it was found that the lower limit of biomass for a successful community description using an Illumina MiSeq approach was $1x10^6$ CFU per mL^{-1}, and that the lower limit at which this concentration of bacteria could be extracted using the most commonly used commercial DNA extraction kit was $1x10^7$ CFU per mL^{-1}.

Antarctic bioaerosol samples were found to have considerably lower biomass than these limits, suggesting that the results obtained were, in part due to technical variation as a result of their low biomass. The choice of bioinformatics pipeline was also investigated for low biomass samples, and found to have no effect on the final outcome. Overall, the study showed that the

Polar atmosphere contains very low biomass and that the pattern of biodiversity in this low biomass environment was both variable and not linked to physical or chemical environmental parameters. Hence, the atmosphere may act as a barrier to dispersal both into and out of the Polar regions.

Table of Contents

List of figures

closed reference OTU picking D) QIIME 2 ASV Greengenes 2013

List of tables

Declaration

I declare that the work contained in this thesis has not been submitted for any other award and that it is all my own work. I also confirm that this work fully acknowledges opinions, ideas and contributions from the work of others.

No ethical clearance for the research presented in this thesis was required.

I declare that the word count for this thesis is 36053 words

Name:

Lewis Paul Cuthbertson

Signature:

Date: September 2019

Chapter 1 - Introduction

1.1. Microbial ecology review

1.1.1. History

During the mid-seventeenth century, the first microscope was developed enabling the first cell to be described in 1665 by English scientist Robert Hooke (4), shortly followed by the discovery of the first bacteria, protozoa, algae, and fungi by Antoni van Leeuwenhoek (5). This progress eventually led to the disproval of the theory of spontaneous generation, first by Lazzaro Spallanzani who showed that chicken broth could be sterilised by heating (6) and finitely by Louis Pasteur in 1860 who's experiments proved that without exposure to exogenous spores and dust, sterile nutrient broths would remain without bacterial growth (7). This evidence lead to the development of germ theory and the first indirect nod to the atmosphere as an environment harbouring microscopic life. Robert Koch then implicitly proved germ theory through his work with Anthrax in 1876 (8), and the first published research articles specifically focused on microbes residing in the atmosphere began to emerge soon after this discovery (9, 10).

The term ecology was first stated in 1866 by Ernst Haeckel (11) who defined it as "the study of all those complex interactions referred to by Darwin as the conditions of the struggle for existence". Microbial ecology is the study of how microorganisms interact with their environment. Whilst the majority of early microbiological research in the late nineteenth century focused on medical application, one scientist, Sergei Winogradsky, began investigating the relationship between microbes and their environment, discovering lithotrophs and researching their use of inorganic compounds in the production of energy. Winogradsky could perhaps be considered the first true microbial ecologist, working alongside Louis Pasteur to initialise and establish the field (12). The implications relating microbial biodiversity to

agriculture, human health, climate change, and even life on other planets are considerable, and as such a great deal of environments have had a significant number of biodiversity surveys ranging from inside homes to Antarctica and outer space (13-16).

1.1.2. Biodiversity

At the core of investigations into the structure and function of microbial communities is biodiversity. Biodiversity is a measure of 'important ecological processes such as resource partitioning, competition, succession, and community productivity and is also an indicator of community stability' (17). The biodiversity of microbes was first studied in the 1960's (18). At present, the biodiversity of a microbial community is measured by two components: species richness (the number of species in a community) and species evenness (how species are distributed within a community). The biodiversity of microbial ecosystems is directly related to biotic interactions between species such as competition for resources and abiotic environmental conditions such as temperature (19), pH (20), humidity (21), and UV radiation (22). Temporal and spatial variation has a considerable impact on the measured biodiversity of an ecosystem (23-27). The relationship between microbes and their environment has led to the concept of microbial biogeography gaining significant interest over the past decade (28-34).

1.1.3. Molecular microbial ecology

Until the advent of molecular techniques, microbial ecology studies were limited to the use of culture dependent techniques. These traditional culture dependent studies were relatively slow due to their reliance on phenotypic identification and also fail to describe the full bacterial biodiversity of environments, as it has been shown that <1% of environmental bacteria are cultivable, with the remaining >99% present in a viable nonculturable (VBNC) state (35, 36). Modern studies employ an omics-centric approach, focusing more on genotypic identification of microbes. Genomics began developing after the discovery of DNA by Miescher in 1869,

which he then named Nuclein (37). Further developments such as the discovery of DNA's role in inheritance (38), that DNA composition is species specific (39), and the double helix structure of DNA (40), all led on to the development of first generation sequencing methods such as the dideoxy technique developed in 1977 by Frederick Sanger (41). The principles underpinning these pioneering techniques are very similar to those of modern next generation sequencing (NGS) technologies.

Woese proposed that life could be classified into a sequence based tree, however acknowledged that there must first be a suitable phylogenetic marker by which to differentiate groups (42). Bacterial rRNA genes were identified as a suitable target to fulfil this role, as they were present kingdom wide, and were highly conserved, but represented enough variation to differentiate between species (43). Specifically, 16S rRNA was presented as the ideal target marker sequence because of its ubiquitous presence in bacteria due to the critical function it has in coding for the small subunit rRNA (44). 16S rRNA is approximately 1500bp in length and is now understood to be made up of ten conserved regions separating nine hypervariable regions (1) (figure 1.1).

The conserved regions of the gene are ideal for performing alignments when investigating phylogenetic relationships between bacteria. One of the main drawbacks of using the 16S gene for molecular ecology comes from the fact that the number of genes present varies from species to species, for example *Escherichia coli* (*E.coli*) strains are known to have seven copies of the gene (45) whereas some bacteria are known to have just one (46). This fact means caution must be taken when drawing quantitative inferences on bacterial communities studied utilising the 16S rRNA gene, although some inference may be drawn from relative abundances. An additional drawback of this gene as a genetic marker is the decreased depth of taxonomic resolution due to its relatively short length.

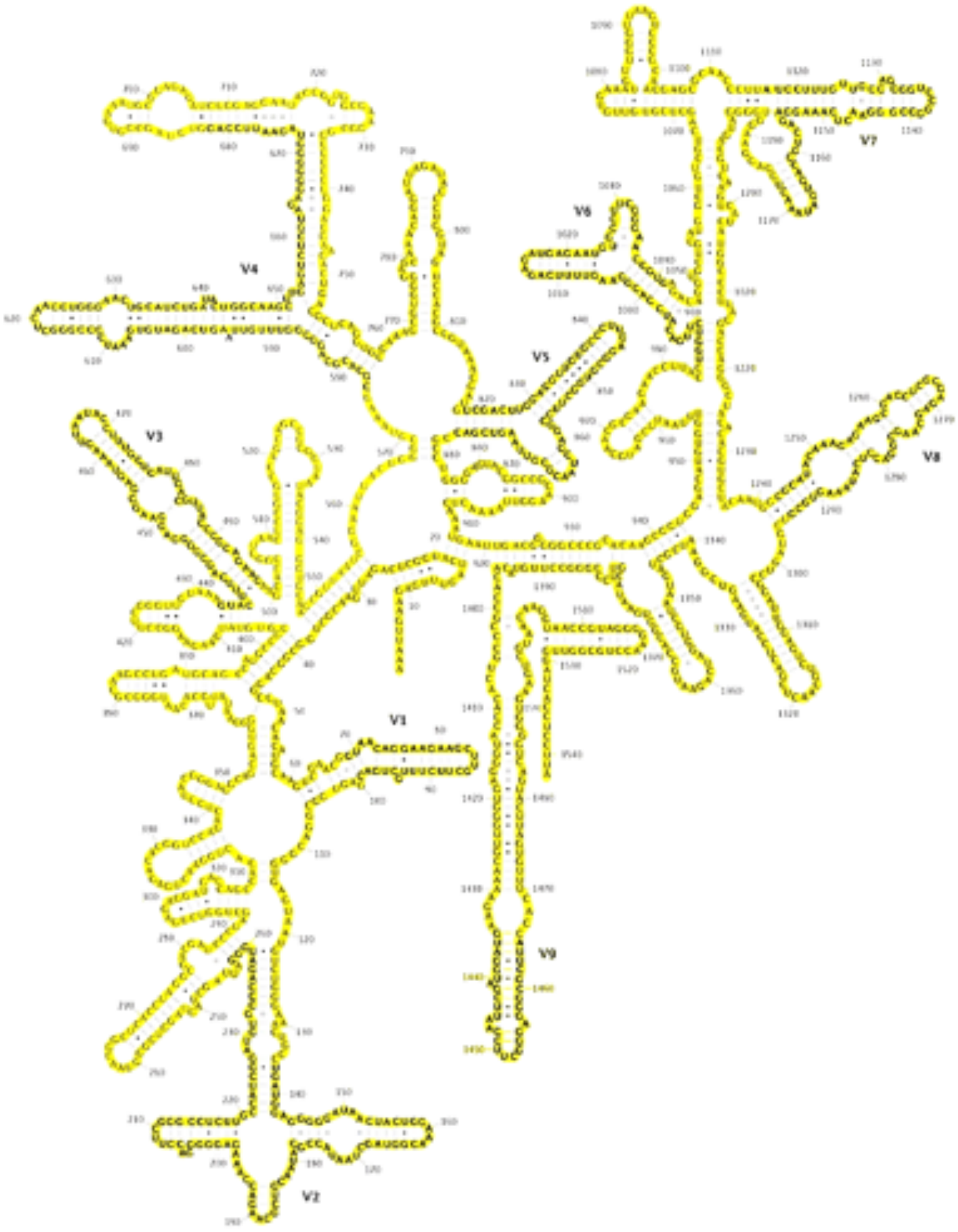


Figure 1.1. Secondary structure of 16S rRNA isolated from *E.coli*. Hypervariable regions in bold. (1)

The success of the gene as a universal marker cannot be understated, the original tree of life published using 16S rRNA contained just 10 phyla (43), however the most recent tree of life published based on the gene contained a proposed 92 phyla (2) (Figure 1.2). The importance of sequencing poorly described environments for difficult to culture microbes is clear (47). The tree highlights the fact that a number of phyla still do not have well characterised representatives, due to the selection against the characterisation of environmental microbes in favour of a bias toward medicinal, agricultural, and industrially relevant microbes (48). As the societal need for novel bioproducts continues to increase, efforts to cultivate microbial life we know is there only through genomic screening will only increase. As fields such as biotechnology continue to develop, our eye will turn more to these uncharacterised microbes and environments (49, 50).

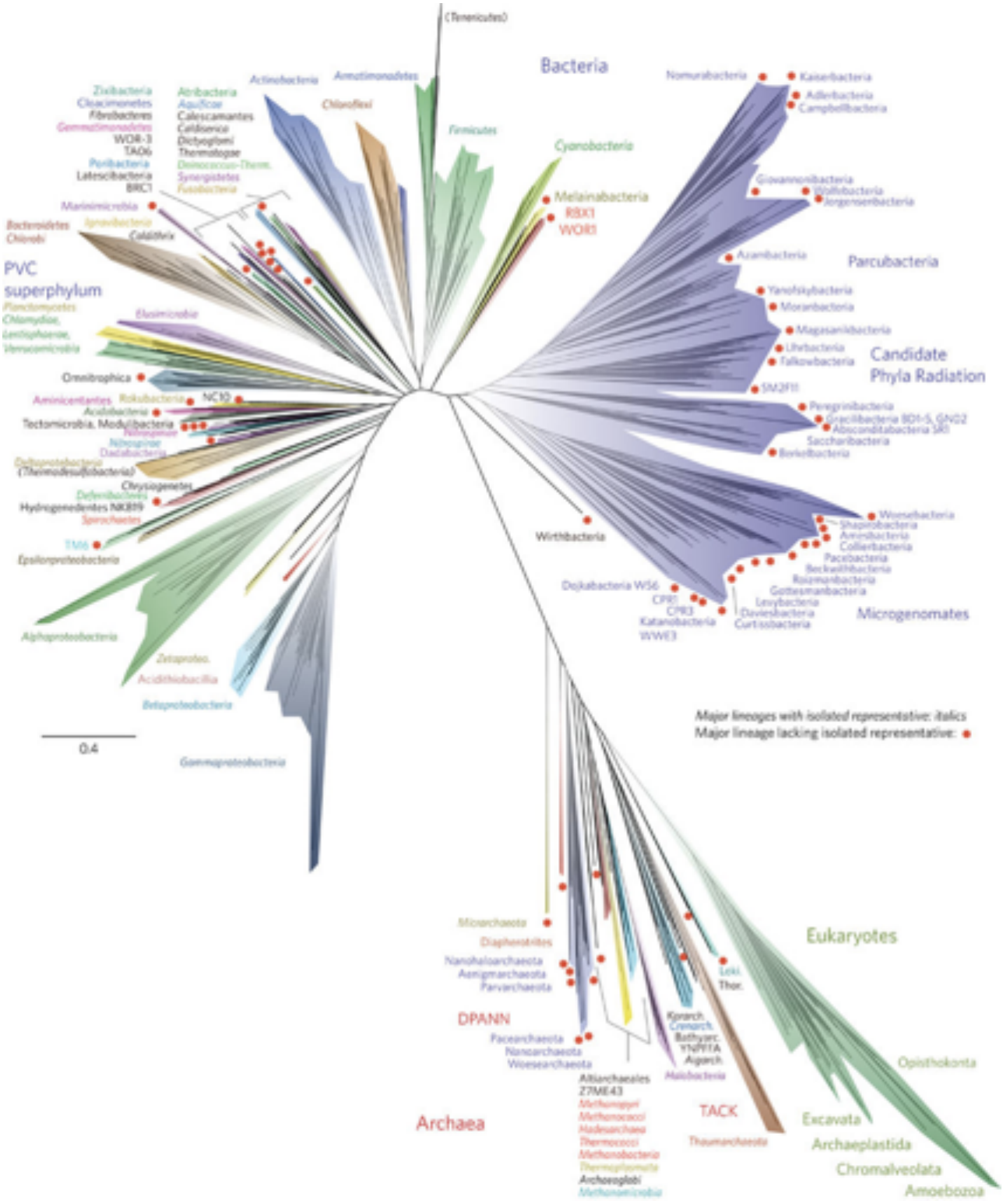


Figure 1.2. The most recent tree of life, containing 92 bacterial phyla, 26 archael phyla, and 5 Eukaryotic super groups. Major lineages are coloured arbitrarily. Well characterised lineages are italicised. Lineages lacking isolated representatives are non-italicised and accompanied by a red dot (2)

The considerable progress in describing the diversity of microbial life in numerous environments can be partly attributed to the recent development of high throughput sequencing technologies (also referred to as next generation sequencing technologies (NGS), massively parallel sequencing, and second generation sequencing technologies). Moving on from first generation sanger sequencing where only one forward and reverse read of a single fragment could be sequenced at a time, the development of these platforms allowing millions of fragments to be sequenced in a single run, has considerably reduced the cost of sequencing whilst more than tripling the throughput (51). This feat has been achieved through innovations such as the use of barcoded primer sets, which span the short hypervariable regions of the 16S rRNA gene, allow vastly increased sample coverage and also allows for the multiplexing of samples on an individual sequencing run (52, 53).

There are a range of sequencing instruments available, such as Illumina MiSeq, or the Ion Torrent PGM. Each technology has its own advantages and disadvantages, for example differing error rates, run times, number of reads, read length, and yield per run, all of which affect the cost of the experiment. There are a lot of studies comparing the different technologies (54-56). The chemistries of each technology are quite similar and continue to be developed. The most recent Illumina MiSeq reagent kit's v3 chemistry has the capability to generate up to 25 million reads per run, and output 15 Gigabases when using a 2x300bp output. The accuracy of these instruments varies, for example the v3 Illumina chemistry gives an inferred based call accuracy of 99.9% (Q score >30) to more than 70% of bases when using the 2x300bp output.

One issue known to occur on the Illumina sequencing platform is cross talk between samples in different lanes, which could impart a degree of contamination in output data files, and is thought to be responsible for up to 2% of reads in particular sets of data (57). The process of sequencing also imparts a natural bias toward samples with higher concentrations of DNA, which means samples with lower concentrations often lose out and fail to gain sufficient

coverage, one means by which this can be accounted for is by normalising samples to a set concentration during library preparation (15, 58).

In its raw format, NGS data consists of the assigned sequencing read and associated quality information. Processing the data manually would be a near impossible, labour and time intensive task. Transforming the data from its raw form to a high quality and interpretable state is more the realm of statistics and computational sciences than microbial ecology, and requires running a series of transformations on the data through bioinformatic pipelines. A number of pipelines are available for processing 16S NGS data, however the most commonly used pipelines are Quantitative Insights Into Microbial Ecology (QIIME) (59) and Mothur (60); when discussing QIIME, unless specified otherwise the version referred to is QIIME 2 (61). Mothur is a pipeline developed and maintained by the Schloss work group, who regularly produce literature pertaining to the quality of the data that can be attained as well as helpful tutorials (https://www.Mothur.org/wiki/MiSeq_SOP). QIIME on the other hand is a pipeline based on the implementation of several well used individually developed tools. The majority of 16S amplicon bioinformatic pipelines are made up of similar steps (figure 1.3), differentiated by the implementation of different methods to perform each step. The developers of each pipeline also provide recommended workflows and SOPs.

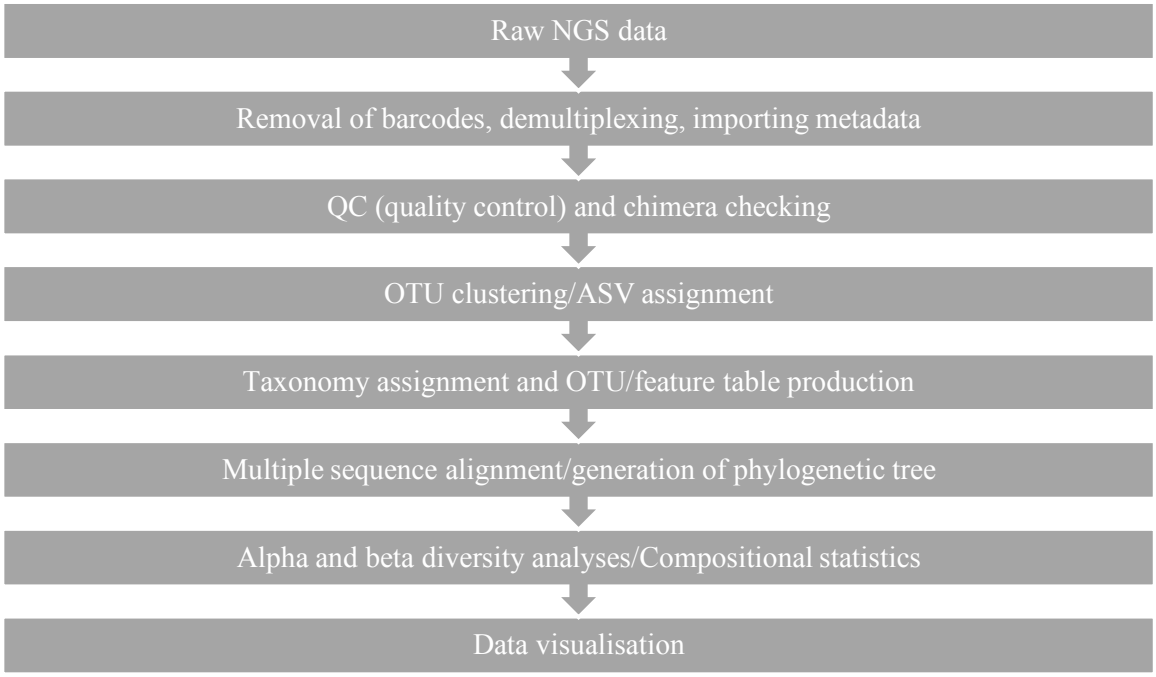


Figure 1.3. Flowchart depicting a typical 16S analysis pipeline from NGS output to data visualisation. (QC = Quality control, OTU = Operational Taxonomic Unit, ASV = Amplicon Sequence Variant).

Once raw reads have been demultiplexed, they can be read into a chosen analysis pipeline. A metadata file is first required in order to provide file names, sample ID's and additional environmental data. This is then followed by stringent QC based upon the quality profile of the sequences output by the NGS instrument utilised. The method of quality filtering utilised by Mothur is carried out in multiple steps. First forward and reverse reads are combined to form contigs, with the stipulation that where a disagreement occurs between the forward and reverse read and both disputing bases have Q scores of <25 the contig will be removed. Following this step, any base-calls which were ambiguous are removed and any contigs which are too long are removed. Sequences are then combined into a file containing just unique sequences and abundance data and aligned against a 16S database, any non-aligned sequences are removed. Sequences are then pre-clustered with the parameter of 1 different base pair per 100 before chimeras are removed with the implementation of VSEARCH (62). Undesirable taxa are then

filtered such as Archaea, chloroplasts, and mitochondria as whilst they may have matched a database, they are not targeted using a 16S primer set. Finally, OTUs are randomly clustered at a dissimilarity cut-off level of 0.03.

The previous version of QIIME, QIIME v1 quality filtered Illumina data by removing reads with more than 1 ambiguous base-call, and reads where the high-quality region of the sequence is less than 75 bases long, where high-quality region is defined a stretch of bases containing no more than 1 quality score less than 1e-5. QIIME v1 then utilised software such as UCLUST (63) to cluster sequences into OTUs and check for chimeric sequence utilising a reference based method, although de novo chimera checking was available through Chimera Slayer (64). The closed- and –open reference method of OTU assignment used by QIIME v1, where sequences are first matched to a reference base and unmatched sequences are then clustered de novo, has recently been shown to potentially produced too many OTUs and exaggerated diversity estimates (65).

QIIME v2 offers two options for quality filtering and identification of true sequence variants, DADA2 which has been shown to produce more real variants and fewer spurious sequences than both the QIIME v1 workflow and Mothur (66) and Deblur which has been shown to have a higher stability (i.e., obtaining the same sOTU (sub-OTU) across different samples) and better resolution of mock communities than DADA2 (67). DADA2 first requires the inspections of the quality profiles of both the forward and reverse reads to trim low quality bases, a data specific parametric error model is then learned, identical reads are combined along with abundance data, and the error model is applied to the dereplicated data, before spurious reads are further reduced by merging overlapping reads. Finally all sequences which can be exactly reconstructed as bimeras from the most abundant sequences are identified and removed. The output of the DADA2 pipeline is referred to as an Amplicon Sequence Variant (ASV). As oppose to a traditional OTU used in QIIME v1 and Mothur, where OTUs are binned

together at a set dissimilarity threshold, improvements to the methodologies both at a sequencing and bioinformatics level allow ASVs to be resolved to a single nucleotide base, offering greater phylogenetic resolution. As well as this, by not randomly binning, as is done with OTU clustering, the results are reproducible (68). Deblur first applies quality filtering based on quality scores as described by Bokulich et al (69) before sequences are truncated at a specified length based on quality information. Deblur uses sub-OTUs as an exact sequence output similar to DADA2's ASVs. The output of both options in QIIME is a feature table populated by unique sequences and their abundance data on a per sample basis.

After clustering sequences into OTUs or assigning them as ASVs, the representative sequences are then assigned taxonomy. Taxonomy assignment is carried out through comparing the sequence to a reference database. There are four regularly used and well populated taxonomic databases. These are the Ribosomal Database Project (RDP) which contains the largest collection of aligned and fully annotated rRNA gene sequences from bacteria, archaea and fungi (70), Greengenes containing bacterial and archaeal sequences (71), SILVA containing eukaryotic, bacterial, and archaeal sequences (72), and the NCBI database (73). There are known classification inconsistencies in microbial nomenclature due to the requirements of the classifying researcher (74). This means that the choice of user deposited database can impact results (75). SILVA and RDP classify only to genus level, whereas NCBI and Greengenes give classifications at species level where possible. It has been shown that the accuracy of taxonomic classification improves when a Naïve Bayes Classifier is trained to the region of the target sequences (76).

Pipelines often include methods by which to perform a multiple sequence alignment on the quality-filtered reads. QIIME v1 performed this task with the implementation of PyNast (77) whereas QIIME v2 uses MAFFT (78). Mothur implements its own alignment method, briefly the general approach is to i) find the closest template for each candidate using kmer searching,

blastn, or suffix tree searching; ii) to make a pairwise alignment between the candidate and de-gapped template sequences using the Needleman-Wunsch, Gotoh, or blastn algorithms; and iii) to re-insert gaps to the candidate and template pairwise alignments using the NAST algorithm so that the candidate sequence alignment is compatible with the original template alignment. When aligning multiple sequences, it is often preferential to mask the alignment prior to phylogenetic analyses to remove any highly variable positions that add noise to a phylogenetic tree (79). In order to produce a phylogenetic tree from the aligned sequences, typically a distance matrix is produced. Mothur generates a distance matrix and then performs the task of producing a phylogenetic tree by implementing the clearcut program (80). In QIIME, the production of a rooted and an unrooted phylogenetic tree is carried out by the FastTree program (81). As oppose to a distance matrix, FastTree stores sequence profiles and utilises neighbour-joining using heuristics to identify candidate joins. This method is preferential when working with large datasets as it greatly reduces the amount of computing power and time required.

Once the raw data has been quality filtered, taxonomic assignments made, and phylogenies inferred, the data is ready for analysis. The simplest measure of diversity is to implement the observed OTUs index, which is a count of the number of OTUs present in each sample. It is commonplace in microbial ecology to assess biodiversity by using a compound diversity index. Compound diversity indices encapsulate both the richness and evenness of a microbial sample. There are an enumerable number of ways to assess both aspects of diversity, and as such there are an infinite number of candidate indices, all of which must focus more upon one of the two aspects, meaning there is no single perfect diversity indices to use (82), as such it is prudent to use a selection of different indices to better describe both aspects of biodiversity.

Both QIIME and Mothur offer a sizeable range of similar options where alpha (within community) and beta (between communities) diversity measurements are concerned. One of the most common alpha diversity indices within microbial ecology literature is the Shannon

Index (figure 1.4 (i)), a non-parametric measure of species richness and evenness. However, despite its popularity, the Shannon index is not without issue. The index is prone to increasing error as the proportion of undescribed species within an environment rises, and as the true richness of a community is never truly known, the use of this index can never truly be without bias (83). Additionally, it would take more than 100,000 species within a sample to gain a H value of more than 5, making the result difficult to interpret (84). The Shannon index tends to better express species richness. As previously mentioned, it would be prudent to compliment the use of this index, with one of the indices which is more centred around dominance (the inverse of evenness). One of the most used of these indices is the Simpson index of diversity (Figure 1.4 (ii)) (85). The Simpson index of diversity describes the probability that the next species drawn from a population is the same as the first. Another means by which communities' diversity can be measured is in its taxonomy.

In order to measure a communities' diversity based upon taxonomy, a phylogenetic tree must be available. The most commonly used phylogenetic index is Faith's phylogenetic diversity (86), where the branch lengths of the phylogenetic tree are summed and expressed. Beta diversity metrics describe the distance between communities by multiple pairwise comparisons. As well as those described, there are a number of different options to choose from, however whilst the choice of diversity indices utilised can directly influence results, data mining to attain the most significant results is bad practice (87).

Often, the range of reads in quality filtered data sets is large, and so in order to fairly assess diversity between samples with different depths, samples can be rarefied. The choice at which depth to rarefy samples to should be made based on rarefaction curves computed based upon the number of OTUs observed or diversity estimate. Data is typically rarefied at the depth all samples asymptote on a rarefaction curve. Whilst rarefying is common practice in microbiome studies, the wastefulness of discarding reads and the effect this has on the ability to detect

differences between populations is now being questioned, with suggestions rarefying should be stopped altogether and replaced with new methods based upon normalisation using statistical mixture models (88).

$$i)\ H = -\sum_{i=1}^{s} (p_i \log_2 p_i)$$

$$ii)\ D = 1 - \sum p_i^2$$

Figure 1.4. i) Shannon Index (s = number of OTUs, P_i = proportion of total community represented by OTU i). ii) Simpson Index of diversity (Pi = the proportion of the total community represented by OTU i) (3).

There are many issues facing modern microbial ecology. Sampling regime, sample storage, DNA extraction method, DNA extraction location, choice of sequencing technology, sequencing depth, choice of primer set, and choice of bioinformatic pipeline can impart bias on biodiversity data (54, 89, 90).

The issue of contamination in microbial ecology studies has long went ignored by the vast majority, however recent studies have shown that, despite it being commonplace to employ strict sterile technique in molecular laboratories, random and reagent contamination is common, with varying contaminant taxa present in commercially available kits used on mass (91-94). The issue of contamination is more pressing for studies involving low biomass samples, as the less concentrated contaminant 'noise' is drowned out by high concentration samples. Our understanding of contamination continues to improve, however the issue remains ignored by the majority of biodiversity studies, with most publications failing to state the use of control samples or alternatively verifying their presence but neglecting to explain how they were utilised (95-97). Whilst there remains no consensus on how to deal with contaminant sequences, studies will continue to exclude this information in an attempt to safeguard the

validity of their results. At present there are only three methods by which to exclude potential contaminants from data where negative controls have been run:

1) The removal of taxa previously described as potential contaminants (91-94).

2) The removal of potential contaminants based on an *ad-hoc* approach, by applying an arbitrary cut off based on the relative abundance of taxa present in negative controls such as 5% or 10% (15, 58, 98) or

3) The removal of taxa identified by a chosen statistical model (99).

Statistical models have been developed based on assumptions, such as reagent contaminants appearing to present a strong inverse correlation with sample DNA concentration after library preparation (92), and total sample DNA is a mixture of contaminating DNA in a uniform concentration and true sample DNA present in a varying concentration across samples (99). An algorithm to remove potential taxa present due to Illumina cross talk has also been suggested (57), but is not routinely implemented. Each of these approaches requires a different set of assumptions, for example removing taxa previously described as contaminants in commercial kits does not account for the possibility that these taxa are a true feature of the sampled environment. Ultimately, contamination is a consistent artefact of modern microbial ecology studies based on NGS sequencing and there remains no consensus on the correct approach to remove potential contaminants. As such it is the role of the researcher to account for and address these on a per study basis during data analysis and interpretation.

1.2. Aerobiology review

The atmosphere forms part of the biosphere and surrounds and interacts with every habitat and organism on Earth, either directly or indirectly (100). It is essential to life. However, life in the atmosphere itself is challenging as organisms must cope with extreme psychrophilic and oligotrophic conditions. Although recent methodological developments in both the collection

and analysis of aerial samples have accelerated our understanding of microorganisms inhabiting the atmosphere, many questions still remain. Here, I present a brief overview of aerobiological studies to date, describe the atmosphere as a habitat for microorganisms, review what is known about bioaerosols with regards to dispersal, diversity and atmospheric processes, and discuss the relevance of these organisms in atmospheric function.

1.2.1. History

The discipline of aerobiology was first founded in the 1930's by Meier during his expeditions with Charles Lindbergh and was defined as the passive transport of biological particles through the atmosphere and its effect on living systems and the environment (101). But it wasn't until the mid 20th century when medical research about microorganisms began to increase that research into aerobiology gained momentum with studies proving that tuberculosis (TB), influenza and streptococcal infections all required aerial transfer (102-104). The number of aerobiological studies has increased decade by decade since the founding of the discipline (figure 1.5), with methodologies changing throughout the course of this time.

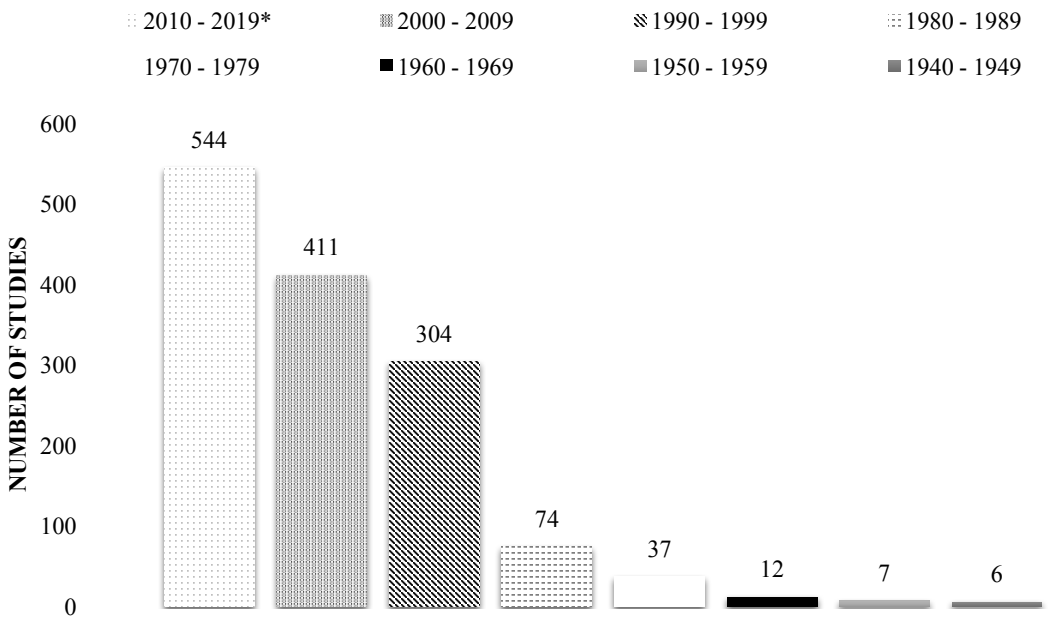


Figure 1.5. Total number of aerobiological studies for each decade since the term was invented, as per the total number of document results per decade based on a Scopus search of the term 'aerobiology'.

Early aerobiological studies employed basic passive techniques such as plate fall assays for the collection of airborne microorganisms or bioaerosols. Such techniques are still very much in use today, as they result in a microbial culture which can be subjected to physiological investigations, unlike many of the molecular based methods, which may only generate DNA sequence data. This was followed by development of active sampling techniques in the early 20th century, using simple impactor and impinger designs. These early designs led to the Hirst spore trap and Anderson air sampler developed during the 1950's and derivatives of these, such as the multi-phase Anderson sampler, were developed shortly after (105-108). Modern techniques have not diverged greatly from these founding methods in terms of the basic principle, besides the inevitable improvements that come with technological advancement in manufacturing precision and effectiveness of the equipment. Typical of many microbiological studies, early aerobiological techniques relied solely on culture dependent methodologies,

restricting observations to the relatively small culturable fraction of the atmospheric biota, which lead to the belief that the atmosphere had significantly lower levels of diversity, a belief that has changed following the development of microscopy and molecular techniques.

1.2.2. The atmospheric habitat

As an environment for microbial life, the atmosphere can be regarded as harsh as any on Earth with the lowest temperatures, highest ultra violet-radiation and extreme oligotrophic conditions. The structure of the atmosphere has been described in detail elsewhere, but in summary (adapted from 109), it can be divided into five main layers. The first is the Troposphere which is the layer at ground level extending up to 14.5 km, it is the densest layer of the atmosphere containing the majority of the Earth's life. It also contains over 99 % of the water vapour in the atmosphere and is where almost all familiar weather conditions occur. The second layer is the Stratosphere which begins directly after the Troposphere and extends up to 50 km, the Stratosphere contains the ozone layer. Above that comes the Mesosphere which extends further up to 85 km and temperatures in this layer reach lows of -100 °C, lower than any temperature ever recorded on the surface of the Earth. This is the highest atmospheric layer in which microscopic life has been found to date (110). Above the Mesosphere comes the Thermosphere and then the Exosphere, respectively, which extend beyond 600 km above the Earth's surface.

The majority of the atmosphere consists of dry air, which provides an almost inhospitable environment for microbial life. Clouds are a combination of condensed water droplets and ice crystals; they provide an important refuge for bioaerosols as they contain liquid water and levels of organic acids and alcohols comparable to fresh water lakes (111).

1.2.3. Bioaerosols

Microorganisms found in the atmosphere are commonly referred to as bioaerosols; the term bioaerosol refers to all living components of the atmosphere. There are multiple types of bioaerosol found in the atmosphere including bacteria, archaea, algae, fungi, viruses, smaller eukaryotes and pollen and these can be viable, dead, dormant or a combination of the three. Indeed, it has been estimated that a bacterial biomass of 40-1800 Gg is aerosolised annually (112).

1.2.3.1. Aerosolisation

Microorganisms enter the atmosphere from exposed terrestrial and marine surfaces. In terrestrial locations, bioaerosols are released from plant, soil and other surfaces when drying reduces bonding forces and this loosely bonded material is disturbed by strong air movements (113). Studies have shown vertical bacterial fluxes in terrestrial locations supporting this theory (114). Marine bioaerosol formation is also directly related to air movement, as the majority of marine aerosols are released by either evaporation or a bubble bursting processes (115). Theories regarding the aerosolisation of microbes are supported in the literature by findings which show direct relationships between the microbiota of the atmosphere and nearby surface level sources. However, contradictory studies also exist which show the potential for long-range atmospheric transport (116).

1.2.3.2. Dispersal

Once aerosolized, there are two main mechanisms by which bacteria are transported through the atmosphere: free-floating and attached to larger particles. Free floating particles in the atmosphere are unlikely to come into contact with other organisms frequently, but particle-associated bacteria living in close quarters and subjected to stress whilst suspended in the atmosphere might be subjected to increased horizontal gene flow (117). It is this hypothetical

horizontal gene transfer and the abundance of bacteria within the atmosphere that has drawn attention to the environment as a potential source of new antibiotics (50). Whilst airborne, it is estimated that bacteria have a residence time of between 2.2 and 188.1 days (118). Average generation times of bioaerosols have been measured to be between 3.6 and 19.5 days, a generation time comparable to that of many marine organisms (111, 112). Bioaerosols have been shown to undergo cross continental transport in plumes of desert dust in Asia (116) and across the Pacific Ocean with organisms of Asian origin detected in North America (119). Bacteria from the atmosphere can be deposited by two key mechanisms: dry deposition and wet deposition. Dry deposition is the process of bioaerosols adhering to plants, water and other ground surfaces with which they come into contact (113), while wet deposition describes the process by which bioaerosols are deposited through precipitation (rain, snow and hail) (120).

1.2.3.3. Atmospheric microbial processes

Despite the long-held belief that the atmosphere acts solely as a reservoir or conduit for microbial life, there are compelling arguments that life in the atmosphere should also be considered a functioning ecosystem. Suspended microbes have been shown to be both metabolically active and capable of reproduction (121), performing multiple functions in the such as ice nucleation (122), cloud formation (123), the degradation of organic carbon based compounds (124), nitrogen processing (125), sulphur oxidation and reduction (120), and photosynthesis (126). It is likely that the atmosphere works as both a conduit and a functioning ecosystem based on this evidence (28), however, further investigation is required before this can be proven definitively.

The majority of metabolic activities in the atmosphere take place within clouds. Bacterial concentrations in cloud water have been described within the range of 10^3-10^5 bacteria ml^{-1} (127). The majority of cloud condensation nuclei (CNN) and ice nuclei (IN) bacteria such as

Pseudomonas syringae, Pseudomonas fluorescens and *Psychrobacter* sp. are psychrophilic. Bioaerosols, like standard aerosols (e.g. mineral dust, sea salt) aid the formation of clouds, by acting as both CCN in suitable warm conditions where relative humidity conditions exceed saturation, and IN at temperatures of -2°C and lower (128); CCN- and IN-associated bacteria also play a role in the initiation of precipitation events (128). The role of CCN and IN bacteria in precipitation has been touted bioprecipitation, a mechanism describing a feedback cycle which enables the wide dispersal of bacteria by wet deposition (129). These psychrophilic ice nucleators have been shown to respond significantly to environmental triggers such as changes in humidity and are ubiquitously present in abundances of 4-490 L^{-1} (127).

1.2.4. Atmospheric biodiversity

Whilst the atmospheric environment is relatively extreme, thriving diverse bacterial communities have also been found in other challenging environments such as hot springs and deserts (28). Microorganisms within the atmosphere are diverse, with airborne microbial communities above both terrestrial and marine environments having been shown to contain more than 100 genera of bacteria, a level of generic diversity comparable with that of soil and marine environments (130-132). One recent study by Barberán et al. (133), collated over 1000 sampling events, found more than 110,000 different species of airborne bacteria in the USA alone, with more than 55,000 species of fungi. This diversity stems from a multitude of adaptions to atmospheric life such as cell pigmentation (134), spore-forming ability (135), cryopreservation (136) and DNA repair mechanisms (137).

Bacterial populations can be seen to decrease in number by as much as half with increasing altitude, however the following viable bacteria and fungi have been found in the Stratosphere at altitudes as high as 77 km: *Mycobacterium luteum*, *Micrococcus albus, Aspergillus niger,*

Penicillium notatum, Circinella muscae, Papulaspora anomala, (110), *Bacillus simplex, Staphylococcus pasteuri* and *Engyodontium album* (138).

The majority of these diverse bioaerosol communities are largely comprised of four main bacterial groups which are the *Actinobacteria*, *Bacteroidetes*, *Firmicutes* and *Proteobacteria*, an observation consistent in both marine and terrestrial environments (24, 131). At the genus level, however, there is more variation, dependent upon environmental conditions, such as proximity to agricultural sites and weather (132). Variation is also directly influenced by season, however seasonal influence varies by location (139, 140). Concentrations of bacteria in the atmosphere generally range from 10^4 to 10^6 cells m^{-3} (141), however, these concentrations are known to vary significantly across all four calendar seasons and can also be affected by weather (wind direction, wind speed, temperature, fog etc.) (112, 139).

1.2.5. Polar aerobiology

The aerosolisation of microbes from cold environments such as the polar regions propels psychrophilic organisms directly into the atmosphere. Psychrophilic bacteria are better suited than most bacteria to atmospheric life as they are already adapted to survive the freezing temperatures of higher altitudes. Bacteria collected from clouds have been shown to be capable of growing and reproducing at 0 °C suggesting the existence of psychrophilic bioaerosols (111). With cloud temperatures often well below 0 °C, any bacterial species residing there, such as the recently discovered novel bacteria *Deinococcus aethius* and *Bacillus stratosphericus* (127), should be considered psychrophilic. Psychrophilic bacteria not only reside in the atmosphere but also play a key role in atmospheric processes, for example the psychrophilic plant pathogen *Pseudomonas syringae* is involved in ice nucleation in clouds.

1.2.5.1. Arctic

Aerobiological studies in the Arctic date back to the coining of aerobiology as discipline, with Meier and Lindbergh collecting aerial samples in flight above the Region (101); this work was followed up by Polunin et al. in the late 1940's (142), though studies of this nature are sparse. The most recent terrestrial study of bioaerosols in the Arctic was carried out by Harding et al. (143) on Ward Hunt Island. This study reported the communities in the air to have significant similarities with communities found in studies of the surrounding Arctic Ocean, drawing the conclusion that local sources contribute a large proportion of communities; the study also found organisms not normally associated with the high Canadian Arctic, from other sources and locations in the Arctic as well as some organisms associated with the Antarctic, supporting the theory of long distance atmospheric dispersal. These findings are consistent with those of previous studies that have stated the dominant groups of bacteria in cold ecosystems to be *Proteobacteria* (*Alpha-*, *Beta-* and *Gamma-Proteobacteria*), *Firmicutes*, *Bacteroidetes*, Cyanobacteria and *Actinobacteria* (144, 145).

1.2.5.2. Antarctic

Some of the first ecology-based aerobiological studies took place in the Antarctic in the early 1900's, however despite a significant sampling effort by Marshall et al. (146-148) there have only been 12 Antarctic aerobiology studies published since 1996. However, critically the transfer of biological material into Antarctica by atmospheric transport has been demonstrated (149, 150). Despite these findings, the small range of studies means that little is still known about the viability, duration of suspension and process of colonization and establishment of these organisms (151). Bacterial genera that are common in both aerial and Antarctic literature are *Staphylococcus, Bacillus, Corynebacterium, Micrococcus, Streptococcus, Neisseria* and *Pseudomonas*. Commonly encountered fungal genera include *Penicillium, Aspergillus,*

Cladosporium, Alternaria, Aureobasidium, Botryotrichum, Botrytis, Geotrichum, Staphylotrichum, Paecilomyces and *Rhizopus* (16).

Bacterial atmospheric residence times in the Antarctic are predicted to be longer than in other environments which implies that long range transport is more likely in the region (118). Evidence currently suggests there is an endemic population of bioaerosols in the atmosphere which are in part, but not entirely, related to the surrounding maritime and terrestrial conditions (152). These results are further supplemented by findings suggesting other characteristics of the Antarctic such as sea ice area may have a negligible impact on local biodiversity of atmospheric microbes (153). Along with long range atmospheric transport, one of the other key inputs of airborne microbes into the Antarctic atmosphere is human activity. The results of aerial studies taken from research stations such as Halley V Research Station, Concordia and Rothera Point have suggested the potential for input from human-derived sources whilst marine input into terrestrial samples is low, the most striking comparison across the majority of Antarctic studies, however, is that the biodiversity is markedly different (152-155). Although contradictory studies exist, another feature of aerial input into the region is that it might be directly affected by seasonality, which has potentially been correlated to an increase of keratinous material in summer regions due to increased bird and seal activity (148).

1.2.6. Biogeography, Microbiome interactions, and Human health

Microbial dispersal in the atmosphere has been considered ubiquitous in line with the hypothesis that "Everything is everywhere but the environment selects" (156). Initially, a significant amount of evidence supported this theory, where organisms with similar phylotypes were shown to be present in similar but geographically separated environments (157). However, recent developments in modern technologies and an increase in the number of studies of microbial communities across significant spatial and temporal scales, has led to the concept

of significant microbial biogeography, which would be in direct conflict with Baas Becking's theory and refers to patterns in the spatial distribution of microbial life from local to continental scales shaped by processes such as dispersal, speciation and extinction (33). Studies have since provided evidence for microbial endemism at local scales in Antarctica, for example, in isolated Antarctic habitats (153, 158, 159).

The atmosphere is key to microbial biogeography, particularly in the cold biosphere, as dispersal provides one of the main exogenous inputs into geographically isolated environments such as the polar regions, however factors such as dispersal, colonization and survival rates during atmospheric transport are poorly understood (23, 151). Little attention has been given to microbial diversity patterns in the atmosphere as the environment has been disregarded as a conduit rather than a habitat (28). Whilst recent studies have begun addressing these issues in the atmosphere, showing for example that marine bioaerosol communities can be distinct from those found in adjacent terrestrial locations (133, 160). Whether microbial biogeography in the atmosphere exists at all is still open to question and requires further research (34). Patterns in diversity have been observed in the atmosphere with genera such as *Polaribacter* sp. and *Psychrobacter* sp. being observed in both Arctic and Antarctic studies (143, 153) and the discovery that bioaerosols over urban environments contain typically higher diversities than those seen in remote locations (161).

Whilst clear patterns are beginning to emerge regarding the atmospheric dispersal of microorganisms and the viability of a number of organisms over extended periods of time in the atmosphere even under the pressures of the environment (113), studies still consistently fail to consider the viability of these source colonists upon arrival in their new environments (33). These colonists have the potential to interact with the microbiomes of the environments in which they are deposited both positively and negatively, for example bacteria suspended and deposited in low nutrient locations can provide nutrients by recycling, a mechanism which can

benefit the ecosystem (162), transversely this situation can also disrupt ecosystems causing events such as algal blooms to occur which can be devastating to the native community (163). Migrating bioaerosols pose a significant pathogenic threat to agriculture due to the heterogeneity of modern day crops (164). Human pathogens such as *Mycoplasma pneumonia, Mycobacterium tuberculosis, Corynebacterium diphtheria, Bordetella pertussis* and influenza virus also utilise the atmosphere as a conduit to spread from host to host. Incidences of both influenza and meningococcal meningitis have been described associated with long range transport during dust storms (165, 166) highlighting the relevance of the atmosphere in the spread of human disease. Today, there is a much wider range of techniques used in aerobiological studies which rely on either the impaction, impingement, membrane filtration, cyclonic or plate fall mechanisms (Table 1.1).

1.3. Aims and hypotheses

This study aimed to characterise the biodiversity of bacterial communities residing within the previously understudied Arctic and Antarctic atmosphere, using standard modern molecular techniques. Commonly used bioaerosol sampling devices were used to collect aerosolised bacteria at Svalbard, a Norwegian island within the Arctic, and on a ship which circumnavigated the oceans surrounding the Antarctic, stopping at a range of sub-Antarctic islands throughout the cruise. The spatial patterns of alpha diversity, beta diversity, and taxonomy, revealed by these sampling regimes were used to investigate the hypothesis that:

i) bacteria are ubiquitously present in the polar atmosphere as heterogeneous communities, due to the harsh selection pressures they face and the isolation of the environments relative to temperate regions

As there is no consensus on sample collection methodology for bioaerosols, a secondary aim of this study was to investigate the amount of variability imparted on bioaerosol communities by sampling methodology. Multiple sampling devices, utilising a range of sampling mechanisms, were employed in polar environments, to address the following expectation:

ii) Bacterial community profiles are not influenced by sampling methodology

Whilst addressing the initial aims, the high variability of datasets generated from similar samples, directed the study toward the proportion of variation in low biomass bioaerosol data that could be attributed to differences in sample processing and analysis methodologies. The most commonly used method of DNA extraction for bioaerosols was assessed by extracting DNA from membrane filters containing a known concentration of bacterial cells. The results of this study were used to explore the hypothesis:

iii) The Qiagen Powersoil DNA extraction kit is highly efficient at extracting bacterial DNA from samples over a range of bacterial concentrations at which bioaerosol samples would be expected to fall within (10^4 to 10^6 cells per m^3 of air (141))

Multiple kit negatives were extracted in order to determine whether kit negative community profiles were consistent as previously described within the literature. DNA extractions of known constitution and kit negatives were used to investigate the reproducibility of 16S amplicon profiles sequenced using Illumina MiSeq, in order to gauge the lowest biomass at which a sample can be successfully extracted and sequenced, addressing the hypothesis that:

iv) Illumina MiSeq is a suitable molecular tool for the investigation of low biomass air sample biodiversity

Table 1.1 Summary of the available aerobiological sampling techniques

Mechanism	Example of sampler	Flow rates (L min^{-1})	Collection media	Pros	Contras	Analysis techniques
Impaction	SAS SUPER 100/180/DUO360	<530	Contact plates, petri dishes, dry vessel, membrane filters (cellulose nitrate, cellulose acetate and PTFE)	- High flow rate/short sampling - Portable - Multiple collection media - Multiple downstream analysis options	- High cost - Desiccation - High flow rate collection bias	- Microscopy - Molecular - Culture
Membrane filtration	Sartorius MD8	~30	Membrane filters (cellulose nitrate, cellulose acetate and PTFE)	- Long sample periods - Multiple downstream analysis options - Wide range of filters - Easy sample storage - Duration does not affect viability - Low cost	- Not portable - Low flow rate / long sampling durations - Self assembly	- Microscopy - Molecular - Culture
Impingement	SKC Biosampler	~30	H_2O, PBS, mineral oil	- No desiccation - Portable - Viable samples - Multiple downstream analysis options - Multiple collection media	- High cost - Poor in cold environments	- Microscopy - Molecular - Culture
Drop plates	N/A	N/A	Agar plate	- Very low cost - Wide variety of sampling media - Viability shown	- Low proportion of microbes shown - Few analysis options	- Culture - Limited microscopy - Limited molecular
Cyclonic (wet and dry)	Bertin Coriolis μ	<300	H_2O, PBS, dry vessel wall	- Very high flow rate/short sample period - Increased viability (wet) - Portable	- Very high cost - Desiccation (dry)	- Microscopy - Molecular - Culture

N/A, not available; PTFE, polytetrafluoroethylene; PBS, phosphate-buffered saline

Chapter 2 - Methodology

2.1 Sample collection

Samples were collected in Newcastle upon Tyne using a membrane filtration setup (Figure 1.6) as follows. A Welch WOB-L vacuum pump (Welch, Mt. Prospect, IL, USA) connected by tubing to a Sartorius filtration unit (Göttingen, Germany) containing a 47 mm × 0.2 μm pore size cellulose nitrate membrane filter (GE Healthcare Life Sciences, Chicago, IL, USA) was ran for a duration between 3-12 hours at a flow rate of 20 L m^{-1}. Once collected, samples were placed in sterile 50ml falcon tubes and stored at -80°C for downstream analysis. Prior to sample collection, the flow rate of the vacuum pump was calibrated. This was done by placing the tubing attached to the vacuum pump working at maximum flow rate into an inversed 2L measuring cylinder and submerging the measuring cylinder into a beaker of water, the time taken to uptake 1.5L of water was then measured in triplicate and this value was taken as the maximum flowrate of the vacuum pump (figure 1.7).

Samples were collected in the Arctic using 3 sampling methods. The first was the aforementioned membrane filtration setup (figure 1.6). The second was passive accumulation onto agar plates; here, R2A agar plates (Sigma-Aldrich, St. Louis, MO, USA) were placed open for 15 minutes (figure 1.8), prior to being incubated for a duration of 10 days at room temperature. The third method utilised was active impaction onto gelatine filters contained within a portable AirPort MD8 (Sartorius, Göttingen, Germany) (figure 1.9).

Samples were collected during the Antarctic Circumnavigation Expedition (ACE) cruise using a membrane filtration apparatus (figure 1.6) as well as two cyclonic impingement devices (figures 1.10 and 1.11). The first of the cyclonic impingement devices was the Bertin Coriolis μ (Bertin Technologies, Montigny-le-Bretonneux, France), the collection cones were filled with sterile DNase and RNase free H_2O (Thermo Fisher Scientific), and the sampler ran at a

flow rate of 300 L m^{-1} for a duration of 50 to 60 minutes. Once sample collection was complete, the collection cones were wrapped in sterile zip lock bags and stored at -80°C for downstream analysis. The second cyclonic impingement device was an SKC Biosampler impinger (Eighty Four, PA, USA), samples were collected in 20ml of sterile DNase and RNase free H_2O, and at a flow rate of 12.5L m^{-1} for opportunistic durations between 2 to 9 hours. Once sample collection was complete, the liquid was transferred from the sampler to sterile 50ml falcon tubes, which were wrapped in zip lock bags and stored at -80°C for downstream analysis. Rainwater samples were also collected opportunistically during the ACE cruise; this was done by placing an ethanol sterilised funnel on top of a sterile falcon tube, the precipitation was then stored at -80°C for further analysis. All sample types were transported at -80°C from the location of collection to Northumbria University, Newcastle upon Tyne, UK.

Figure 1.6.

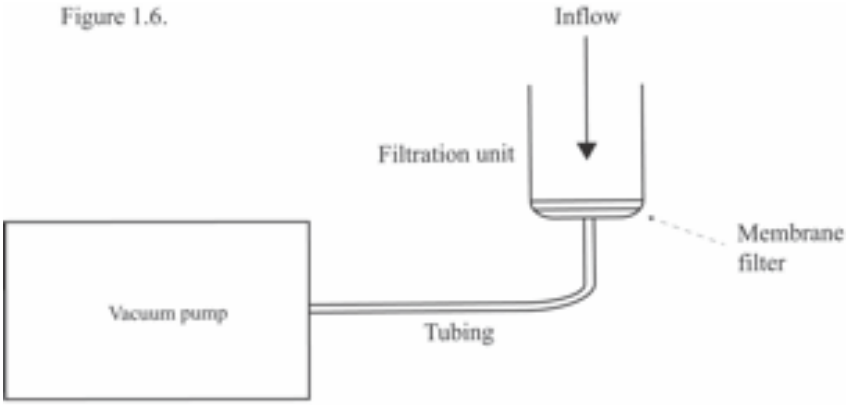


Figure 1.7

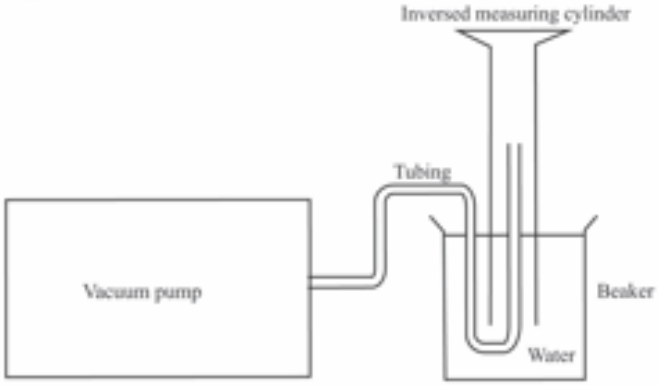


Figure 1.8

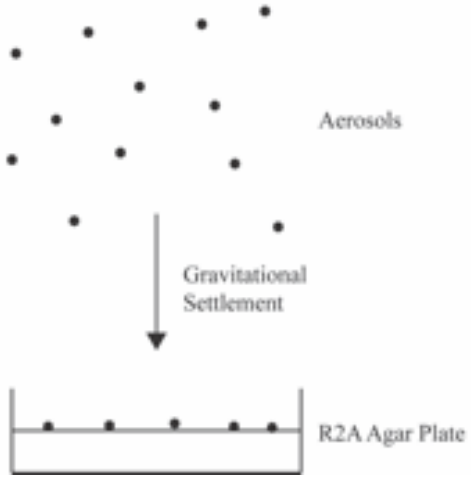


Figure 1.9

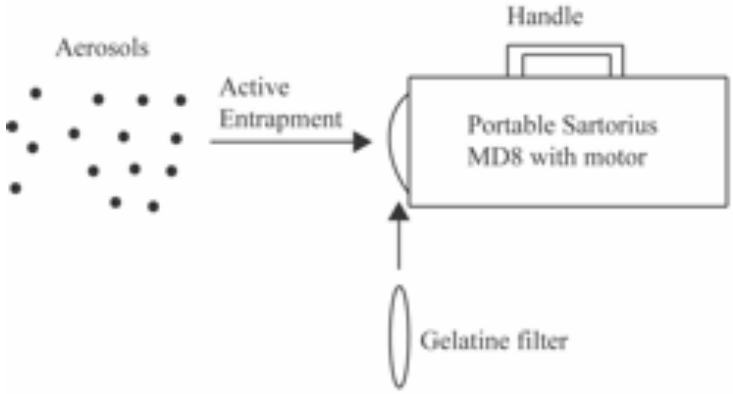


Figure 1.10

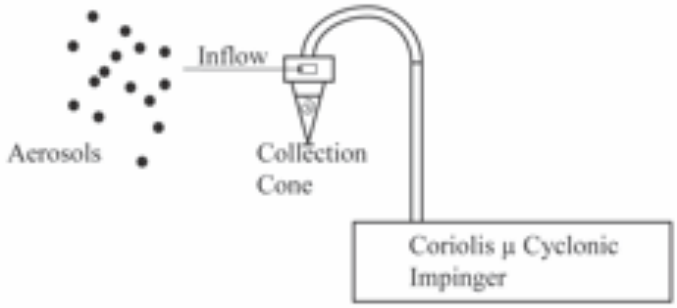
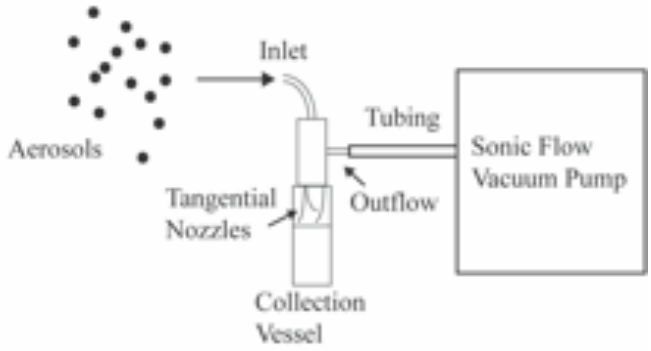


Figure 1.11



Illustrated sample collection mechanisms. Figure 1.6) membrane filtration setup, Figure 1.7) vacuum pump calibration, Figure 1.8) Passive accumulation, Figure 1.9) portable Sartorius MD8, Figure 1.10) Coriolis µ Cyclonic impinger, Figure 1.11) SKC BioSampler cyclonic impinger

2.2 Wet Lab Protocols

2.2.1 DNA Extraction protocol

All bacterial DNA extractions were carried out within an Envair Bio2+ class II microbiological safety cabinet (Lancashire, GB). Prior to all extractions, the cabinet was exposed to UV for 1 hour and wiped thoroughly with 10% NaClO in order to reduce the risk of sample contamination from prior cabinet use. Barrier pipette tips were used to minimise the risk of contamination during pipetting. Equipment was wiped with 70% ethanol prior to being placed into the cabinet to reduce the risk of personal contamination.

Extraction of bacterial DNA for all bioaerosol sample types was carried out using the Qiagen PowerSoil kit (Qiagen, Hilden, Germany) as per the manufacturer's instructions. This kit was chosen due to its frequent use in prior studies, cost effectiveness, and efficiency. The kit proved to be superior or on par with the Qiagen PowerLyzer PowerSoil kit (Qiagen, Hilden, Germany)(see appendix III), the Qiagen PowerWater kit (Qiagen, Hilden, Germany) and a non-kit Proteinase K based method (167). Whilst samples were stored together, extraction batches were chosen at random in order to inhibit any potential false patterns appearing within the dataset.

In order to facilitate the use of the Qiagen PowerSoil kit with all types of air sample, samples collected on filter papers were first dissected into quarters using an ethanol and flame sterilised scalpel and a sterile petri dish. The dissected quarter filter was then placed directly into a labelled bead tube for extraction. The remaining filter was stored at -20°C. For water based samples stored in falcon tubes and collections cones, samples were transferred to sterile 15ml falcon tubes and centrifuged for a duration of 20 minutes at 5000g. Following centrifugation, the supernatant was removed leaving 1ml, within which the formed pellet was re-suspended. This 1ml was then loaded directly into a labelled bead tube for extraction. Where samples

contained more than 15ml liquid, they were combined after centrifugation, and the previous steps were repeated.

The bead tubes in which samples were transferred into contain a buffer which begins dissolving humic acids and aids in protecting nucleic acids from degradation. Following on from loading the sample, 60µL of solution C1, previously warmed to 60°C to remove precipitate, was added. Solution C1 contains Sodium dodecyl sulphate (SDS) and other unnamed disruption agents which aid in lysing the cell walls of present bacteria. In addition to this, the high SDS concentration helps break down fatty acids and lipids associated with the cell membranes of several organisms. Samples were then briefly vortexed, before all samples were secured horizontally to a MoBio bead beating vortex adaptor pad (Carlsbad, CA, USA), prior to bead beating at maximum speed for 15 minutes. For instances where 24 bead tubes are attached at a time, it is suggested that a bead beating duration of 15-20 minutes is optimal; due to the low expected bacterial yield, the lower limit of 15 minutes was chosen in order to reduce the potential sheering of DNA. Cells are mechanically lysed during the bead beating phase, allowing the SDS and other agents to further lyse cell walls and break down membranes. Following bead beating, samples were centrifuged at 1000 x g for 30 seconds to form a pellet containing the beads and cell debris.

The entire volume of supernatant containing DNA, cellular proteins and some sample debris was then transferred to a sterile eppendorf where 250µL of solution C2 was added before samples were vortexed briefly and then incubated for 5 minutes at 4°C. Solution C2 is a highly saline solution which precipitates non-DNA organic and inorganic material including humic substances, cell debris and proteins. Samples were then centrifuged at 10000 x g for 60 seconds to pellet the inhibitor containing precipitate. 600µL of supernatant was then transferred to a sterile eppendorf, 200µL of solution C3 was added prior to brief vortexing at high speed and a

further 5 minutes incubation at 4°C. Solution C3 is a higher concentration inhibitor removal solution, and also increases the ion content of the solution to aid in the salting out of DNA in future steps. Following incubation, samples were then centrifuged at 10000 x g for 60 seconds at room temperature to pellet any remaining inhibitors, before 750µL of supernatant was transferred to a sterile eppendorf.

1200µL of solution C4 was then added to the supernatant before samples were briefly vortexed at high speed to homogenize. Solution C4 contains both isopropanol and guanidine salt; together, they facilitate the precipitation of DNA out of solution. 675 µL of this solution was then spun at 10000 x g for 60 seconds at room temperature, through a silica membrane containing micro spin filter in order to bind the precipitated DNA to the membrane. The remaining cellular components here are filtered out as the amount of isopropanol in the solution is insufficient to precipitate the remaining proteins out of solution. The entire solution was passed through the filter in 3 separate repetitions.

The liquid containing intracellular debris was discarded. A 500µL volume of Solution C5, an ethanol based solution with the purpose of removing any remaining salts from the filter membrane, was then loaded onto the silica based filter where the DNA was bound, and this was centrifuged at 10000 x g for 30 seconds. The flow through was then discarded, and the filter was then centrifuged again at 10000 x g for 60 seconds. The purpose of the second centrifugation is to remove any residual ethanol, as any ethanol carryover could inhibit downstream molecular analyses. The silica membrane was then carefully removed in order to make sure no ethanol is carried over and then placed into a sterile eppendorf.

The final stage of the extraction is to elute the membrane bound DNA. This was done by loading a 50µL volume of solution C6, a salt-free TE buffer solution, directly to the centre of the membrane, and centrifuging the membrane at 10000 x g for 30 seconds. The manufacturers

protocol at this stage suggests an elution volume of 100µL, however a lower elution volume was chosen to increase the concentration of DNA in the final solution. The spin filter was then discarded, and the DNA in solution was stored at -20°C for further analysis. At least one kit negative was undertaken per batch of extraction.

2.2.2 Thermo Scientific Nanodrop™

A Thermo Scientific Nanodrop™ spectrophotometer (Waltham, MA, USA) was used to gain an insight into the quality of DNA extracts. Following the successful loading up of the Nanodrop software, 2µL of Molecular Grade Nuclease Free H_2O was loaded onto the pedestal to initialise the device. 2µL of eluent (solution C6) was then loaded and used as the measurement blank. 2µL of each sample was then loaded directly onto the pedestal in triplicate and the absorbance of light at 260nm was measured. The output total DNA concentration in ng/µL was noted, as well as the A260/280 and A260/230 ratios measuring purity were noted. The sampling pedestal and arm were cleaned with sterile lens tissue between each sample loading. The ideal A260/280 ration is 1.8, whereas the ideal A260/230 ratio is expected to be in the range of 2.0 – 2.2. A low A260/280 ratio may represent residual reagents such as guanidine from extraction, whereas a high ratio may represent a poor blank. A high A260/230 ratio may be indicative of residual biological contaminants such as carbohydrates or reagents such as EDTA.

2.2.3 Concentrating of DNA using a rotational evaporator

Where samples showed low yields of DNA and/or poor A260/280 or A260/230 ratios, they were placed open into a Christ 2-18 CDPlus Rotational vacuum concentrator (Osterode am Harz, Germany) in order to evaporate off any residual contaminants, whilst leaving the DNA precipitated and in higher concentration due to the lower volume of eluent. The device was then set to 60°C and samples were ran for 10 minute intervals to make sure all of the sample

did not evaporate. Samples were vortexed at high speed to re-suspend precipitated DNA and stored at -20°C for further analyses.

2.2.4 Polymerase Chain Reaction

Polymerase chain reaction (PCR) amplification of aerial samples collected and extracted at Northumbria University, UK, was carried out in order to ascertain the success of DNA extraction methods by confirming the presence of an amplifiable concentration of bacterial DNA. All PCR reaction preparations were carried out in a PCR hood, namely the C.B.S Scientific Optimiser PCR Workstation (San Diego, CA, USA), in order to reduce the risk of contamination. The hood was exposed to UV for at least 60 minutes prior to use and cleansed with 10% NaClO. Filter tips were used to prevent pipette based contamination and full PPE was worn at all times. Samples and PCR reagents were stored on ice in order to reduce degradation. An Eppendorf Mastercycler (Stevenage, UK) was used to conduct PCR reactions. The universal 16S rRNA primer set 27F (5' AGA GTT TGAT CMT GGC TCA G 3') and 1492R (5' TAC GGY TAC CTT GTT ACG ACT T 3') were utilised at a concentration of 0.2µM for this reaction, as they span almost the full length of the gene.

A New England Biolabs (Hitchin, UK) *Taq* PCR kit was used. Standard 25µL reactions were made as follows:

2.5µL of 10X TAE Buffer

0.5µL of dNTPs

0.5µL Forward Primer

0.5µL Reverse Primer

0.125µL Taq DNA Polymerase

18.375µL Molecular Grade Nuclease Free H_2O

1.5µL $MgSo_4$

1µL template DNA (added at the end)

PCR master mix volume was calculated and prepared in a sterile 2ml eppendorf on ice as follows: volume of each component of PCR (excluding template DNA) x (number of samples to be amplified + 2 controls (both positive and negative) + 2 reactions to account for pipetting error). A negative reaction control of Molecular Grade Nuclease Free H_2O was used and the positive control used was DNA extracted from a subculture of *E. coli* K12. The master mix was then vortexed and pulsed in a centrifuge to gather all components in the bottom of the eppendorf, and held on ice. 24µL of vortexed master mix was then aliquoted out into each sterile labelled PCR eppendorf. The template DNA was then added to the master mix before samples were loaded into the thermocycler and amplified as per the program in table 1.2.

Cycle step	Temperature	Time	Cycles
Initial Denaturation	95°C	30 seconds	1
Denaturing Annealing Extension	95°C 53°C 68°C	30 seconds 60 seconds 90 seconds (60 seconds per kb)	30
Final Extension	72°C	300 seconds	1
Hold	4°C	∞	

Table 1.2. Standard PCR program for the amplification of the 16S rRNA gene.

Following amplification, samples were removed from the thermocycler and stored at -20°C for downstream molecular analyses.

2.2.5 Agarose gel electrophoresis of amplified bacterial DNA

In order to ascertain whether the PCR amplification of samples collected and extracted using a range of methodologies at Northumbria University, UK was successful, the amplified product was ran on a 1% (w/v) agarose gel. Briefly, 1XTAE solution was prepared by diluting a 50X stock solution 20-fold. The gel was then made by combining 0.5g of agarose powder with 50µL of 1XTAE solution and boiling by microwave until the powder was fully dissolved. Once slightly cooled, 5µL of SYBR Safe DNA gel stain was added (Eugene, OR, USA), the gel was poured into a casting tray and placed in a cold room at 4°C to speed up the setting process. The gel was then removed from the casting tray, placed in an electrophoresis tank, and submerged in 1XTAE. Pipette mixing was then used to combine 10µL of PCR product with 2µL of loading dye (Thermo Fisher Scientific, MA, USA), and this was loaded into a well. A 1kb plus DNA

ladder was loaded into the first well of each agarose gel in order to ascertain the size of the amplified product (figure 1.12). The gel was run at 120V for 45 minutes. Following this, the gel was removed from the tank and placed into a Bio-Rad Gel Doc™ XR+ system (Herts, UK), where it was visualised under a UV light. Successful amplification was noted where i) fragments were the correct length ii) there was no amplification of the negative control and iii) the positive control showed strong amplification.

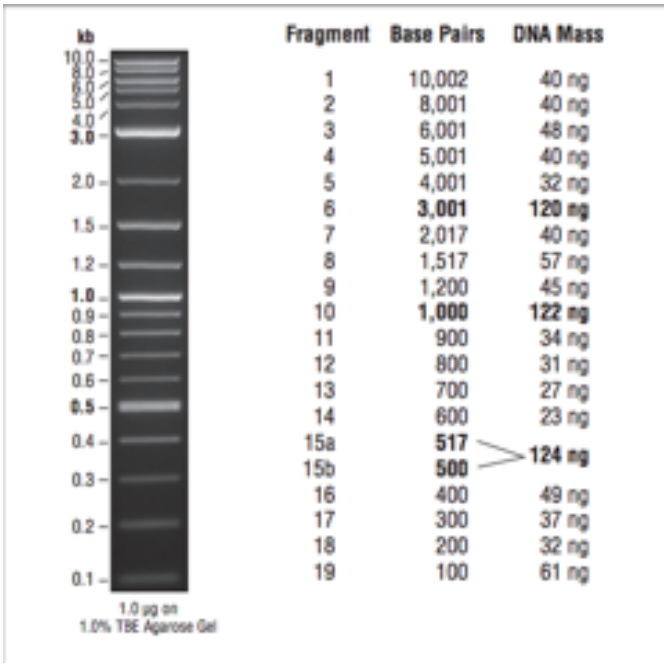


Figure 1.12. 1kb plus DNA ladder used for gel electrophoresis.

2.2.6 Quantification of double stranded DNA by Qubit

Due to the poor performance of the Nanodrop™ spectrophotometer for low biomass samples containing less than 10ng/µL of DNA, and its inability to differentiate between double stranded and non-double stranded DNA, the Qubit™ dsDNA HS Assay was utilised to quantify DNA extracts, as it is highly sensitive and can measure a range of sample concentrations from 10 pg/µl to 100 ng/µl, and is also selective for double stranded DNA, meaning PCR product could also be measured. First, one 500µl thin walled PCR tubes was labelled per sample, plus one

for the low standard (0 ng/µL in TE buffer) and one for the high end standard (10 ng/µL in TE buffer). The Qubit® dsDNA HS Reagent was diluted 1:200 in Qubit® dsDNA HS Buffer to make the working solution and vortexed briefly. 190µl of working solution was then added to the two PCR tubes labelled for the standards. 10µl of each standard was then pipette into each tube and vortexed. 195µl of working solution was added to each sample tube, along with 5µl of sample before being vortexed. The samples were then placed in a dark cupboard and incubated for 2 minutes. The standards were placed into the Qubit® 2.0 Fluorometer and read, before sample readings were taken. All preparations were carried out in a UV and bleach treated C.B.S Scientific Optimiser PCR Workstation.

2.2.7 16S rRNA quantification by Real Time qPCR

In order to ascertain the presence and total copy number of 16S rRNA genes within the aerial samples, a real time quantitative polymerase chain reaction (RT-qPCR) was carried out using the Qiagen QuantiNova SYBR Green kit (Hilden, Germany). Sybr Green is a non-specific fluorescent dye that intercalates between nucleotide bases and binds all double stranded DNA. All work was carried out in a pre-treated C.B.S Scientific Optimiser PCR Workstation and filter tips were used throughout.

The QuantiNova SYBR Green kit contains a Sybr Green master mix, consisting of the Sybr Green Dye, a dNTP mix, buffer, and QuantiNova DNA polymerase. QuantiNova DNA polymerase is a hot-start polymerase which allows the reaction to be set up at room temperature. Hot-start polymerase' contain a chelating antibody which remains attached until the initial denaturation. Primer set BACT1369F (5' CGG TGA ATA CGT TCY CGG 3') and PROK1492R (5' GGW TAC CTT GTT ACG ACT T 3') (168) were chosen; this set was chosen due the fact it spanned a 123bp region of the 16S rRNA gene. RT-qPCR reactions were

set up in DN-ase, RN-ase, and human DNA free white 8-well 200µL Biorad strip caps (Hercules, CA, USA). Reactions were set up as follows to a total reaction volume of 20µL:

Sybr Green master mix – 10µL

Forward Primer – 0.6µL (0.3µM)

Reverse Primer – 0.6µL (0.3µM)

Template DNA – 5µL

Molecular Grade Nuclease Free H_2O – 3.8µL

For each preparation, a master mix was made as follows: (the total number of samples x volume of each reagent + 6 standards + 1 NTC) x total number of replicates + 2 reaction' for pipetting error. All reagents and DNA was vortexed thoroughly prior to pipetting using a Vortex Genie 2 (Scientific Industries inc, Bohemia, NY, USA).

In order to carry out the RT-qPCR, standards were made containing a known quantity of the gene of interest. *E.coli* K12 was grown in LB broth at 37°C for 48 hours. 1ml of growth was then aliquoted into a sterile eppendorf and spun at 10000 x g for 1 minute. The supernatant was removed and one loop of the cell pellet was streaked onto an LB agar plate which was incubated for 48 hours at 37°C. Two filled loops of cells were transferred directly to multiple bead tubes, and DNA was extracted as described in chapter 2.2.1.

The DNA extracts were quantified by Qubit (chapter 2.2.6). The quantified value was then used to prepare 6 gDNA standards for the RT-qPCR reaction ranging from $3x10^1$ to $3x10^6$ copies of the 16S rRNA gene. The calculations required to prepare these standards are shown as a working example below. Once the $3x10^6$ standard was prepared, this was then diluted 1:10

five times to prepare the remaining standards. The standards and gDNA extract were then stored at -20°C for further use.

Working example of q-PCR standard preparation calculation using E.coli K12

Total weight of *E.coli* K12 genome ($5.08x10^{-15}$g) = Number of base pairs in *E.coli* K12 (4639221 bp) x average weight of a nucleotide base pair (x $1.096x10^{-21}$)

Total weight of *E.coli* K12 genome converted to pg ($1x10^{12}$) = $5.08x10^{-3}$

Weight of *E.coli* K12 genome containing 1 copy of 16S rRNA ($7.26x10^{-4}$pg) = Total weight of the *E.coli* K12 genome ($5.08x10^{-3}$ pg) / the number of copies of 16S rRNA in the entire *E.coli* K12 genome (7 copies)

Desired total copies of 16S rRNA in standard		**Weight of *E.coli* K12 gDNA containing 1 copy of gene (pg)**		**Required weight of *E.coli* K12 gDNA to be added to standard (pg)**
$3x10^{6}$	X	$7.26x10^{-4}$	=	2178
$3x10^{5}$	X	$7.26x10^{-4}$	=	217.8
$3x10^{4}$	X	$7.26x10^{-4}$	=	21.78
$3x10^{3}$	X	$7.26x10^{-4}$	=	2.178
$3x10^{2}$	X	$7.26x10^{-4}$	=	0.218
$3x10^{1}$	X	$7.26x10^{-4}$	=	0.022

Table 1.3. Calculations for the mass of gDNA required for each qPCR standard

Required weight of gDNA to be added to standard (pg)		Volume of DNA to be added to reaction (µL)		Required volume of gDNA stock in each standard (pg/µL)
2178	/	5	=	435.6
217.8	/	5	=	43.56
21.78	/	5	=	4.356
2.178	/	5	=	0.436
0.218	/	5	=	0.044
0.022	/	5	=	0.004

Table 1.4. Concentration of gDNA required to be added in a volume of 5µL to the sybr green reaction

Dilution of Stock gDNA extract of E.coli K12 to make 100µL of $3x10^6$ standard

$C_1V_1 = C_2V_2$

435.6 (pg/µL) x 100 (µL) / 1.5 x 10^4 (pg/ul) = 2.9µL of E. coli K12 gDNA + 97.1µL H_2O

Cycle step	Temperature	Time	Cycles
Initial Denaturation	95°C	120 seconds	1
2-step cycling Denaturing Combined Annealing/Extension	 95°C 60°C	 5 seconds 10 seconds	40
Hold	4°C	∞	

Table 1.5. Thermocycler conditions for qPCR.

Once samples were loaded into the thermocycler, the program was setup as shown in table 1.5. Cycle threshold (Ct) values were plotted against a log linear scale to create a standard curve. An r^2 value of >0.98 was used as a cut off for acceptable standards alongside a reaction efficiency of 90-110%. A linear regression formula was then implemented by the software to plot sample Ct values and assign a total copy number. The melt curve was also analysed to check for any non-amplicon specific binding such as primer dimer.

2.2.8 Illumina MiSeq Sequencing by Synthesis (performed in part by NUomics)

2.2.8.1 PCR amplification and tagmentaion of bacterial DNA

The Schloss wet lab SOP was used to prepare samples for sequencing (https://github.com/SchlossLab/MiSeq_WetLab_SOP/blob/master/MiSeq_WetLab_SOP.md) . All amplifications were carried out in a C.B.S Scientific Optimiser PCR Workstation.

Barcoded primers 515F (5’ GTG CCA GCM GCC GCG GTA A 3’) and 806R (5’ GGA CTA CHV GGG TWT CTA AT 3’) were used to target a 253bp amplicon of the 16S rRNA gene (169). PCR reactions were set up in individual wells of a 96 well plate for each sample plus a negative PCR control of Molecular Grade Nuclease Free H_2O and a ZymoBIOMICS Microbial Community Standard (Irvine, CA, USA) positive control as follows:

Accuprime DNA polymerase master mix - 17µL

Forward Primer - 1µL

Reverse Primer - 1µL

Template DNA - 1µL

The barcoded dual index primer set allowed for the multiplexing of samples on the run. Following the pipetting of reagents, the 96 well plate was briefly vortexed at high speed, then spun down to accumulate reagents and sample in the bottom of each well prior to placing the 96 well plate into a thermocycler. PCR conditions were as follows:

Cycle step	Temperature	Time	Cycles
Initial Denaturation	95°C	120 seconds	1
Denaturing Annealing Extension	95°C 55°C 72°C	20 seconds 15 seconds 300 seconds	30
Final Extension	72°C	600 seconds	1
Hold	4°C	∞	

Table 1.6. Thermocycler conditions for standard MiSeq amplicon library amplification.

2.2.8.2 PCR product clean-up

Samples were both normalised and cleaned using a SequalPrep Normalisation Kit (Thermo Fisher Scientific). The kit normalises to a maximum yield of 25ng where total mechanical saturation of amplicon product occurs to the silica walls of each well; the mechanical fixing of DNA to the wells of the Sequal plate filters out any excess reactants such as Polymerase or dNTPs. 18µL of barcoded PCR amplicon and 18µL of binding buffer was mixed in a well of a new sterile SequalPrep™ Normalization plate by vortexing, then centrifuged and incubated for 60 minutes at room temperature. Following incubation, the liquid was aspirated from the wells being careful not to scrape the sides of each well whilst pipetting. 50µL of SequalPrep™ Normalization Wash Buffer was then mixed in by pipetting to remove and residual unbound non-target biological material. 20µL of SequalPrep™ Normalization Elution Buffer was then added to each well and mixed by pipetting to elute the DNA from the silica walls of each wells.

Finally, the plate was vortexed briefly at high speed and centrifuged to gather all product in the bottom of each well.

2.2.8.3 Pooling and quality control of normalised PCR amplicon

Following on from normalisation and clean up, 5µL of product from each sample was combined into four separate pools in new wells of a 96 well plate. The Bioanalyzer high sensitivity DNA kit (Agilent technologies, UK) was used to assess fragment size in each pool in triplicate as follows; 9µL of gel was loaded into the gel loading wells of a Bioanalyzer chip and held under pressure for 60 seconds, the pressure was then released and gel was subsequently added to the remaining wells. 5µL of high and low standard marker was added to all sample wells and the DNA ladder well. 1µL of ladder was added to the ladder well, and 1µL of pooled amplicon added to the sample wells. The chip was then vortexed and placed into the Bioanalyzer device where fragment size was assessed and recorded. The DNA concentration of each library was assessed by Qubit (chapter 2.2.6). The fragment length measurement and the concentration measurement was then used to calculate the nM concentration of the library using the following equation:

$$[DNA]\, nM = \frac{[DNA]\ ng/\mu L}{660\ g/mol \times average\ fragment\ length\ (bp)}$$

The amplicons were then diluted to 2nM using Molecular Grade Nuclease Free H_2O and pooled together into a single library.

2.2.8.4 Sample loading and sequencing on Illumina MiSeq

The MiSeq V2 Reagent kit (500 cycles) was used to run the pooled library. Prior to loading the library into the MiSeq cartridge, 5µL of sequencing library was mixed with 5µL of 0.2N NaOH to denature the double stranded DNA into single stranded DNA. This mixture was then diluted

with 990µL of HT1 buffer, and this diluted mixture was then further diluted with 300µL of HT1 buffer to a final concentration of 5pM. 60µL of the concentrated library was removed, and 60µL of 5pM Phi-X was added. The entire 600µL 5pM library was then loaded into the cartridge. 3.2µL of sequencing and index primers were then mixed into each well by pipetting. The cartridge was loaded into the MiSeq. The flow cell, pre-cleaned with lint-free cloth and 18.2MΩ water followed by 100% ethanol, was inserted into the loading dock and the sequencing run was initiated.

2.2.9 Picogreen quantification of PCR amplicons

The Quanti-iT™ Picogreen dsDNA Assay (Invitrogen, Thermo Fisher Scientific, MA, USA), was used to quantify sequencing PCR amplicons. 200µL of 1X TE was prepared per sample from the provided 20X stock plus an additional 600µL for the DNA standards. 98µL of the prepared 1X TE was then added to each sample well of a black, thin walled PCR plate (Anachem, Manchester, UK). 600µL of DNA standard was then diluted 1:50 with 12µL of DNA and 588µL of 1X TE to make a 200ng/100µL solution. The DNA standard was then pipette mixed with 1X TE into the 96 well plate in the following concentrations:

Concentration (ng)	Volume of 1XTE (µL)	Volume of 200ng DNA standard (µL)
200	0	100
150	25	75
100	50	50
50	75	25
20	90	10
2	99	1

Table 1.7. Concentration dilutions for Picogreen standards

2µL of vortexed sample was then pipette mixed with the 98µL 1XTE. 100µL of 1:200 diluted Picogreen dye was added to each well containing sample or standard and pipette mixed thoroughly. 100µL of 1XTE and diluted Picogreen dye was also ran in two wells as a negative control. The plate was then incubated in the dark for 10 minutes prior to being excited at 485nm and 535nm and read on a plate reader. A standard curve was then created using the DNA standards, and the equation of the line of best fit was used to calculate sample DNA concentration.

2.2.10 Ampure XP bead cleanup of PCR amplicons

Ampure XP beads were used at a 1.0:1 ratio with PCR product to size select fragments of a similar size to the target amplicon in order to reduce primer dimer and non-target amplicons (figure 1.13). 10µL of PCR amplicon was combined in a sterile eppendorf with 10µL of vortexed Ampure XP beads and incubated at room temperature for 5 minutes. The eppendorf

was then placed on a magnetic stand to separate the bead fixed DNA from the liquid. The supernatant was then removed with care not to disturb the magnetised bead pellet. The pellet was then washed twice with 20µL 80% Ethanol and left to air dry in a PCR hood. DNAse/RNAse free water was then used to resuspend the DNA from the pellet.

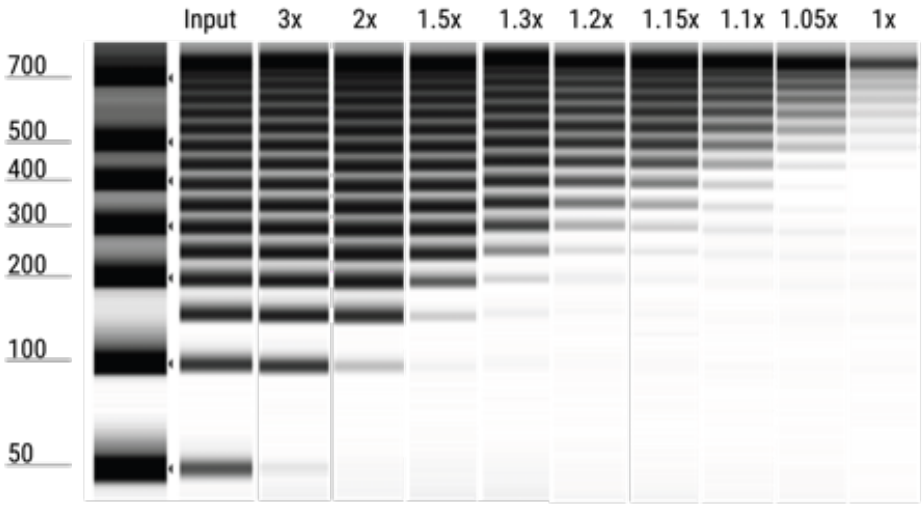


Figure 1.13. Ampure XP bead ratios for specific size selections

2.2.11 DAPI staining of filters

DAPI counts of liquid suspended samples were performed by first pipetting 1ml of sample into a sterile eppendorf and incubating with 100µL of 10µg/mL DAPI solution (Sigma-Aldrich, St. Louis, MO, USA) for 15 minutes. Once stained, the suspension was filtered onto a 0.2 µm pore size, hydrophilic polycarbonate membrane, 25 mm diameter (Merck KGaA, Darmstadt, Germany), placed onto a microscope slide containing a drop of immersion oil, and covered with a coverslip also containing a drop of immersion oil. The prepared slide was then wrapped in aluminium foil prior to viewing on a Nikon fluorescence microscope at 400nm. Negative control slides were made using DNA/RNAse free water and counts were subtracted from samples. Each slide was then imaged and cell counts were performed using ImageJ (170) to calculate a running mean; running means are calculated by plotting mean cell number against

total number of counts until the trend line is flat. Running mean values are then used to calculate total bacterial load per filter using the following calculations:

Field of view diameter = Eyepiece field number / total magnification

Field of view radius = Diameter of field of view / 2

Field of view area = πr^2 (π = pi (3.142), r^2 = radius of field of view squared)

Filter paper radius = Diameter of filter paper / 2

Filter paper area = πr^2 (π = pi (3.142), r^2 = radius of filter paper squared)

Fields of view per filter paper = Filter paper area / Field of view area

Total bacteria per filter = Fields of view per filter paper x Mean bacteria per field of view

2.3 Sample analysis

2.3.1 Sequence processing in QIIME2 and contaminant screening

Fastq files generated by 16S MiSeq sequencing were processed in QIIME2 version 2019.4 (61). Samples were imported into QIIME2 as an artefact in un-joined paired end format still containing their quality data. Following import, the QIIME2 artefact was screened in order to ascertain the number of sequences per sample and the quality of each base call. Samples were then truncated at base 247 as the quality of base call dropped considerably thereafter. The DADA2 pipeline (66) was then implemented from within QIIME2. Once samples have had the low quality base calls removed, DADA2 dereplicates the fastq files to reduce computing time whilst retaining the quality data for each dereplicated sequence. DADA2 then calculates per sequencing run error rates based on data features and uses these to infer sample composition by use of the dada algorithm to remove any substitution errors.

Once sequence variants have been inferred for both forward and reverse reads, they're combined, and any reads without a forward and reverse match are removed from the analysis. Singletons and chimeras are also removed using the DADA2 pipeline. Additionally, the DADA2 pipeline wrapped into QIIME2 removes any phiX reads. Following amplicon sequence variant assignment within DADA2, taxonomy was assigned using QIIME2's Naive Bayes classifier pre-trained on a Greengenes 13_8 99% OTUs database, trimmed only to include the V4 region of 16S. The taxonomy table and sequence variant tables were then uncompressed, formatted and used to screen for potential contaminant sequences as follows. Multiple reagent negatives including kit and PCR negatives were sequenced alongside true samples and the profiles of each negative control were used to screen true samples for negative controls using the following condition: in true samples where the total number of reads exceeded the total number of reads within each negative control by 2X for individual taxa, the taxa were retained, where the taxa were present at below 2X reads in a true sample vs negative control, the taxa were removed. This rule was followed on a per sample basis.

2.3.2 Biodiversity and statistical analyses in R Studio

Following decontamination in Microsoft excel, the table of sequence variants and taxonomic information was read into R Studio and transformed into a Phyloseq object (171). Taxa with sums > 0 were removed as well as any non-bacterial, mitochondrial and chloroplast taxa. Samples which summed to fewer then 1000 reads were also removed from analysis.

2.3.3 Calculating the mean, mode, median, range and standard deviation

The arithmetic mean was calculated as the sum of all numbers divided by the total numbers within the dataset. The mode was calculated as the most common number within the dataset. The median of a dataset was calculated by organising all values from high to low and taking the central value of the dataset. The range of the data was calculated by subtracting the smallest

number within the dataset from the largest. The standard deviation was calculated to measure the spread of the data from the arithmetic mean by calculating the mean, subtracting the mean from each value and squaring the result, calculating the mean of the squared differences and finally taking the square root of the mean.

2.3.4 Students t-test

The student's t-test (172) was used to show significant difference between two groups of continuous, normally distributed data, such as 16S rRNA copy number. This statistical test was implemented in the R '*stats'* package (R_Core_Team, 2014).

2.3.5 Kruskal-Wallis rank sum test

Means of continuous variables with non-normal distribution were compared using the Kruskal-Wallis rank sum test (173) employed in the R '*stats'* package (R_Core_Team, 2014). The test converts observations to ranks to normalise different samples sizes and account for abnormal-distribution. An alpha of < 0.05 was used determine significant differences.

2.3.6 Assessing sampling depth

Rarefaction curves were used to ascertain whether sufficient sampling depth had been achieved to fully describe per sample species richness by randomly subsampling each community prior to diversity analyses. Samples for which the rarefaction curve did not reach asymptote were removed.

2.3.7 Relative abundance

Following rarefaction, raw sequence variant counts were transformed to relative abundances for the purpose of normalisation. This method of normalisation accounted for difference in

sampling depth whilst additionally reducing sample wastage (88). The relative abundance of an individual taxa within a community was calculated as follows:

$$\frac{Total\ reads\ per\ taxa}{Total\ reads\ per\ sample} \times 100$$

2.3.8 Alpha diversity

Alpha diversity, or per sample diversity, was calculated using *vegan*' package for community ecology in R (Oksanen *et al.*, 2015). The alpha diversity metrics observed OTUs and Shannon index were calculated for each sample and plotted using the R packing '*ggplot2*' (Whickham, 2009). The observed OTUs metric is simply a count of the total number of taxa within a sample and is therefore a qualitative measure of community richness. Shannon's diversity index is a quantitative measure of community richness and is influenced most by sample evenness.

2.3.9 Beta diversity

Beta diversity, or between sample diversity was calculated using the '*vegan'* package for community ecology in R (Oksanen *et al.*, 2015). The Bray-Curtis dissimilarity index (Bray & Curtis, 1957) was used. This index is a quantitative index, making use of abundance data, to quantify the compositional dissimilarity between samples.

2.3.10 PCoA

Principle co-ordinate analysis (PCoA) was used to better visualise multi-dimensional data such as the Bray-Curtis dissimilarity matrix. Values and plots were produced using the Phyloseq package in R (171).

2.3.11 Heatmaps

Heat maps of the top 50 most relatively abundant taxa were generated using the Phyloseq package in R (171). The weighted Bray-Curtis dissimilarity beta diversity metric plotted as a PCoA was used to generate each heat map.

2.3.12 Differential abundance testing

Differential abundance testing was carried out using the DeSeq2 package in R (174). DeSeq2 fits negative binomial generalized linear models between groups and tests for significant difference using the Wald test, controlling false discovery rates (FDR) using the Benjamini-Hochberg method.

2.3.13 Permanova

Pairwise permutational analysis of variance (PERMANOVA) was used to check for significant dissimilarity between the Bray-Curtis dissimilarity of < 2 multivariate communities. The Adonis function of the Vegan package in R was used to carry out this statistical test (Oksanen *et al.*, 2015).

Chapter 3 - Characterisation of Arctic Bacterial Communities in the Air above Svalbard

3.1 Introduction

Microbial dispersal in the atmosphere represents a key biological input, directly influencing the gene pool (153). The dispersal rate of bacteria in the atmosphere has been shown to be directly linked to weather events, such as dust storms, that lift large amounts of microbial matter into the atmosphere (116). There are two mechanisms by which bacteria are transported through the atmosphere: free floating and attached to larger airborne objects. Free floating bacteria in the atmosphere are unlikely to come into contact with other microorganisms frequently; however, bacteria associated with larger airborne particles could be subject to increased horizontal gene flow (117). In fact, it is this horizontal gene flow and the abundance of bacteria within the atmosphere which has drawn attention to the environment as a potential source for new antibiotics (50).

Whilst several studies have focused on the movement of bacteria through the atmosphere, the majority of these studies have failed to consider the viability of these colonists upon arrival in their new environments (33). Microbial matter can be transported through the atmosphere potentially at a global scale, allowing long distance colonization. A large number of bacteria also remain viable for extended periods of time in the atmosphere, even under intense selection pressure (113). These viable microorganisms carry out multiple functions whilst suspended in the atmosphere; these include cloud formation by ice nucleation (120, 175), nitrogen processing (125), the degradation of organic carbon-based compounds (124) and photosynthesis (176). Viable colonists have the potential to interact with microbiomes at the site of deposition in an antagonistic or synergistic way. For example, suspended nitrifying bacteria that are deposited in nutrient poor locations could provide a novel source of nutrients

benefitting the ecosystem; conversely, the same mechanism can prove disruptive in other circumstances, causing toxic algal blooms, which can be devastating (163). Migrating bacteria also pose a potential pathogenic threat to human health, global ecosystem stability (166, 177, 178) and agriculture due to the homogeneity of modern day crops (164).

Atmospheric bacterial abundance generally ranges from 10^4 to 10^6 cells per m^3 (141), but, this varies throughout the year (139), and can be affected by weather (wind direction, wind speed, temperature, etc.) (112). Bacterial abundance can decrease by as much as half with increasing altitude, although viable bacteria have been found in the stratosphere at altitudes as high as 7.7 km (116, 179). Bacteria found in the atmosphere are diverse. Airborne bacterial assemblages in both terrestrial and marine environments contain more than 150 genera of bacteria (130-132), a level of diversity comparable to other nutrient poor environments such as Antarctic snow which has been shown to contain in the region of 250 genera of bacteria (180). Barberán et al. (133), collated over 1000 sampling efforts and found more than 110,000 different species of airborne bacteria in the USA alone.

Most bacterial communities in the atmosphere comprise four main phyla: *Actinobacteria*, *Bacteroidetes*, *Firmicutes*, and *Proteobacteria*, a fact that remains consistent in the atmosphere surrounding both marine and terrestrial habitats (24, 131). However, aerial microbial diversity at genus level is more variable and depends on environmental conditions, such as proximity to agricultural sites, meteorological conditions and season (132, 139).

Patterns of diversity in airborne bacterial communities are central to the emerging field of atmospheric biogeography. Indeed, until relatively recently whether microbial biogeography existed in the atmosphere at all was contentious (34). However, an increasing number of studies have shown the inter-continental dispersal of bacteria across continents separated by both political (Europe and Asia) and geographical (North America and Asia) borders (133, 181). Furthermore, distinct geographical features give rise to distinct airborne microbial

communities, for example marine coastal communities are different to continental terrestrial ones (133, 160). Despite these findings, atmospheric biogeography has received little attention as the atmosphere is considered a transport route rather than a stable habitat (28). The development of aerobiology as a field and improved techniques should help understand whether at the ecological level, microbes interact and evolve within the atmosphere, as they do in other habitats.

The Arctic can be defined as the area above the Arctic Circle. The Norwegian Arctic archipelago of Svalbard is one of the northernmost inhabited locations in the world at 79° N. Svalbard is characterised by its remarkably low human population with only 2185 registered Svalbard inhabitants in 2015 (182). This low population density translates into reduced anthropogenic environmental alterations such as those linked to agriculture. Thus, the Arctic represents an optimal location to study natural patterns of airborne dispersal and its influence shaping natural communities. Aerobiological studies in the Arctic date back as far as the late 1940s (142). Studies of this nature are sparse between these early efforts and the present, with very few studies taking advantage of novel molecular techniques.

To the best of our knowledge, the only recent terrestrial study of bioaerosols (airborne particles of biological origin) in the Arctic was carried out by Harding et al (143), on Ward Hunt Island located in the Canadian high Arctic. Harding et al. found similarities between air and snow communities and those bacterial communities found in the surrounding Arctic Ocean, drawing the conclusion that local sources are the largest contributors which influence bacterial community assemblages. Their study also found organisms not normally associated with the high Canadian Arctic, microbes from other Arctic locations, as well as some Antarctic microorganisms, supporting the theory of long distance atmospheric dispersal. These findings are consistent with those of previous studies that have stated the dominant groups of bacteria in cold ecosystems to be *Proteobacteria* (*alpha*, *beta*, and *gamma*), *Firmicutes*, *Bacteroidetes*,

and *Actinobacteria* (144, 145). However, while aerobiological studies in the Arctic are scarce, the number of studies in the Antarctic has increased (150, 153). To this end, a comparative analysis of aerobiological data over the Arctic and the Antarctic will allow the study of bipolar diversity and potentially, the global atmospheric distribution of microbes.

Organisms in the Arctic atmosphere are exposed to extremely low temperatures and hurricane strength winds, seasonal freeze–thaw cycles, extreme exposure to UV and extremely low levels of nutrients. Thus, organisms inhabiting this region are referred to as extremophiles and tend to exploit features such as the ability to form spores, which allow them to survive the harsh conditions. Similar to those microbes inhabiting the Arctic, organisms surviving in the atmosphere also endure extreme temperatures, UV exposure and poor nutrient levels.

Sampling techniques for terrestrial and aquatic microbial ecology studies are highly variable but based on common principles, established and used consistently. In contrast, a wide range of techniques are currently available in aerobiology, despite the low number of studies in the field. In general, these sampling methods involve impaction, impingement, membrane filtration or the drop plate mechanism, the results of which are not directly comparable due to strong methodological biases. Furthermore, the strength of the bias is still unknown, due to the lack of studies comparing different methodologies, although recent efforts have been made towards establishing a standard methodology (151).

Analytical techniques can also vary considerably among studies, compromising comparability even further. To date, most aerobiological studies use colony-forming units (CFU) count per unit volume of air sampled to measure the density of cultivable microorganisms in the atmosphere. These studies report density changes over space, time and varying environmental conditions; however, culture based studies only provide a partial picture of the overall microbial diversity (28). Culture dependent studies are also biased towards gram-positive bacteria, while molecular based studies show the opposite trend, with a large proportion of

gram-negative bacteria populating the aerial environment (23). For this reason, fluorescence microscopy is increasingly used for cell counts and taxonomic identification, combined with molecular techniques such as high throughput sequencing. Temporal, spatial and meteorological variations also lead to differences in the aerial communities identified (25, 161), reducing further the ability to describe biogeographical patterns.

Set against this background, in this study, the influence of different sampling techniques, sampling location and total sample volume on the identification of aerial bacterial communities in the Arctic was explored in order to test the following hypotheses, based on culture dependent and independent analytical methods, thus presenting a preliminary picture of the microbial community in the air over Svalbard.

i) Bacterial communities are ubiquitous in the atmosphere around Svalbard can be accepted
ii) Bacterial communities in the air above Svalbard are homogeneous
iii) sampling methodology does not impact the seen biodiversity of Arctic bioaerosol communities

3.2 Methodology

3.2.1 Site Description

Airborne microbial samples were collected in July 2015 above Svalbard (Figure 1.14). Svalbard is home to a relatively small human population and plays host to very few mammals. The majority of the human population of Svalbard resides in Longyearbyen; implying that, were samples subject to human influence, it would most likely occur here. The west coast of Svalbard is influenced by the Atlantic Ocean and is affected by warmer currents than the East Coast, oriented towards the Barents Sea.

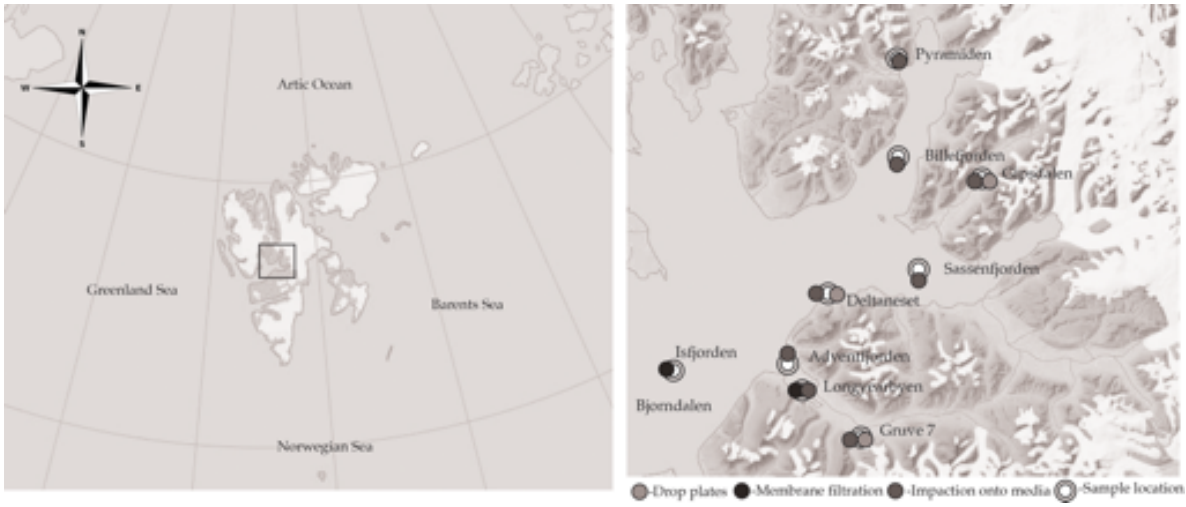


Figure 1.14. Svalbard location and sampling sites (map adapted with courtesy of the © Norwegian Polar Institute (http://www.npolar.no/no/).

Samples were collected between 6 and 23 July 2015 above both marine and terrestrial locations using a range of techniques (Table 1.8). Marine samples were collected aboard the research ship (Viking Explorer) and aboard a zodiac. The terrestrial sites were on the roof of The University Center in Svalbard (UNIS (78°13′ N, 15°39′ E)) located in central Longyearbyen, Mine (Gruve) 7, Deltaneset, Gipsdalen and Bjørndalen; these locations were chosen to represent a large terrestrial geographic range. The marine sites were located in the surrounding fjords at Billefjorden, Isfjorden, Sassenfjorden and Adventfjorden bay (Figure 1.14).

Sample Location	Environment	Sampling Mechanism	Date	Flow Rate (L m^{-1})	Duration (min)
Bjørndalen	Terrestrial	Drop plates	13 July 2015	-	15
Deltaneset	Terrestrial	Impaction onto media	13 July 2015	50	20
		Drop plates	13 July 2015	-	15
Gipsdalen	Terrestrial	Impaction onto media	13 July 2015	50	20
		Drop plates	13 July 2015	-	15
Longyearbyen	Terrestrial	Impaction onto media	16 July 2015	30, 50	20, 40, 60, 80
		Membrane filtration	06, 19, 21–23 July 2015	~20	30, 60, 120, 300, 3 days
Mine (Gruve) 7	Terrestrial	Impaction onto media	13 July 2015	50	20
		Drop plates	13 July 2015	-	15
Adventfjorden	Marine	Impaction onto media	17 July 2015	50	20
Billefjorden	Marine	Impaction onto media	17 July 2015	50	20
Isfjorden	Marine	Membrane filtration	11 July 2015	~20	480
Sassenfjorden	Marine	Impaction onto media	17 July 2015	50	20

Table 1.8. Summary of sample locations and regimes

3.2.2 Meteorological data

Seven-day back trajectory models were calculated for sampling days where sequencing was carried out at air mass arrival heights of 10 m, 500 m and 1500 m (Figure 1.15) using National Oceanic and Atmospheric Administration (NOAA) Hysplit Model (183) and the Global Data Assimilation System (GDAS1) archived data file. In general, pockets of air at all altitudes arrived from a northerly (Arctic Ocean) direction, however high altitude air pockets at 1500 m were more easterly influenced than the lower altitudes. On 6 July, the low altitude air masses (10 m, 500 m) were easterly. Temperatures averaged 8°C across all sampling days with only one precipitation event totalling 0.1 mm occurring on 17 July. Wind speed varied between 10 and 22 kmh^{-1} and humidity averaged 67% (Table 1.9).

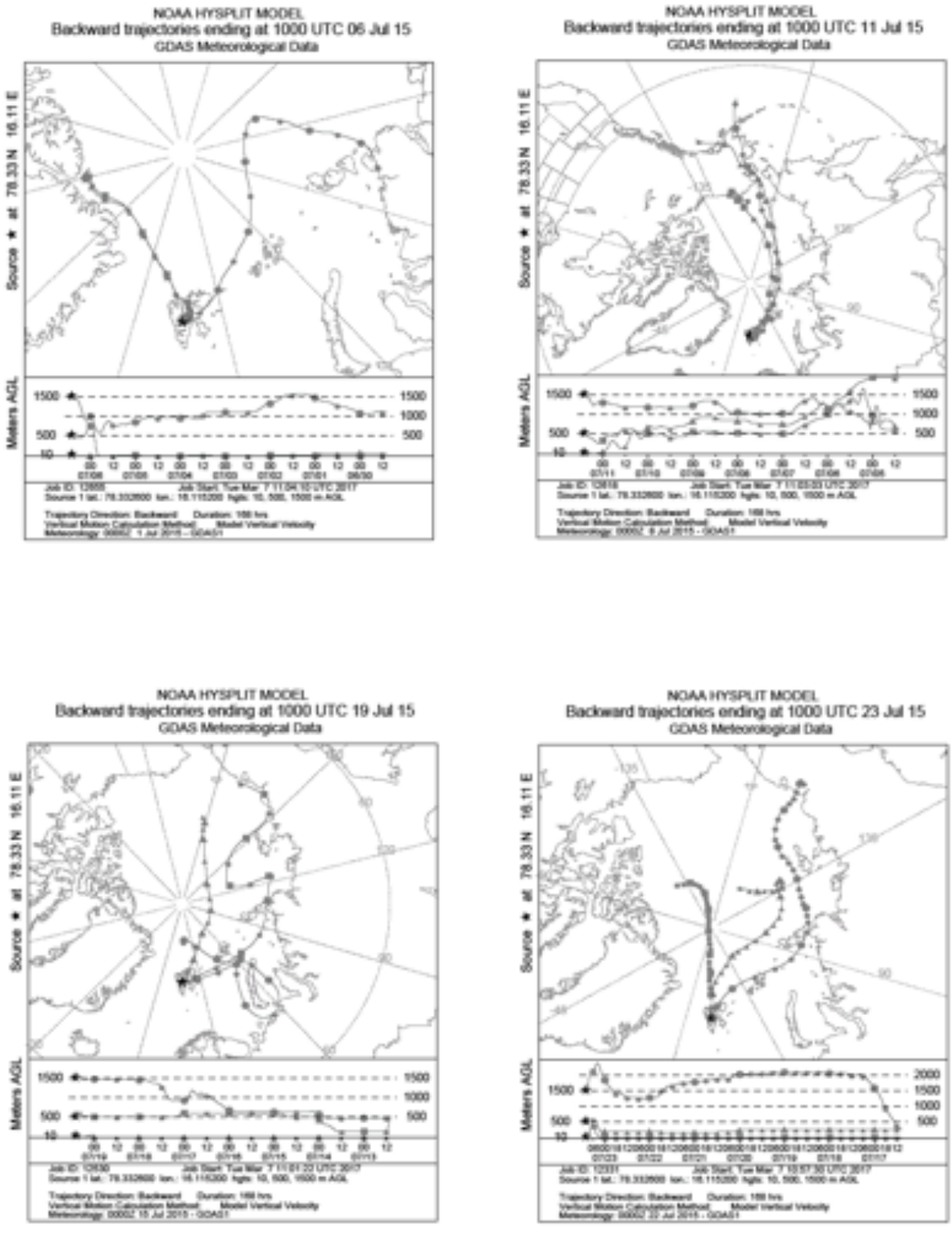


Figure 1.15. Back trajectory models were calculated using the NOAA Hysplit Model. Three arrival heights were used 10 m (transect marked by triangles), 500 m (transect marked with squares) and 1500 m (transects marked with circles). Sampling location is marked by a black star.

	06 July 2015	11 July 2015	13 July 2015	16 July 2015	17 July 2015	19 July 2015	21 July 2015	22 July 2015	23 July 2015	21–23 July 2015 (Average)	Average across All Sampling Days
Average temperature (°C)	8	10	8	6	8	8	10	8	6	8	8
Total precipitation (mm)	0	0	0	0	0.1	0	0	0	0	0	0
Average wind speed (kmh^{-1})	13	20	10	20	14	22	18	12	12	14	16
Average humidity (%)	63	75	90	57	68	59	65	68	61	65	67
Pressure (hPa)	1025	1022	1019	1009	1013	1015	1015	1013	1006	1012	1016

Table 1.9. Meteorological conditions on sampling days at Svalbard airport (The Weather Company (Atlanta, GA, USA))

3.2.3 Culture dependent

Drop plates containing R2A media (Sigma-Aldrich, St. Louis, MO, USA) were placed open at Gipsdalen, Mine (Gruve) 7, Deltaneset and Bjørndalen for 15 min; plates were incubated for 10 days at room temperature; following incubation the plates had colony counts and distinct colony counts taken.

Additionally, a portable AirPort MD8 (Sartorius, Göttingen, Germany), comprising a disposable gelatine filter membrane, was used to compare sampling efficiency and cultivability at two flow rates and different sampling volumes. Sampling sites were chosen to compare with terrestrial plate drop sites but also to assess for the differences at marine sites. Terrestrial samples were collected at Mine (Gruve) 7, Deltaneset, Gipsdalen and central Longyearbyen (UNIS roof) and marine samples at Billefjorden, Sassenfjorden and Adventfjorden, respectively. The sampler was used at respective flow rates and durations ranging 30–50 L m^{-1} and 20–80 L m^{-1} on 13, 15, 16 and 17 July 2015. The gelatine filters collected at all sites were placed directly onto the surface of R2A agar plates (Sigma-Aldrich, St. Louis, MO, USA). These plates were then incubated at room temperature for 10 days. Total CFU and distinct colony numbers were counted.

3.2.4 Culture independent

As gelatine filters are not amenable to culture independent techniques (due to the presence of gelatine), airborne bacteria from both terrestrial and marine sites were collected via membrane filtration. A Welch WOB-L vacuum pump (Welch, Mt. Prospect, IL, USA) was set up at a flow rate of ~20 L m^{-1} connected to Sartorius filtration unit (Göttingen, Germany) containing a 47 mm 0.2 μm pore size cellulose nitrate membrane filter (GE Healthcare Life Sciences, Chicago, IL, USA).

A marine sample was collected at Isfjorden on the 11 July 2015 with a respective sample duration and volume of 8 h and ~9600 L and the terrestrial sample was taken in central Longeyearbyen (UNIS roof) at the following dates, durations and volumes, respectively: 6 July 2015 for 30 (600 L), 60 (1200 L), 120 (2400 L) and 300 (6000 L) min; 19 July 2015 for 30 (600 L), 60 (1200 L), 120 (2400 L) and 300 (6000 L) min; and 21–24 July 2015 for three days (~86,000 L) continuously (Table 1.8).

The cellulose nitrate membrane filters were sent to MrDNA (MrDRNA, Shallowater, TX, USA) for extraction and sequencing. DNA was extracted from samples using the MoBio PowerSoil kit (MoBio, Vancouver, BC, Canada) following the manufacturer's protocol with an additional 1 min of bead beating to account for the filter paper. Extracted samples were then amplified using 16S rRNA universal primers 27Fmod (AGRGTTTGATCMTGGCTCAG) and 519Rmodbio (GWATTACCGCGGCKGCTG) and barcodes were attached at the 5′ end. A 28-cycle PCR using the HotStarTaq Plus Master Mix Kit (Qiagen, Germantown, MD, USA) was carried out under the following conditions: 94 °C for 3 min, followed by 28 cycles of 94 °C for 30 s, 53 °C for 40 s and 72 °C for 1 min, after which a final elongation step at 72 °C for 5 min was performed. After amplification, PCR products were checked in by running a 2% agarose gel in 1X TAE buffer to determine amplification success.

Samples were then pooled based on their molecular weight and DNA concentrations, purified and Illumina DNA libraries were prepared. Paired end sequencing of the V4 region was then performed on a MiSeq following the manufacturer's guidelines. The resultant data were analysed using QIIME v1.9.1 (59). The 776,315 raw sequence reads were quality trimmed and checked for chimeras using USEARCH 6.1 (63), clustered at an identity threshold of 97% and assigned to Operational Taxonomic Units (OTUs) using UCLUST (63) and the Greengenes reference database (71) was used to assign taxonomy. Sequences were then aligned using PyNAST (77) and a phylogenetic tree was built using FastTree (81).

3.2.5 Statistical analysis

Statistical analyses were performed using PAST (184) to test for differences in means, medians, variances and distributions and MS Excel (2013) to calculate correlation coefficients, the coefficient of variance and produce graphs of the analyses; statistical tests were carried out at an assumed significance of alpha: 0.05. When calculating diversity indices, to avoid statistical bias due to differences in sequencing depth all samples were normalised to a depth of 26,190 reads. Rarefaction curves, diversity indices (Shannon and Simpsons reciprocal), Bray–Curtis OTU and unweighted UniFrac phylogenetic distance metrics, and PCoAs were produced using QIIME (59).

3.3 Results

3.3.1 Culture Dependent

Viable bacteria were found in all of the samples. Clear differences were apparent in the mean CFUs from the two culture dependent methods used (Figure 1.16). A Kruskal–Wallis test for equal medians of CFUs and morphologically distinct CFUs was undertaken to assess the drop plate replicates, the result did not show significant differences (Kruskal–Wallis plate fall CFU: $p = 0.095$, plate fall morphologically distinct CFUs: $p = 0.123$).

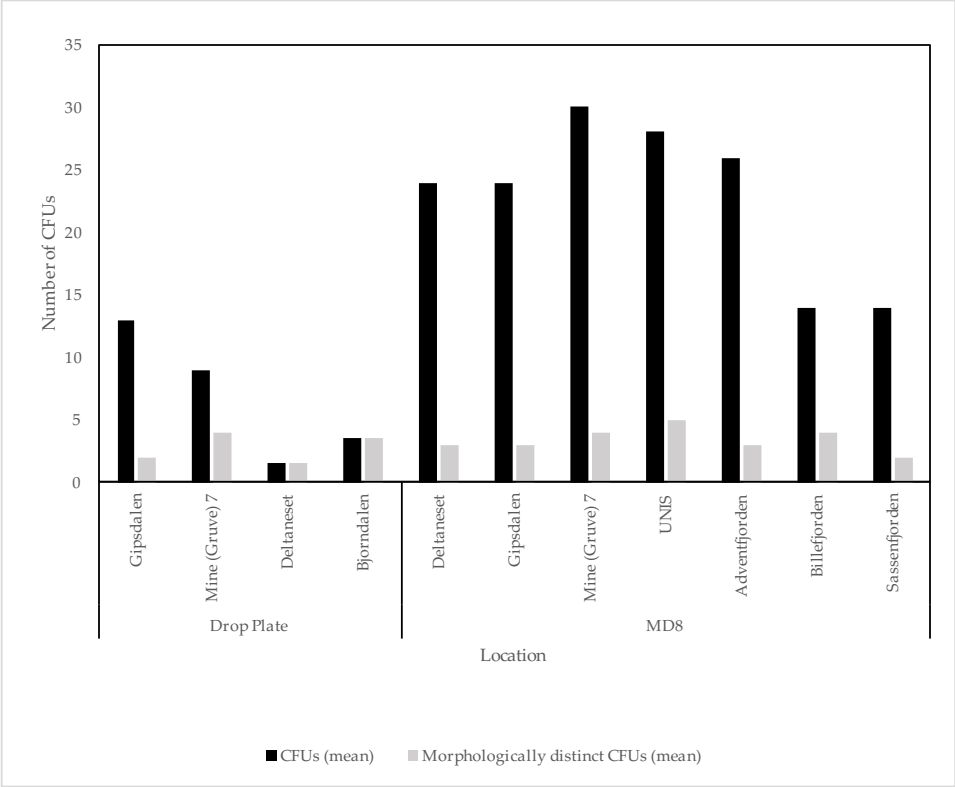


Figure 1.16. Mean colony-forming units (CFU) and morphologically distinct CFU counts for drop plate and Sartorius MD8 data

Comparing the differences in drop plate and MD8 results for the locations where data were available for both methods, the MD8 showed much higher CFU yields, although this difference is not obvious when looking at the number of morphologically distinct CFUs i.e. CFUs of different appearance (Figure 1.16). Statistical analyses show a significant difference of mean CFUs sampled at the same location using different methods ($p < 0.05$); no differences in variances, medians or coefficient of variations, but a significant difference in equality of distributions (Kolmogov–Smirnov: $p < 0.05$). For the morphologically distinct CFUs, however, there were no significant differences for any of the mentioned parameters. When looking at the overall variance and efficiency of both culture dependent methods, only considering the MD8 samples collected at 50 L m^{-1} for 20 min i.e., 1000 L sampling volume (Figure 1.17), there were obvious differences in the mean CFUs, but not for morphologically distinct CFUs.

An independent *t*-test comparing the two methods showed a significant difference in the mean CFUs from drop plate and MD8 samples ($p < 0.001$), no significant differences in variances, but with significant differences in coefficients of variation ($p < 0.005$), medians (Mann–Whitney U $p \leq 0.001$) and distributions ($p < 0.001$). Looking at the statistical analysis of the morphologically distinct CFUs, there was no significant difference in the means from drop plates and the MD8 ($p > 0.05$), with no significant differences in variances, medians, distributions, or coefficients of variation. These results show that there is a significant difference between the two methods, the MD8 yielding a larger number of CFUs.

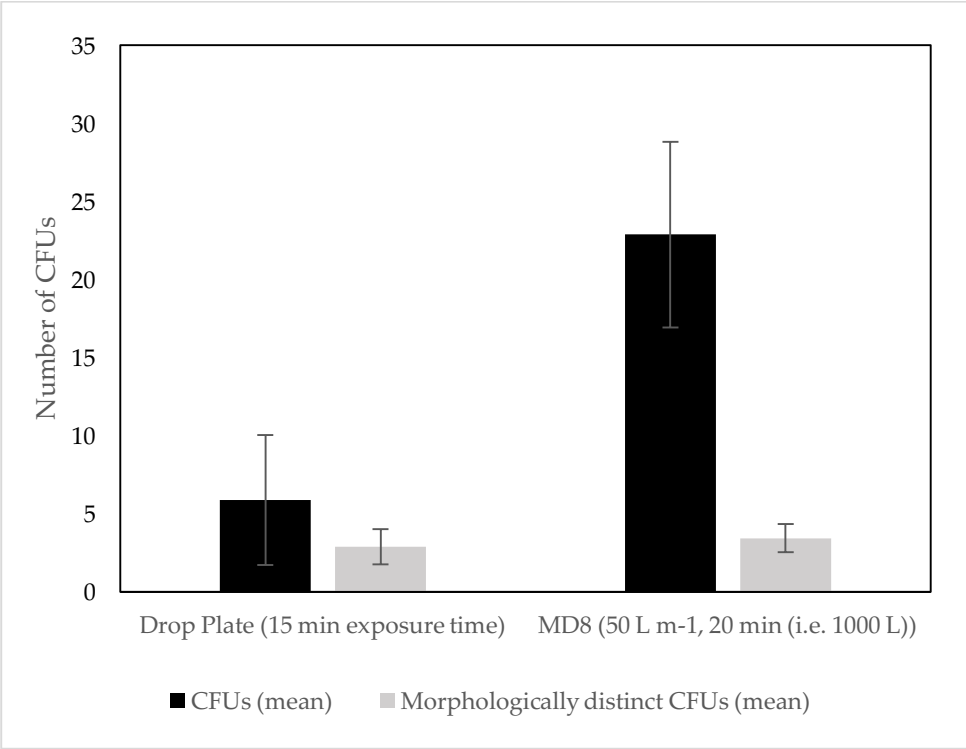


Figure 1.17. Mean CFUs and mean morphologically distinct CFUs for drop plates and 1000 L MD8 samples

Comparing MD8 results for terrestrial and marine samples, the mean CFUs were lower at marine sites than terrestrial sites (Figure 1.18). Statistical analysis showed no significant difference in CFUs for marine and terrestrial sites ($p = 0.070$). This also held for the morphologically distinct CFUs, there was no significant difference for either of the mentioned parameters between terrestrial and marine sites.

To account for the variable scales across the samples, the normalised coefficient of variation was used, and this showed the highest variation in the CFUs in the drop plate and MD8 samples collected at UNIS. In the MD8 marine sample the variability of morphologically distinct CFUs was the highest. The CFUs at the terrestrial sites varied the least (Figure 1.19). At UNIS, where volume and flow rate were varied, a clear trend of increasing CFUs with increasing volume was evident (Figure 1.20). The 30 L m^{-1} flow rate sample, however, had slightly higher CFUs, despite lower volume. There was a clear correlation between CFUs and the sampled volume of air ($R^2 = 0.933$, Figure 1.20A), not including the 30 L m^{-1} sample, and still a very high positive correlation of ($R^2 = 0.906$) when this sample was included. For the morphologically distinct CFUs, there was a slight negative correlation ($R^2 = -0.256$) in number of CFUs with increasing volume (Figure 1.20B), leaving out the exceptional value of 30 L m^{-1} showed a considerable negative correlation ($R^2 = -0.640$).

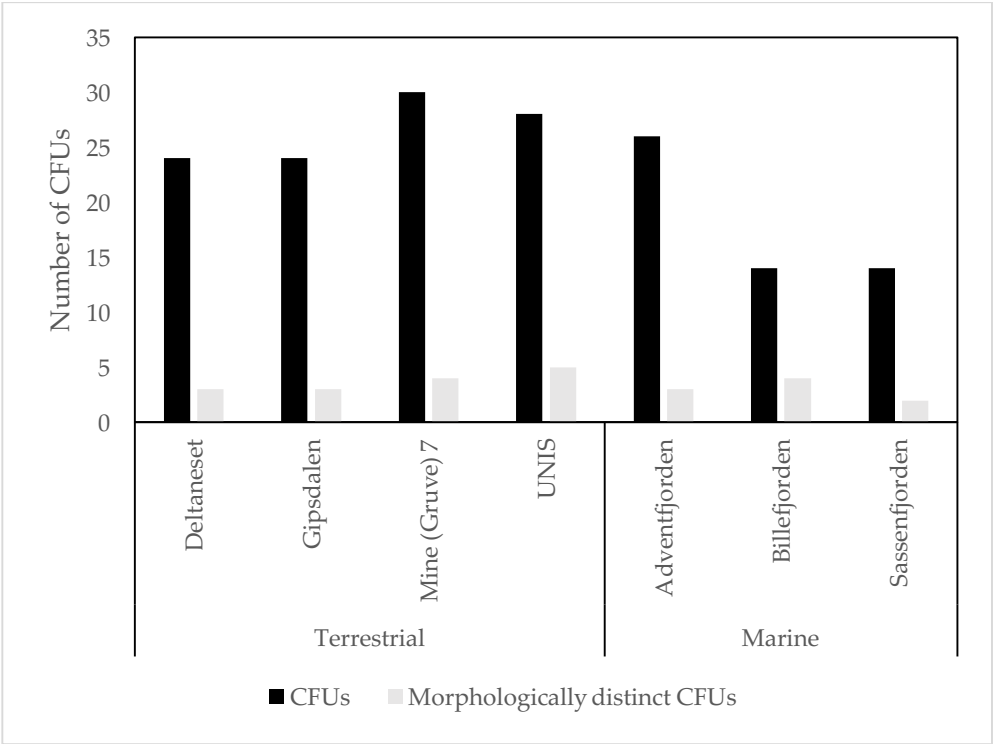


Figure 1.18. Counts of total (black) and morphologically distinct (grey) CFUs in each sample separated by environment (terrestrial and marine)

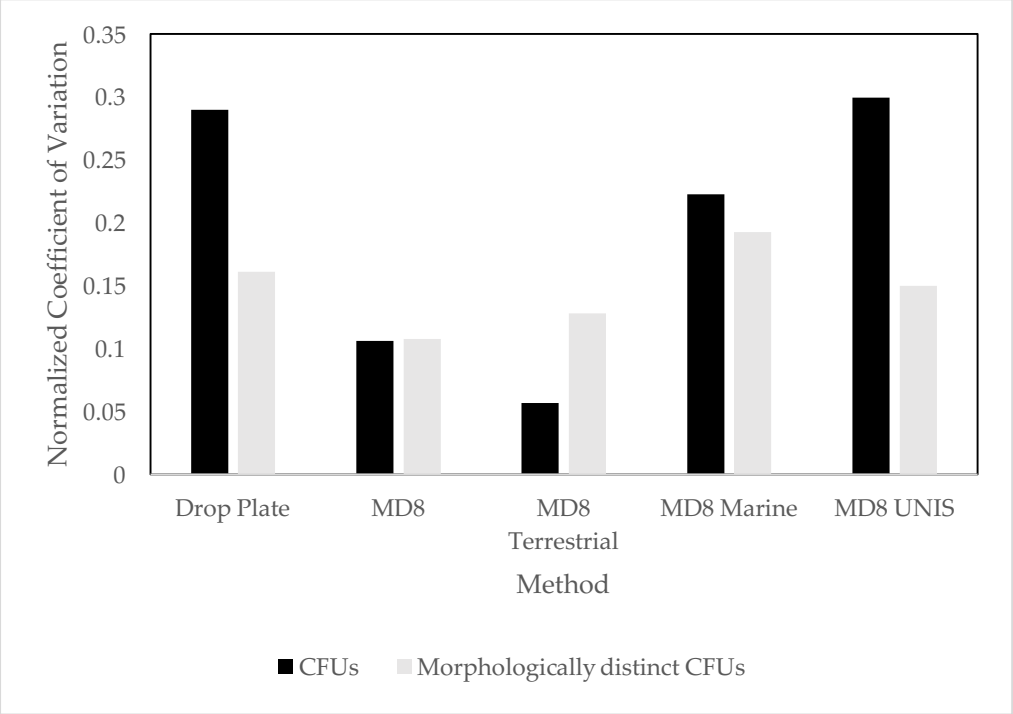


Figure 1.19. Normalised coefficient of variation (a ratio of mean and standard deviation without unit, to compare different scales, here normalised to account for small sample size)

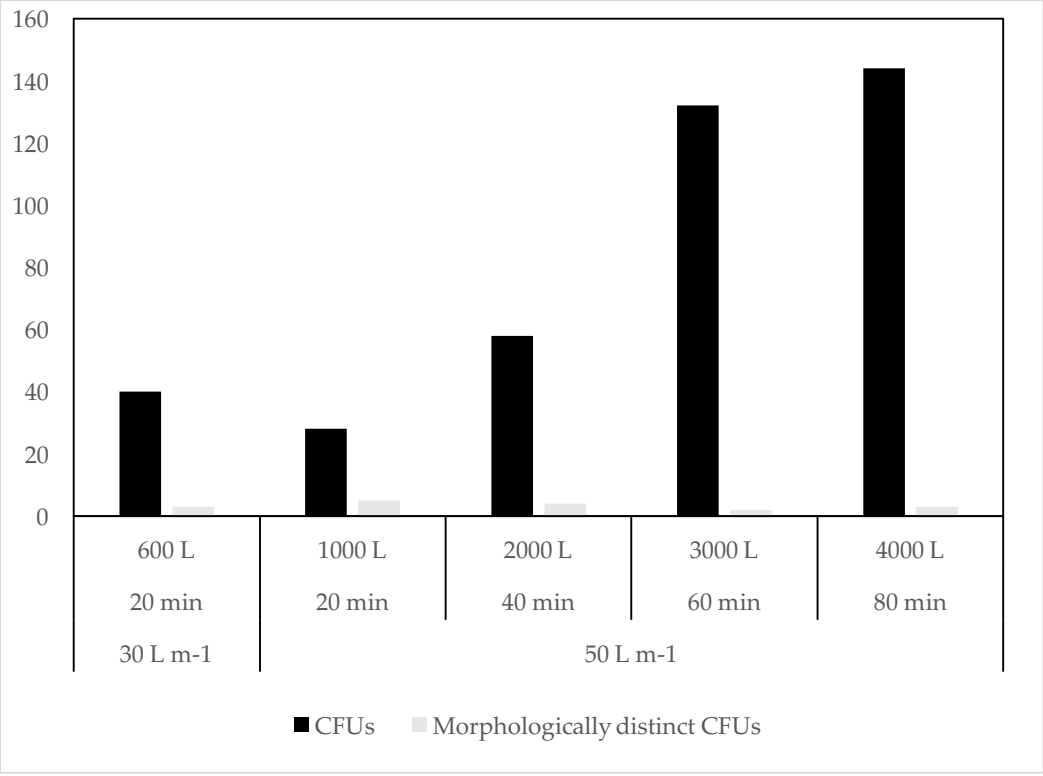
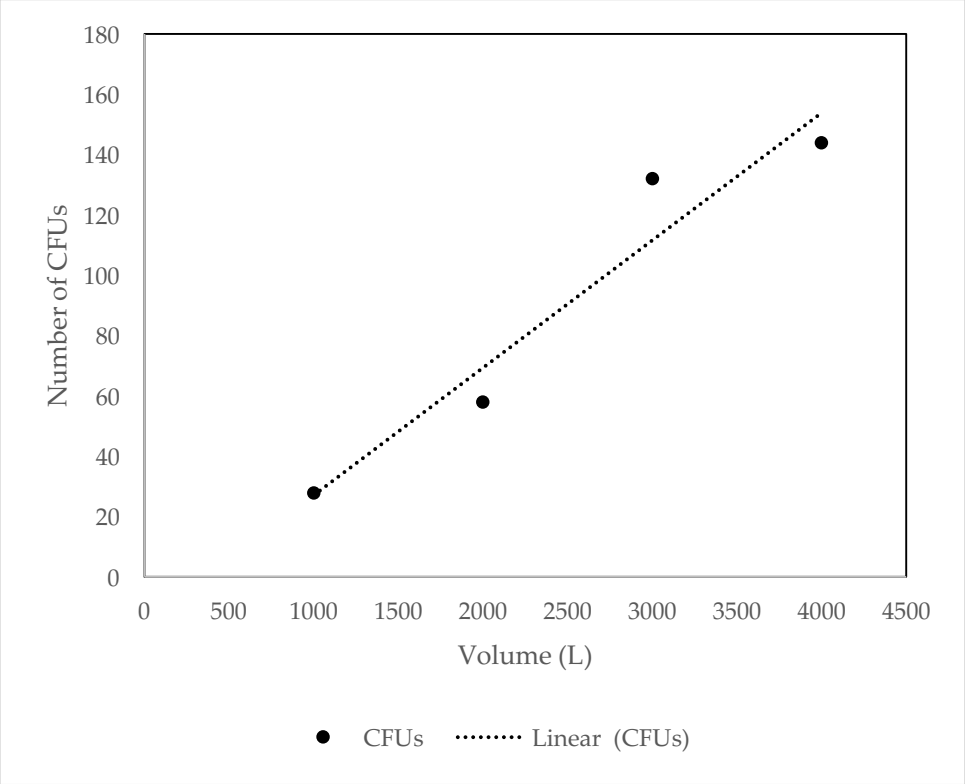


Figure 1.20. MD8 samples with increasing sample volume at UNIS



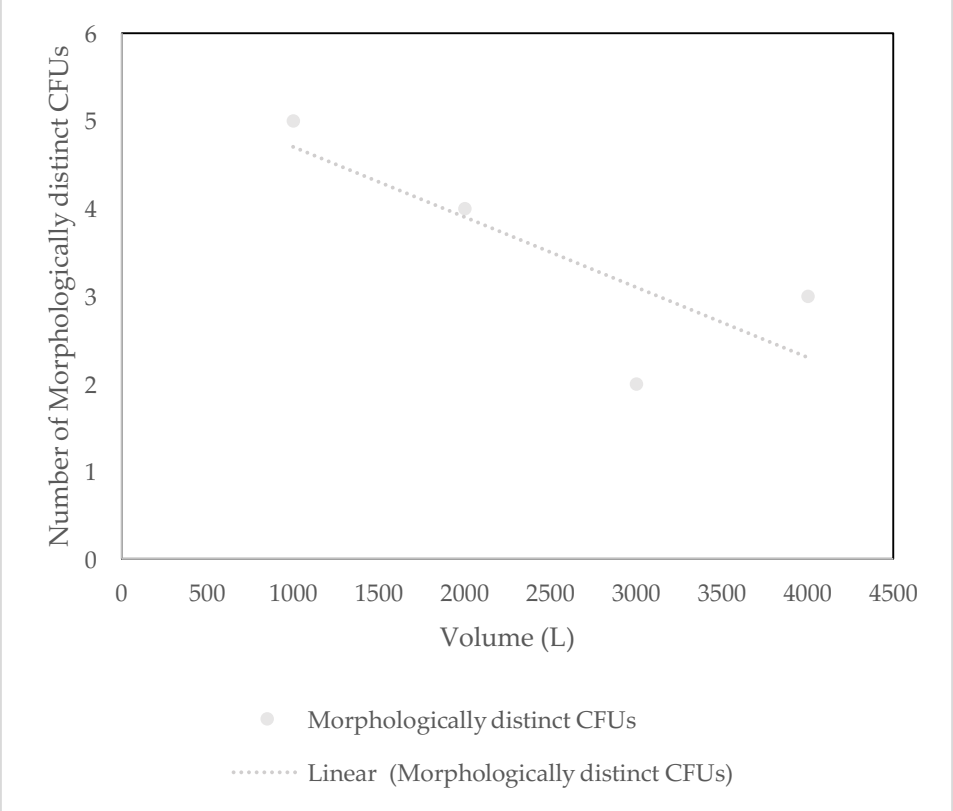


Figure 1.21. (**A**) CFUs sampled against total volume of air (excluding 30 L m^{-1} sample). (**B**) CFUs sampled against total volume of air (all samples)

3.3.2 Culture Independent

3.3.2.1 Bacterial diversity

Targeted amplicon sequencing of the 16S rRNA V4 region resulted in 776,315 total reads across the 10 samples, which were then quality filtered and checked for chimeras leaving 685,583 reads. The range of reads per sample ranged from 32,651 (recorded in the 60 min sample from 6 July) to 145,488 reads (recorded in the 30 min sample collected on 19 July). Samples were then rarefied to 26,190 reads (lower than the smallest sample); rarefied samples averaged 5015 OTUs (range 4143–6402). Rarefaction curves for all of the normalised samples did not reach asymptote suggesting the full extent of the diversity present was not reached for all samples (Figure 1.22A).

The Shannon diversity index, a proxy for richness and evenness, was similar in all samples (Figure 1.22B). The results showed that all samples shared similar levels of diversity (Shannon index range 7.66–9.28); the most diverse sample based on the Shannon index was the 30 min sample taken on Day 2 whilst the least diverse sample based on this metric was the 60 min sample on Day 1. The dominance Simpsons reciprocal index showed a larger difference in the degree of diversity between samples, showing the marine sample to be the most diverse with a Simpsons reciprocal value of 114.13 whilst the lowest diversity was seen again in the Day 1 60 min sample with a value of 20.35 (Figure 1.22C).

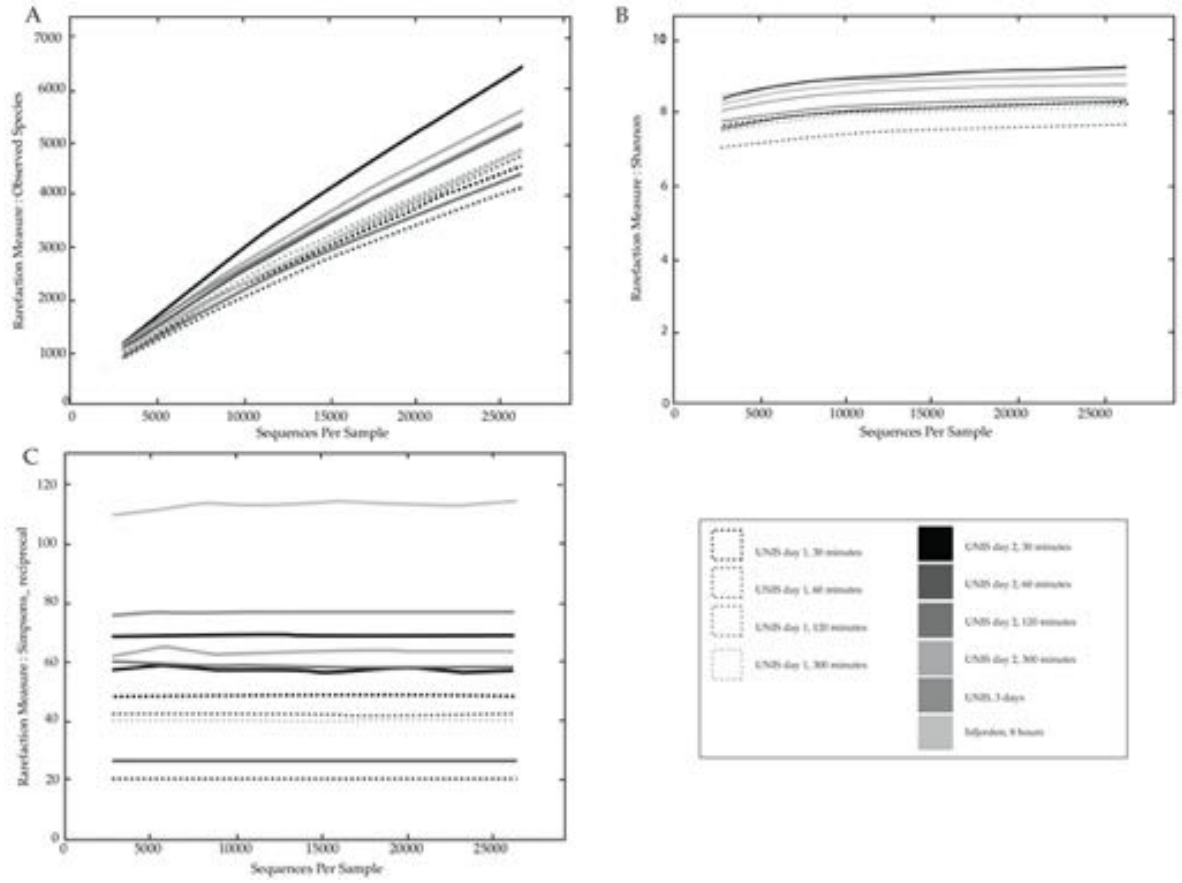


Figure 1.22. α-Diversity measures: (**A**) Rarefaction curves for observed species; (**B**) Shannon index; and (**C**) Simpsons reciprocal index

The differences in OTU diversity between the communities was measured using the Bray–Curtis dissimilarity index (Figure 1.23A) and an un-weighted UniFrac was used to estimate the phylogenetic distance between different communities (Figure 1.23B), the variation across all PCoA axis was low. Both metrics showed no distinct pattern between sampling days; however, sampling location did have an effect and different sampling durations showed minor clustering between the 60 and 120 min durations on Day 1. All samples reported differing levels of richness and evenness (Figure 1.23A) and showed considerable phylogenetic distances with the greatest distance in the three-day sample (Figure 1.23B).

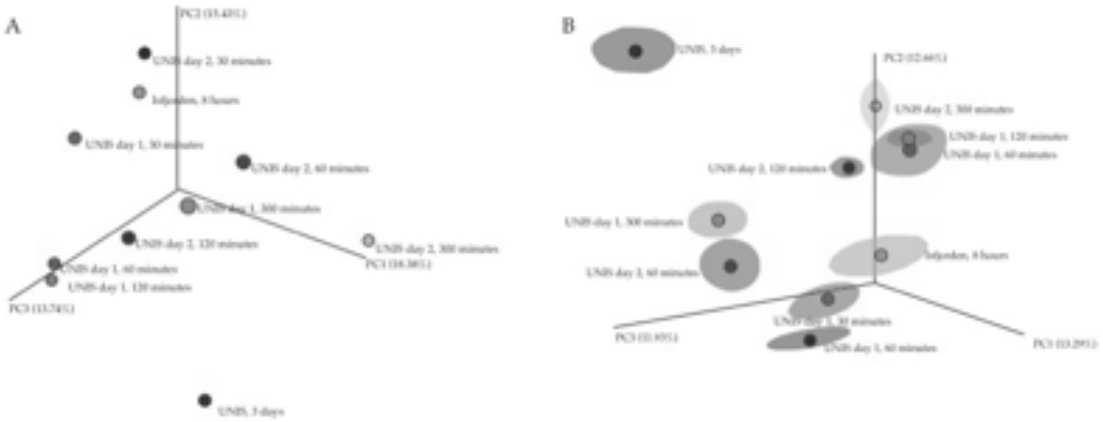


Figure 1.23. Jackknifed β-diversity metrics: (**A**) Bray–Curtis Index; and (**B**) Unweighted UniFrac

3.1.1.1 Taxonomy

Twelve phyla in total were detected within the samples: *Proteobacteria*, *Firmicutes* and *Actinobacteria* were present in all of the samples at differing but high relative abundances and were the visibly dominant phyla (Figure 1.24); *Bacteroidetes*, *Chloroflexi* and *Cyanobacteria* were also present in all samples; and *Cyanobacteria* and *Bacteroidetes* were present in sporadically large relative abundances, however, in general, these three phyla were present at <1% (Figure 1.24).

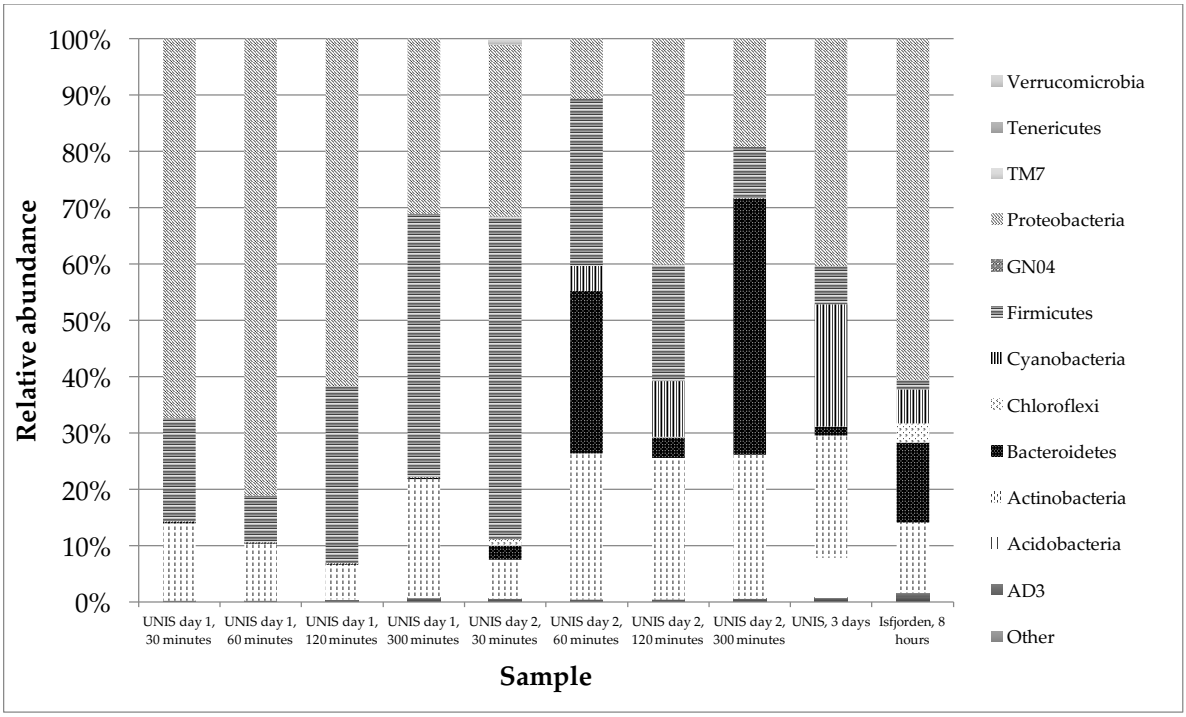


Figure 1.24. Phyla level relative abundances (%) of bacteria in all culture independent samples

Proteobacteria, *Firmicutes* and *Actinobacteria* represented ~99% of the Day 1 sample set in which there were 10 phyla present in total (Figure 1.24). *Proteobacteria* showed the largest total and range of relative abundance on this sampling day. On Day 2, there were 12 phyla present. *Proteobacteria*, *Firmicutes* and *Actinobacteria* remained the three key phyla at a total average relative abundance of 75%. The decrease in relative abundance from Day 1 was mirrored in the 60 and 300 min duration samples by an increase in the average relative abundance of *Bacteroidetes*. The three-day sample contained 10 phyla as for Day 1, but showed similar phyla and relative abundances to the 60 min sample on Day 2. *Acidobacteria* were present at 6% in this sample, but they were present at <1% relative abundance in all other samples. The marine sample collected at Isfjorden contained 10 distinct phyla, the same number present in the Day 1 and three-day sample.

There were 196 genera in total, 58 of which were present in all samples. The marine sample taken at Isfjorden contained the highest number of distinct genera with 148, whilst the

terrestrial 120 min sample on Day 2 contained the lowest number of genera at 100. On Day 1, the average number of genera present was 190 whilst on Day 2 the number dropped to 113. In the three-day sample at UNIS there were 130 genera, more than in any of the other eight samples collected at that location. *Pseudomonas*, *Staphylococcus*, *Propionibacterium*, *Delftia* and *Corynebacterium* spp. made up the five most relatively abundant genera. *Pseudomonas* was the most common and relatively abundant genera representing 18% of the full sample set, however, they were only the most abundant genera in the Day 1, 30 min sample. *Pseudomonas, Acinetobacter, Corynebacterium, Staphylococcus, Deltia, Cloacibacterium, Arthrobacter, Sphingomonas, Alcanivorax, Comamonas, Streptomyces* and *Brevibacterium* spp. were all regularly present in the top 10 most abundant genera in each sample (Table 1.10). Members of the order *Lactobacillales* and *Alcaligenes* were both present in all four Day 1 samples but just one Day 2 samples whilst *Microbacterium,* a genus of the *Microbacteriaceae* family and a member of the *Intrasporangiaceae* family were present in all four Day 2 samples but just one Day 1 sample. There were 15 genera specific to Day 1 and 17 specific to Day 2. The three-day sample recorded six genera specific to that sample. The marine sample recorded the highest number of sample specific genera with 17.

UNIS Day 1, 30 min		**UNIS Day 1, 60 min**		**UNIS Day 1, 120 min**		**UNIS Day 1, 300 min**	
OTU assignment	Relative abundance	OTU assignment	Relative abundance	OTU assignment	Relative abundance	OTU assignment	Relative abundance
Pseudomonas	38%	*Pseudomonadaceae*	39%	*Pseudomonadaceae*	31%	*Corynebacterium*	17%
Acinetobacter	13%	*Corynebacterium*	22%	*Staphylococcus*	22%	Unassigned	16%
Bacillales	10%	*Pseudomonas*	13%	Unassigned	12%	*Acinetobacter*	14%
Corynebacterium	9%	Unassigned	8%	*Streptophyta*	9%	*Gaiellaceae*	14%
Staphylococcus	8%	*Micrococcus*	7%	*Bacillales*	9%	*Pseudomonadaceae*	9%
Pseudomonadaceae	8%	*Bacillales*	3%	*Pseudomonas*	6%	*Pseudomonadaceae*	9%
Delftia	7%	*Gammaproteobacteria*	2%	*Pseudomonadaceae*	3%	*Staphylococcus*	7%
Propionibacterium	5%	*Gaiellaceae*	2%	*Burkholderiales*	3%	*Bacillales*	5%
Unassigned	0%	*Nocardioidaceae*	1%	*Acetobacteraceae*	2%	*Acetobacteraceae*	3%
Cloacibacterium	0%	*Acinetobacter*	0%	*Sphingomonas*	1%	*Alcaligenaceae*	2%
UNIS Day 2, 30 min		**UNIS Day 2, 60 min**		**UNIS Day 2, 120 min**		**UNIS Day 2, 300 min**	
OTU assignment	Relative abundance	OTU assignment	Relative abundance	OTU assignment	Relative abundance	OTU assignment	Relative abundance
Unassigned	49%	*Oxalobacteraceae*	27%	*Oxalobacteraceae*	26%	*Comamonadaceae*	36%
Corynebacterium	18%	*Acinetobacter*	15%	*Staphylococcus*	20%	*Candidate division TM7*	24%
Alcaligenaceae	7%	*Arthrobacter*	14%	*Arthrobacter*	11%	*Alcanivorax*	6%
Staphylococcus	6%	*Corynebacterium*	11%	*Brevibacterium*	10%	*Bacillaceae*	5%
Arthrobacter	2%	*Alcaligenaceae*	11%	*Candidate division TM7*	9%	*Enterococcus*	3%
Comamonadaceae	2%	*Nocardioidaceae*	4%	*iii1-15*	8%	*Gaiellaceae*	3%
Acinetobacter	2%	Unassigned	3%	*Corynebacterium*	4%	*Brevibacterium*	3%
Candidate division TM7	1%	*Streptomyces*	3%	*Comamonadaceae*	3%	*Comamonas*	3%
Gaiellaceae	1%	*Staphylococcus*	2%	*Weeksellaceae*	3%	*Staphylococcus*	2%
Pseudomonas	1%	*Cloacibacterium*	2%	*Comamonas*	1%	*Acidovorax*	2%
Isfjorden, 8 h		**UNIS, 3 Day**					
OTU assignment	Relative abundance	OTU assignment	Relative abundance				
Oxalobacteraceae	39%	*Oxalobacteraceae*	22%				
Bacteroides	10%	*Alcanivorax*	14%				
iii1-15	6%	*Burkholderiales*	7%				
Delftia	5%	*Corynebacterium*	6%				
Burkholderia	4%	*oc28*	6%				
Achromobacter	4%	*Candidatus Aquiluna*	5%				
Caulobacteraceae	3%	*Bacillaceae*	5%				
Comamonadaceae	3%	*Microbacteriaceae*	5%				
Staphylococcaceae	2%	*Streptomyces*	4%				
Burkholderiales	2%	*iii1-15*	4%				

Table 1.10. Top 10 most abundant OTUs in each sample labelled at their highest resolution

3.4 Discussion

3.4.1 Culture dependent

All culture dependent samples recorded growth, showing that viable microbes are common in the atmosphere, at both terrestrial and marine locations, around Svalbard. The number of viable bacteria measured in the air was considerably lower than the number measured in other environments (e.g., surface ice, and cryoconite holes) tested using the same media on Svalbard where the number of CFU can be tenfold higher (185). These results suggest that the atmosphere represents an extremely selective environment, although it is worth noting that only 0.2%–2% of the culturable bacteria in the atmosphere are typically recovered by culture dependent studies (186, 187). Generally, marine studies tend to present more CFUs than terrestrial samples (188). Despite this there were no significant differences between these two environments; however, by normalising coefficients of variations in the sample a clear difference was visible between the two environments. In our case, the highest number of viable bacteria was in the samples taken at UNIS, consistent with the diversity of activity in that location.

The number of cultivable bacteria increased with the increase in sample volume when using the MD8. This contradicts previous studies which showed no effect of sample volume on total CFU counts (189). In addition, decreasing the flow rate from 50 L m^{-1} to 30 L m^{-1} increased the number of cultivable bacteria recovered, possibly due to the decreased impact stress placed on captured bacteria (190).

Whilst culture dependent studies provide useful information about the proportion of viable bacteria in the atmosphere, it is generally considered that only around 1% of the total bacteria present in the atmosphere are culturable (191). Dormancy may represent an important survival mechanism for bacteria in the atmosphere; therefore, a considerably larger proportion of viable non-culturable bacteria (VBNC) would also be expected and may have been overlooked in previous studies based on culture techniques alone. The reliance on CFU counts and inability

to describe VBNC bacteria limits the value of culture dependent techniques from an ecological perspective.

3.4.2 Culture independent

Culture independent studies using sequencing can provide more information about the diversity and taxonomic composition within an environment. Despite the ability of culture independent studies to generate useful information, they also have major drawbacks, as they contain little information about the viability of the bacteria in the environment. Thus, combining both culture dependent and culture independent methods, provides a better insight into both the structure and viability of bacterial communities.

Previous research on bacteria in the atmosphere outside the Arctic has linked temporal and spatial variation to changes in the diversity and abundance (26). Despite these factors impacting bacterial communities in other Arctic ecosystems such as soil (27), there are no studies to date which investigate these patterns in the atmosphere in this region. Temporal variation (sampling day) did appear to have an effect on community structure, as the composition of the dominant Day 1 phyla was clearly different to that on the other three sampling days. Spatial variation (marine and terrestrial) also appeared to have an effect, although this was less pronounced than the temporal variation, as the dominant phyla present in both the marine and terrestrial samples was consistent. Our results suggest that day of sampling (temporal) is more important than location (spatial) with regards to sample diversity most likely due to changes in meteorological conditions such as wind direction which appeared to produce distinct communities at the phylum level (Figure 1.24).

Duration also appeared to have an effect on the taxonomy of the communities, because whilst the dominant groups of phyla remained constant, the relative abundances varied considerably with changing duration. Although this variation could relate to confounding factors such as the time of day the samples were taken and the duration of sampling. The phylum level patterns

seen in the 3-day sample collected were similar to those seen in a single 3-day sample collected in 2017, suggesting a stable phylum-level community (see appendix I).

3.4.3 Diversity

Samples did not cluster into distinct groups based on OTU or phylogenetic relationships, showing no direct link between diversity and sampling duration, location or day (Figure 1.23A,B). On Day 1, 60 and 120 min samples clustered based on both relationships, likely due to the samples sharing similar relative abundances of *Delftia*, *Ralstonia* and *Pseudomonas*. A higher Simpson reciprocal value was seen on the third sampling day (76.61) taken at UNIS, suggesting that sampling for a longer duration increases the diversity of bacteria captured. The marine sample was considerably more diverse than the terrestrial samples when taking into account dominance (Figure 1.22C), further supporting the idea that the distinct geographical features of marine coastal locations when compared to terrestrial ones give rise to more varied communities (133, 160). Meteorological conditions such as wind speed, humidity and pressure are known to directly impact community structure (112); however, during our study, these conditions remained relatively constant, which could explain the similar levels of diversity of the samples shown by the Shannon index (Figure 1.22B).

3.4.4 Taxonomy

A maximum of 12 phyla were found in air samples from Svalbard; however, the number of phyla varied among samples. The pattern found on Day 1 was the most distinct with three phyla dominating the day. The distinctiveness of the pattern on Day 1 was likely due to easterly winds from a low altitude air mass leading up to and during this sampling occasion. During the other sampling days, the predominant wind had a main westerly component. The 12 phyla could be separated into two groups: the primary phyla *Proteobacteria*, *Firmicutes* and *Actinobacteria*; and the remaining phyla that were present in sporadic relative abundances. This pattern is consistent with previous studies in cold ecosystems (24, 131, 144, 145), and of bioaerosols in a range of environments (139, 153, 161, 192); *Bacteroidetes* could be considered a primary

phyla, as they were present in considerable relative abundance in all of the samples apart from on Day 1, suggesting their source is to the east of Svalbard due to the back trajectory of the prevailing wind direction. The primary phyla are probably well adapted to atmospheric life, e.g., *Firmicutes* are well known for their ability to form spores in low nutrient conditions (193). *Actinobacteria* have a higher GC content than other bacteria (194), which is a useful defence against the increased UV exposure faced by bioaerosols, and *Proteobacteria* are known to fill a multitude of niches due to the metabolic diversity of the group (195).

The number of phyla occurring in the air above Svalbard is considerably lower than that described for urban environments, with studies reporting the number of distinct phyla present to be as high as 38 (192), likely due to differences in the environments. It is notable that *Deinococcus* were not present in any of the samples, a group of bacteria normally associated with atmospheric studies, both in the Arctic and elsewhere (26, 143). *Bacilli* sp. were responsible for a large proportion of the *Firmicutes* present in the sample, the source of which in the terrestrial samples was likely the surrounding soil (25). There also appeared to be a relationship between the *Actinobacteria* and the *Pseudomonadales* whereby as the relative abundance of one increased, the other decreased as has been found previously (196). Interestingly there was a spike of *Acidobacteria* in the three-day sample which could suggest this phylum is best adapted to survive the threat of desiccation caused by sampling for longer periods.

At the genus level, the patterns were much less distinct. Of the 196 genera, only 58 were present in all samples. The five most relatively abundant genera (*Pseudomonas, Staphylococcus, Propionibacterium, Delftia* and *Corynebacterium* spp.) are all either polar associated or ubiquitous. *Delftia* spp. have been described at multiple Arctic locations including Svalbard and Greenland where they are associated with surface ice (197, 198) whilst *Propionibacterium* spp. are typically associated with marine sediment in the Arctic Ocean (199, 200).

Pseudomonas spp. are ubiquitous and present in almost all polar studies, however, on Svalbard they are mainly described in fjords (201), indeed, a new psychrophilic species of *Pseudomonas* was recently described from the same region (202). *Corynebacterium* spp. have previously been found in soils from the Canadian high Arctic (203). *Staphylococcus* sp. were frequently present, but are not routinely described in environmental Arctic studies and could be human or animal associated. *Acinetobacter* spp. are also commonly found in the top 10 most relatively abundant bacteria in all the locations. *Acinetobacter* spp. have been found in glacial snow and ice in mountainous locations outside the Artic (204), however, are mainly associated with marine environments such as fjords in Svalbard (201). *Alcanivorax* spp. and members of the *Oxalobacteraceae* family were also common, they appeared on the days dominated by easterly winds and did not appear on the day dominated by westerly winds (205, 206). Members of the *Oxalobacteraceae* family have also been described in Arctic soils (207).

Polaribacter sp., a bacterium associated with polar sea ice, was present in the marine sample suggesting that the Arctic Ocean provides a source of bacteria to the atmosphere. Many of the regularly occurring marine psychrotrophs, included in the *Pseudomonas, Acinetobacter, Alcanivorax, Psychorbacter* genera and members of the *Oxalobacteraceae* family are associated with the degradation of hydrocarbons in the Arctic (208), which are abundant in Svalbard fjords. The number of distinct phyla recovered on Svalbard (12) was higher than the number recovered over Ward Hunt Island (WHI) in the Canadian high Arctic (143) where six distinct phyla were found. Several of the 14 genera described in the air on Ward Hunt Island (WHI) were also present on Svalbard, including *Cytophagales, Lactobacillus, Staphylococcus, Janthinobacterium, Pseudomonas* and *Polaromonas,* which were mentioned but excluded as a chimeric sequence in that study. Bipolar comparisons also give an insight into both long-range transport and biogeography. Thus, Pearce, Hughes (153) described the presence of

Acidovorax, Acinetobacter, Cloacibacterium, Pseudomonas and *Sphingomonas* at Halley station in Antarctica, all of which were present at varying relative abundances in Svalbard air.

3.5 Concluding remarks

Abundant viable bacteria from a reduced range of bacterial phyla were found in the air above Svalbard, therefore the hypothesis that bacterial communities are ubiquitous in the atmosphere around Svalbard can be accepted. The communities described were fairly homogeneous across sites, suggesting a distinct aerial community above Svalbard, thus the statement bacterial communities in the air above Svalbard are homogeneous can be partially accepted. Airborne bacterial abundance was lower than that described from other Arctic environments, such as soil or the ice surface. The most relatively abundant taxa were polar associated, suggesting that the largest input into the atmosphere on Svalbard was of local origin. The overall diversity of the phyla present in the air above Svalbard was less diverse than in other locations such as urban environments, but was similar to that described previously in the Arctic on WHI. The key phyla remained consistent across studies. Bacterial community biodiversity was impacted by sampling regime, therefore the hypothesis that sampling methodology does not impact the seen biodiversity of Arctic bioaerosol communities must be rejected.

Further studies using metatranscriptomics would provide a deeper insight into the ecological role and metabolic activity of airborne bacteria, and potentially their ability to sustain activity, colonize and alter the environment at their final destination. Future studies investigating the biodiversity of the airborne microbes present in the Arctic will provide an insight as to whether an indigenous community is truly present.

Chapter 4 - Microbial biodiversity of the air around Antarctica

4.1 Introduction

Knowledge of the structure and function of microbial communities on continental Antarctica has rapidly increased in recent decades, with studies focusing on environments such as subglacial lakes (209), dry valleys (152, 210), and surface snow (211). Despite this, the understanding of bacterial dispersal into the continent, and of endemic microbial communities is limited, due in part to a low number of aerobiological studies in the region (151). The impact of changing environmental conditions on the biodiversity of Antarctic bacterial communities is not well published, despite the relevance of this information with regards to climate change; although some studies have found significant relationships between environmental variables, such as air temperature, and the diversity of bacterial communities (212).

Despite the remoteness of the Antarctic continent, several potential dispersal pathways for microbial transport exist. The surrounding Oceans provide the most obvious long distance route for bacteria into the continent, but there is little evidence of marine associated bacterial taxa in the Antarctic atmosphere despite the relative vicinity of the sampling sites to the surrounding seas (152, 153). The Antarctic circumpolar current (ACC), where the cool water of the Antarctic sea meets the warmer waters of the Pacific, Atlantic, and Indian oceans is thought to form a barrier for marine dispersal into the region. The westerly winds, located between -30°S and -60°S, move aerosolised material towards the Antarctic continent from Africa and South America. The ability for these air masses to facilitate the intercontinental transport of microbes has previously been suggested, with evidence showing the presence of bacteria endemic to South America in ancient Antarctic ice sheets (213). The Antarctic circumpolar vortex, a bi-product of the ACC, due to the mixing of cold and warm air masses

at -60°S, is thought to act as a physical barrier to incoming aerosolised particles, limiting dispersal to the region (214).

There are some 10^{30} bacteria and archaea residing in oceans across the globe (215). Microorganisms use bubble bursting as a transfer mechanisms between marine and aerial environments; it has been suggests that microbes possess the ability to manipulate the duration of a bubble to enhance their own dispersal (216), therefore some degree of relationship between marine and aerial bacterial communities is likely. Global studies of marine bacteria have found a markedly similar biodiversity worldwide, describing ocean prokaryotic communities to be dominated by *Gammaproteobacteria*, *Alphaproteobacteria*, *Actinobacteria*, *Thaumarchaeota* and *Deltaproteobacteria*, with *Gammaproteobacteria* being the most relatively abundant of the core phyla; *Actinobacteria* were the only phyla to differ significantly in relative abundance between oceans, being more abundant in the Pacific Ocean (217). One study of the Southern Atlantic Ocean found a total of 30 phyla; samples were dominated by *Proteobacteria*, mainly of the *Gammaproteobacteria* class representing close to 50% of all samples collected (218).

The landing of aerosolised dust, and the aerosolisation of dust by air movement means there is an intimate relationship between local microbiomes and airborne bacteria. Antarctic microbiomes are now known to be diverse. Perhaps the most commonly studied terrestrial ecosystems in the cryosphere are Antarctic soils. Antarctic soil communities are relatively heterogeneous with the *Actinobacteria*, *Proteobacteria*, *Bacteroidetes*, *Acidobacteria*, *Gemmatimonadetes*, *Deinococcus-Thermus*, and *Cyanobacteria* phyla observed frequently, but at varying site dependent relative abundances (219). Studies of soils in maritime regions of the continent have found *Acidobacteria*, *Bacteroidetes*, *Gemmatimonadetes*, *Proteobacteria*, *Actinobacteria*, *candidate division AD3*, *Chloroflexi*, *Firmicutes* and *Cyanobacteria* to be the dominant phyla, and also found that the alpha diversity of the soil communities was positively related to air temperature (212).

Antarctic surface snow communities are dominated by *Proteobacteria*, mainly of the alpha, beta, and *gamma* classes; Fusobacteria, *Firmicutes*, Armatimonadetes, and *Actinobacteria* were also dominant phyla in continental Antarctic samples, whilst *Bacteroidetes* were present in sub-Antarctic Island samples (211). Within the atmosphere, *Proteobacteria*, *Bacteroidetes*, *Firmicutes*, and *Actinobacteria* have been shown to be the dominant phyla in Antarctic air over the McMurdo dry valley (214). *Proteobacteria* were mostly represented by the *Alpha-*, *Beta-*, and *Gammaproteobacteria* classes whilst *Firmicutes* were mainly *Bacilli* and *Clostridia* (152).

The aim of this study was to investigate the spatial variability of airborne bacterial communities within the atmosphere above the oceans surrounding continental Antarctica, in order to assess whether the following hypotheses were acceptable:

i) Bacteria are ubiquitous in the atmosphere surrounding the Antarctic

ii) Aerosolised bacterial assemblages in the air surrounding the Antarctic are homogeneous, due to the extreme selectivity and remoteness of the environment

iii) Local sources contribute considerably to the composition of bioaerosol samples around the Antarctic

iv) Marine bioaerosol communities differ from terrestrial communities compositionally

v) Airborne bacterial communities present above or below -60°S will harbour distinct patterns of biodiversity

4.2 Methodologies

4.2.1 Sample and metadata collection

Air samples were collected across the 3 oceans surrounding the Antarctic continent, aboard the R/V Akademik Tryoshnikov over an 85-day period between December 22nd 2016 and March 16th 2017, whilst the ship circumnavigated the Antarctic continent. Leg 1 departed Cape Town, South Africa where the ship sailed through the Indian Ocean towards Hobart, Australia; during leg 2, the ship navigated the Pacific Ocean, travelling from Hobart, to Punta Arenas, Chile; finally, during leg 3, the ship travelled from Punta Arenas, back to Cape Town via the Atlantic Ocean. During the course of the circumnavigation, the ship travelled both inside the region contained by the Antarctic Circumpolar Current (ACC) and outside; air samples were also collected at the sub-Antarctic islands of Kerguelen, Marion, Crozet, Heard, Bouvet, Siple and South Georgia (Figure 1.25).

A total of 75 successfully amplified and sequenced samples, comprised of 71 marine and 4 terrestrial, were collected via a membrane filtration apparatus set up as described in chapter 2.1 and in earlier studies (220). Samples 12 (Kerguelen), 56 (Siple), 73, and 47 (both South Georgia) were all collected on land. 4 rainwater samples were also collected, on days 12 (r01), 15 (r02), 17 (r03), and 27 (r04) using a sterilised filter funnel as described in chapter 2.1.

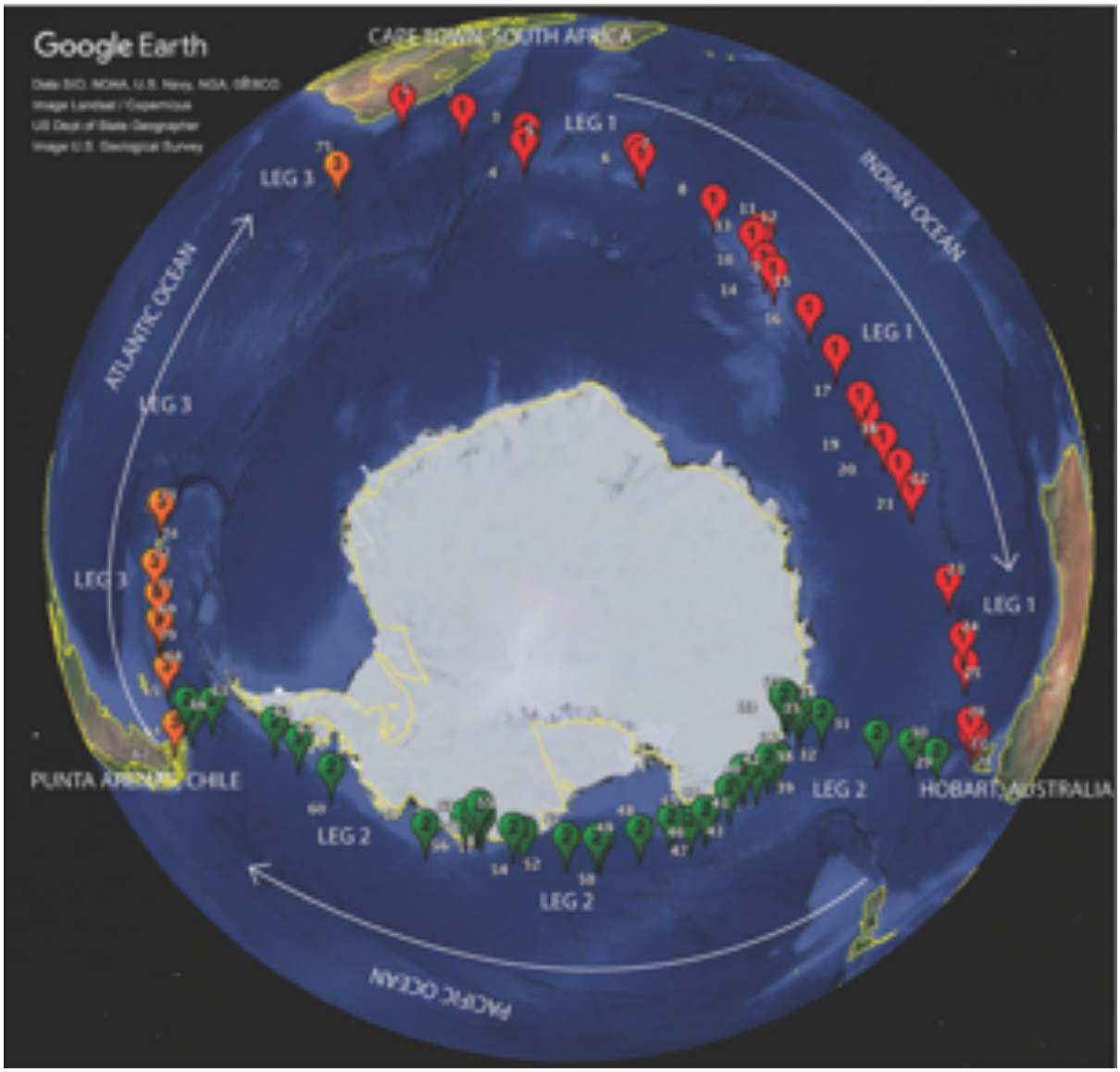


Figure 1.25. Antarctic circumnavigation expedition sampling regime. Red marker = Leg 1, Indian Ocean; Green marker = Leg 2, Pacific Ocean; Orange marker = Leg 3, Atlantic Ocean. Image courtesy of Google Earth.

Marine samples were collected aboard the Akademik Tryoshnikov, the sampling unit was set up on top of the monkey island to reduce the influence of sea spray and potential human bacterial sources; similarly, for terrestrial samples the filtration unit was positioned at a height of 1.5m to reduce the impact of local turbulence. The target sampling duration was 12 hours, however sampling duration range was 108 varying between 1 and 109 hours, dependent on

circumstance. GPS co-ordinates and weather data (table 1.11) was collected continuously throughout the voyage by the Vaisala weather station aboard the ship. Once collected, samples were stored in the ships on board -80°C freezer. Samples were removed from the ship at Bremerhaven, Germany and placed directly into a freezer van for transport back to the British Antarctic Survey, UK. Samples were then transported at -20°C to Northumbria University, UK where they were stored at -20°C until processing.

Variable	Average	Unit
Latitude	Median	Decimal degrees
Longitude	Median	Decimal degrees
Average wind direction	Mode	Degrees
Average wind speed	Mean	Ms^{-1}
Minimum wind speed	Mean	Ms^{-1}
Maximum wind speed	Mean	Ms^{-1}
Cloud level	Mean	Metre
Sky coverage	Mean	Octants
Relative humidity	Mean	%
Temperature	Mean	Degrees C
Dew point	Mean	Degrees C
Pressure	Mean	Millibars
Solar radiance	Mean	NA
UV	Mean	NA

Table 1.11. Weather variables collected by onboard weather station

4.2.2 DNA extraction

DNA extraction was performed on samples using the Qiagen PowerSoil kit (Qiagen, Hilden, Germany) as described within chapter 2.2.1.

4.2.3 Targeted amplicon sequencing

PCR amplification of the V4 region was carried out as described within the methods chapter, the amplified product was then sequenced on an Illumina MiSeq as described in chapter 2.2.8 with the following alterations. PCR was instead performed at 40 in order to increase amplification in low biomass samples; the DNA concentration of each PCR product was individually quantified using the quant-it Picogreen dsDNA assay kit (Thermo Fisher Scientific, MA, USA) and this quantification was used to normalise the library by dilution using a Hamilton robotics microlab star (Birmingham, UK). Finally, samples were cleaned up individually using Ampure XP beads to remove primer dimer and non-target length amplicons prior to pooling.

4.2.4 Sequence processing and analysis

Fastq files generated by 16S Illumina MiSeq were processed into an OTU and taxonomy table in QIIME2, then screened for contaminants using Microsoft Excel (2013) as described in chapter 2.3.1. The OTU table, taxonomy table and metadata files were then read into R studio (R_Core_Team, 2014) and converted into a Phyloseq object (171) for statistical analyses. Sequences were agglomerated at the taxonomic rank of class due to a lack of comparable studies within the region of sampling. Sequences with no class level taxonomy were disregarded from the analysis along with samples with incomplete metadata and samples with fewer than 1000 total reads. Diversity metrics, differential abundance testing, core microbiome, and statistical analyses were carried out as described in chapter 2.3. For analyses, samples were stratified by both leg of expedition and whether they were collected after crossing the Antarctic circumpolar current -60°S as far as -74°S (post-ACC) or before crossing the ACC from -38°S to -60°S (pre-ACC).

4.3 Results

4.3.1 Sampling depth

A total of 1,068,464 reads were retained across the 75 samples. The range of reads was 95,719, with the smallest and largest samples containing 1278 and 96,997 reads respectively. The mean number of reads was 14,246 with a standard deviation of 19,619. Sufficient sampling depth was achieved for all amplicon libraries (Figure 1.26) as shown by each curve reaching asymptote when samples were rarefied to 1000 reads (lower than the smallest sample).

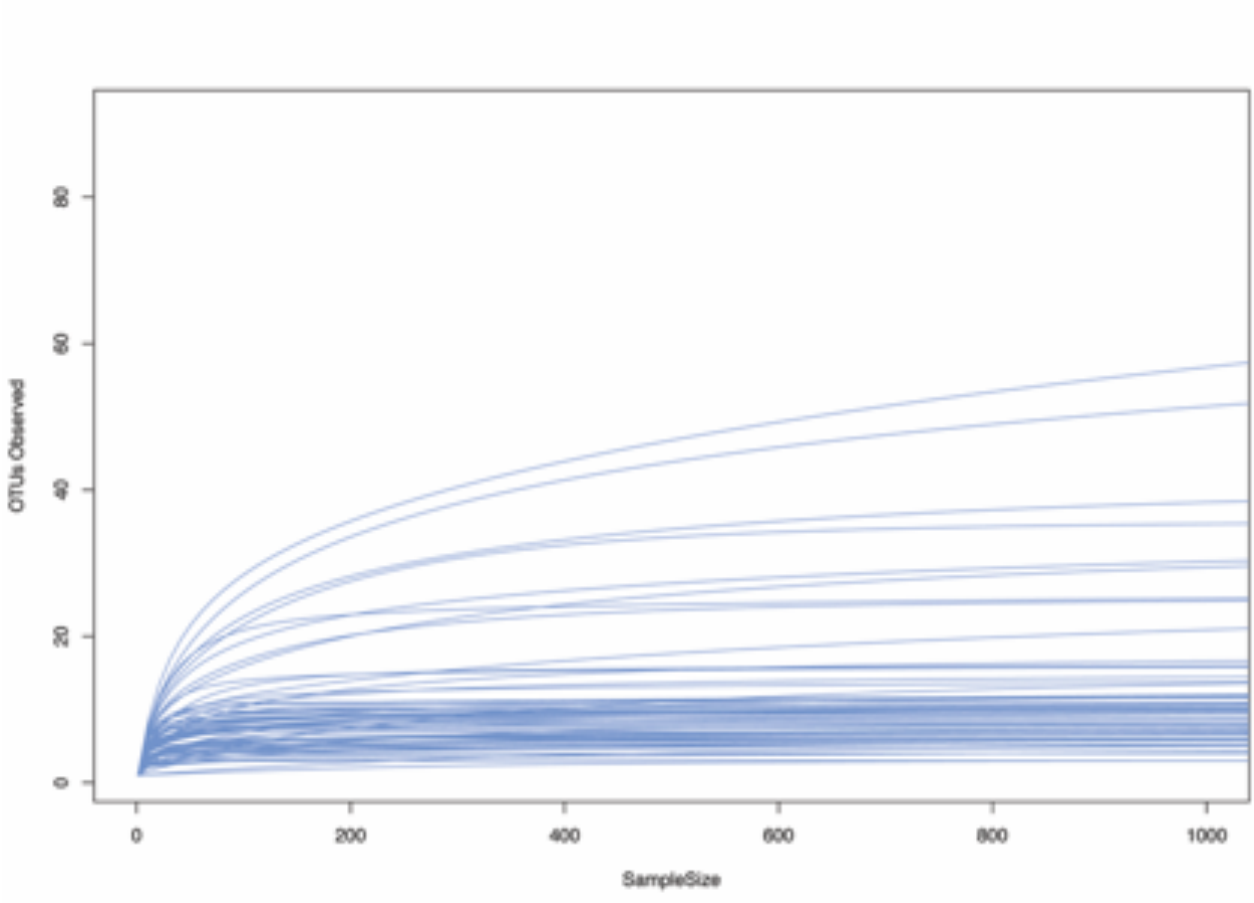


Figure 1.26. Rarefaction curves showing sufficient sampling depth for class level at 1000 reads

4.3.2 Alpha diversity

Class level comparison of the observed OTUs alpha diversity metric by Pairwise Wilcoxon Rank Sum Test with Bonferroni correction showed no significant difference (p=1) between each expedition leg, with the median values for legs 1, 2, and 3 being 9, 8, 8 respectively (figure 1.27). Furthermore, there was no significant pairwise difference between each expedition leg for the Shannon index values (p=1), for which the respective median values were 1.48, 1.34, and 1.43 (figure 1.27). Temperature was significantly lower during the course of leg 2, than during leg 1 or 3 (figure 1.28). Univariate linear regression analysis for all test variables revealed no positive relationship between any of the test variables and either observed OTUs or Shannon Index. The strongest positive relationship was between temperature and Observed OTUs however this relationship was not significant (p=0.09) and the model only described 3% of the variance in the dataset (figure 1.29).

4.3.3 Beta diversity

Principle coordinate analysis of the weighted Bray-Curtis dissimilarity of the communities showed no clear distinction between clusters of each expedition leg (figure 1.30). There was no significant difference between the weighted Bray-Curtis dissimilarity based on expedition leg (pairwise PERMANOVA P value = 1). Terrestrial and marine samples did not cluster independently (PERMANOVA P value = 0.90). Marine samples collected upon approach or departure from corresponding terrestrial sites did not cluster closely to their corresponding terrestrial sample.

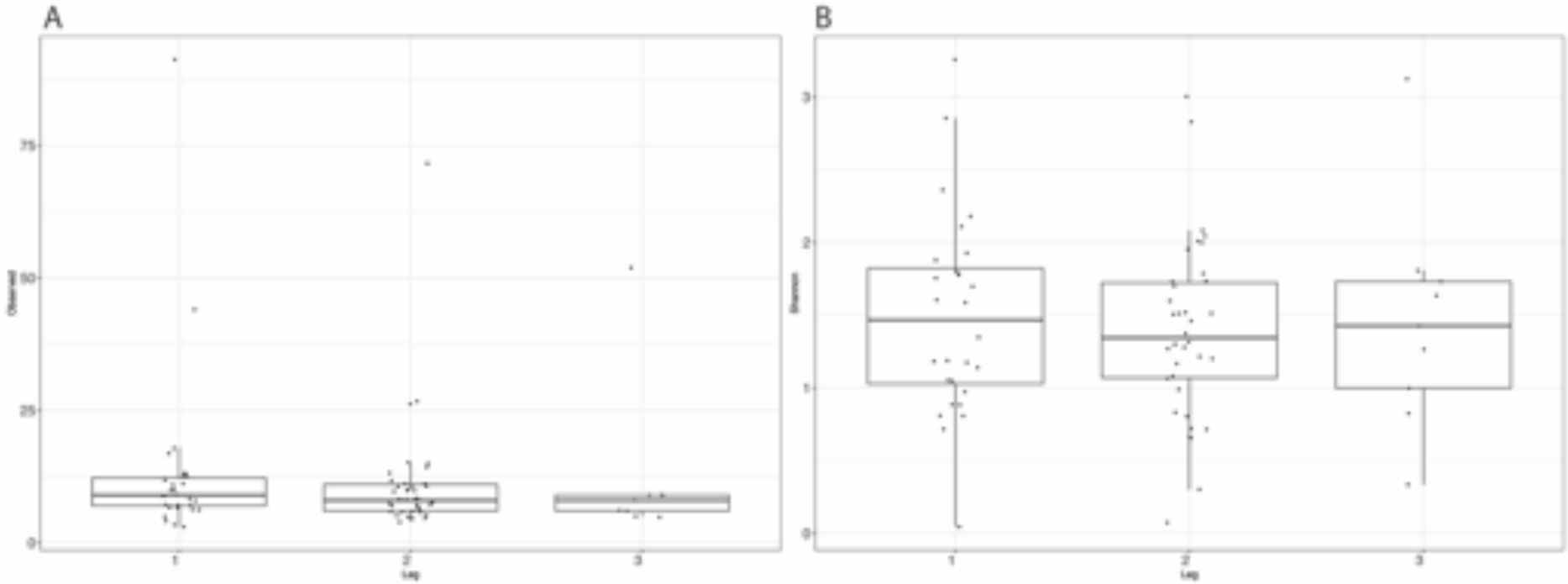


Figure 1.27. Boxplots showing alpha diversity metrics. Samples grouped by expedition leg (1-3). A) Observed OTUs for each leg. B) Shannon Index for each leg

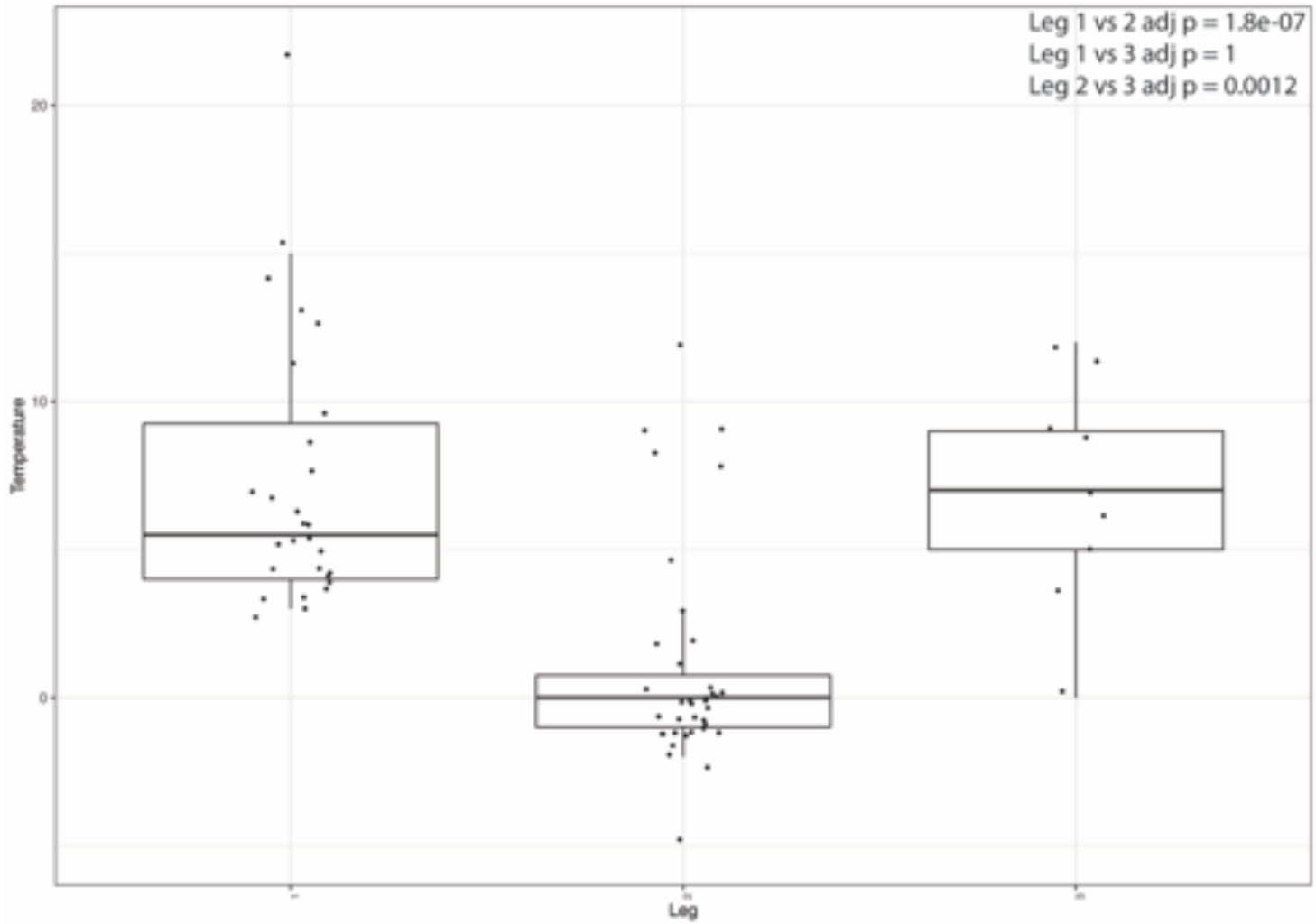


Figure 1.28. Boxplot showing average temperature during sampling for all expedition legs

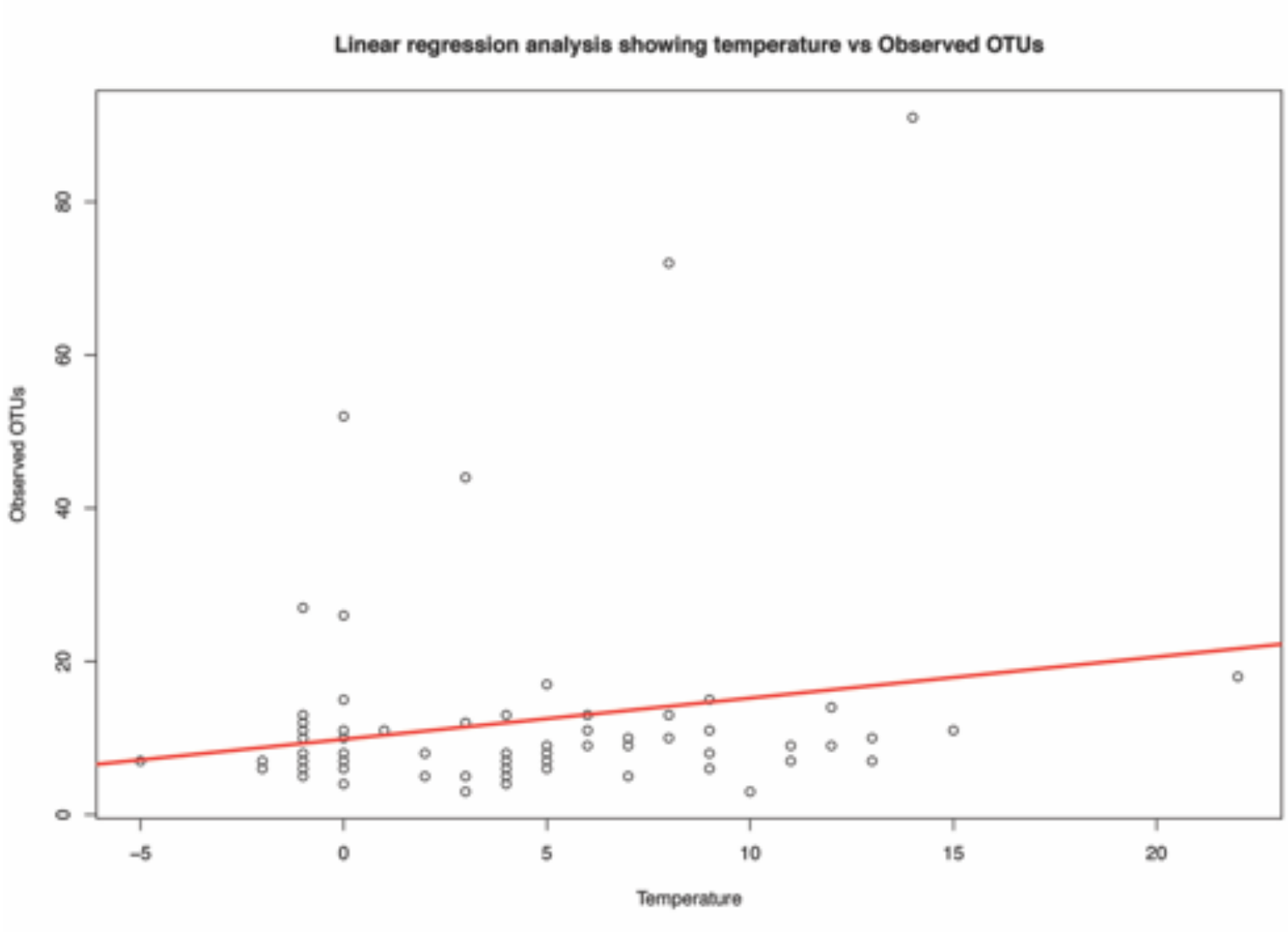


Figure 1.29. Univariate linear regression analysis of number of Observed OTUs against temperature

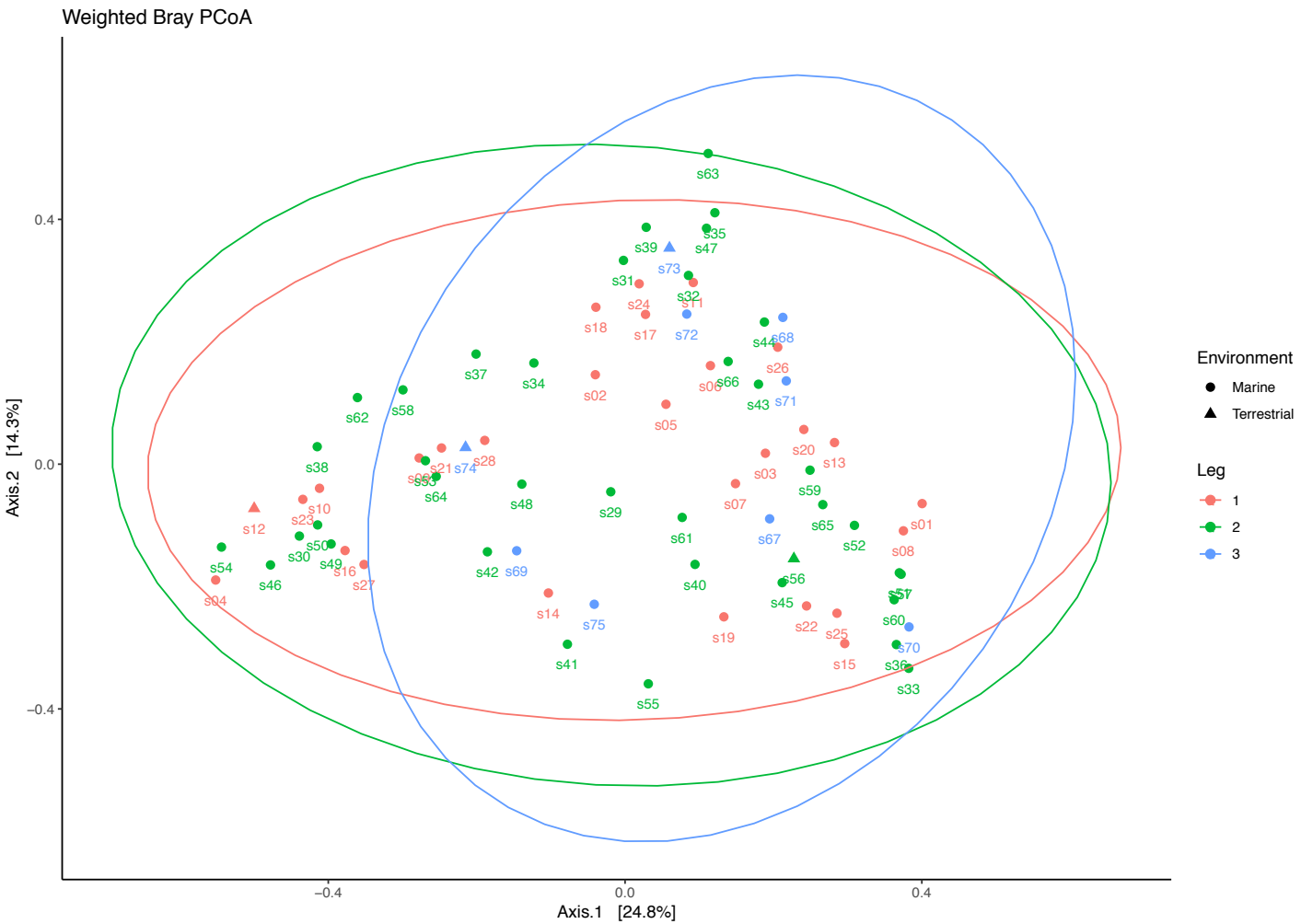


Figure 1.30. Principle Coordinate Analysis displaying the Bray-Curtis dissimilarity between samples. Samples are coloured by expedition leg and shaped by sampling environment. Leg 1 (red), Leg 2 (green), and Leg 3 (blue) are shown alongside 95% confidence ellipse

4.3.4 Taxonomy

45 phyla were present in total. Of the 10 most abundant phyla, *Proteobacteria* appeared most frequently throughout the 3 expedition legs (figure 1.31), however their relative abundance varied considerably. *Firmicutes*, *Actinobacteria* and Bacteroides then followed as the most prevalent phyla. The remaining most abundant phyla appeared more sporadically varying relative abundances. There were a total of 114 classes represented across all samples. *Gammaproteobacteria* were the dominant member of the *Proteobacteria* phylum on the majority of sampling occasions (figure 1.32). *Clostridia* are the dominant class within the *Firmicute* phylum being present on over half of the sampling occasions. The *Actinobacteria* class was also present frequently and on all legs of the expedition, with relative abundances frequently varying from below 10% up to above 95% of the sample.

Of the top 50 bacterial classes present across all sampling days, none were present on all sampling days. Gamm*proteobacteria* were present the most frequently and at the highest average abundances as shown in figure 1.33, followed by *Bacilli*, *Clostridia*, and *Actinobacteria*. The remaining classes which made up the top 50 most abundant group appeared and disappeared, with no clear visible pattern across the longitudinal transects.

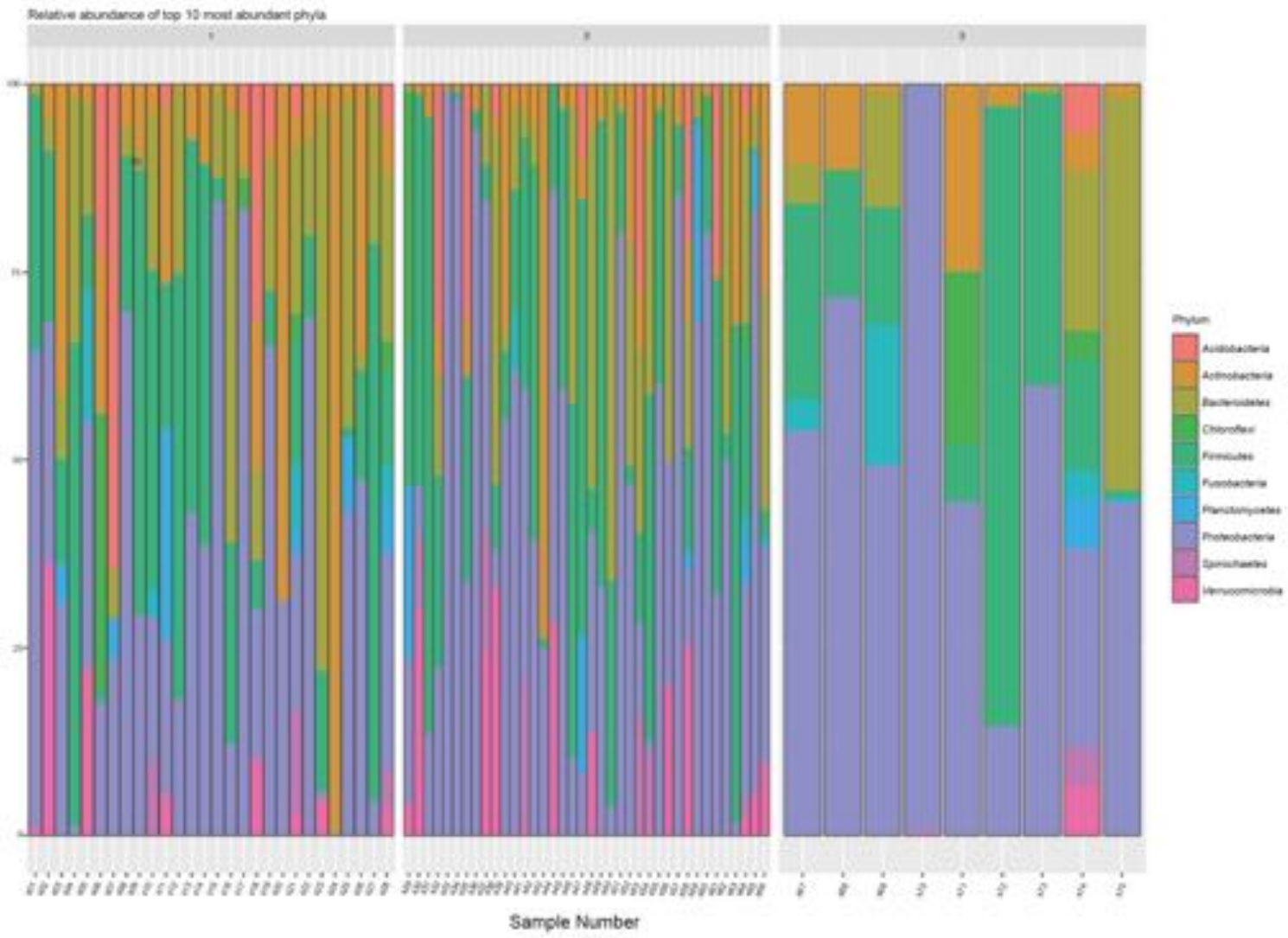


Figure 1.31. Stacked bar showing the relative abundance of the top 10 most abundant phyla

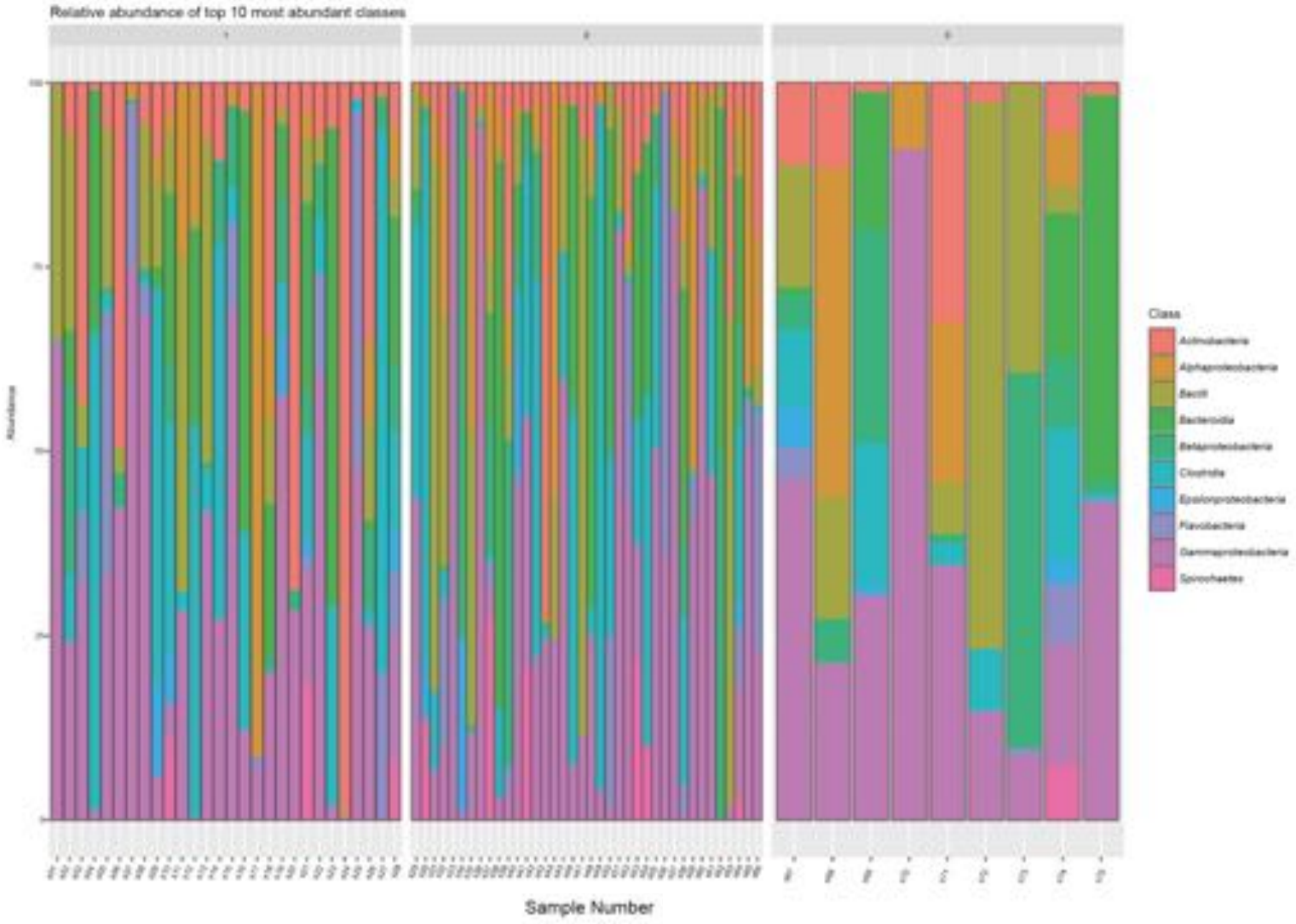


Figure 1.32. Stacked bar showing the relative abundance of the top 10 most abundant classes

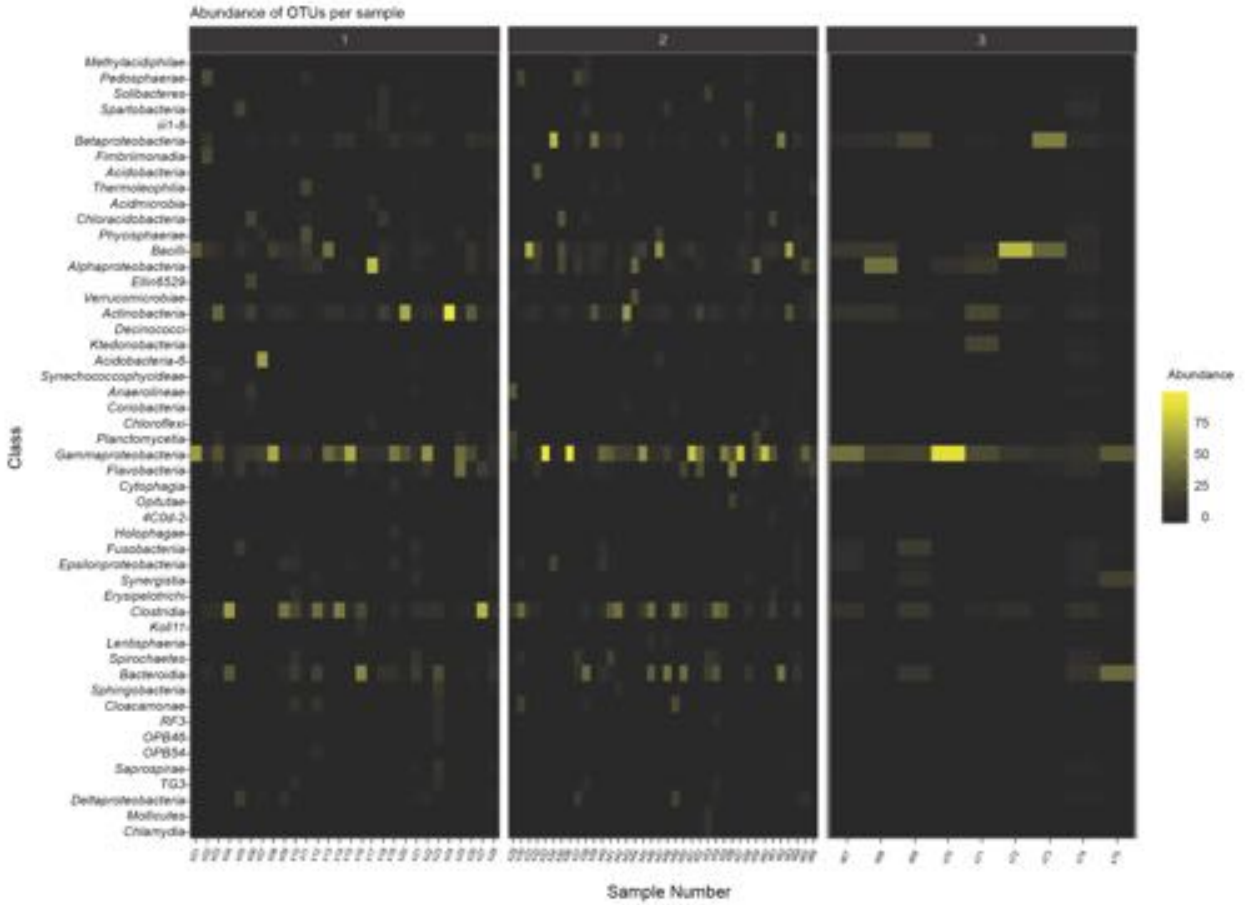


Figure 1.33. Heat map showing the relative abundance of the top 50 most abundant bacterial classes in longitudinal order, faceted by leg of expedition

4.3.5 Differential abundance

There were no differentially abundant classes identified between leg 1 and leg 2 of the expedition. Five classes were identified as being differentially abundant (p=<0.05) between leg 1 and leg 3 (Figure 1.34A), all of which were at a higher relative abundance in leg 1, these classes were *Acidobacteria-6*, *Anaerolineae*, *Planctomycetia*, *Chloracidobacteria*, *and Spartobacteria*. There were 6 total classes of significantly differentially abundant bacteria identified when comparing leg 2 vs leg 3 (Figure 1.34B), which were *Deltaproteobacteria*, *Spirochaetes*, *Thermoleophilia*, *Chloracidobacteria*, *Cloacamonae*, and *Spartobacteria*.

4.3.6 Core microbiome

The core microbiome, as defined as any taxa present at a relative abundance > 0.01% in at least 80% of samples, was identified for each sampling leg and the expedition as a whole. During leg 1, *Actinobacteria*, *Clostridia*, and *Gammaproteobacteria* were identified with *Gammaproteobacteria* holding the highest relative abundance (figure 1.35A). Leg 2 had the largest number of classes making up the core microbiome, with the addition of *Bacilli* to the core microbiome shown in leg 1 (Figure 1.35B). Leg 3 had the lowest number of core classes (2), which were *Actinobacteria* and *Gammaproteobacteria* (Figure 1.35C). Looking at the expedition as a whole, *Actinobacteria*, *Clostridia*, and *Gammaproteobacteria* made up the core microbiome; *Clostridia* and *Gammaproteobacteria* relative abundance increased during the course of the expedition, whilst *Actinobacteria* were least relatively abundant during leg 2. No genera were identified as core community members during any of the 3 expedition legs, nor when looking at the expedition as a whole. This remained the case when the core microbiome threshold was dropped from 80% to 50%.

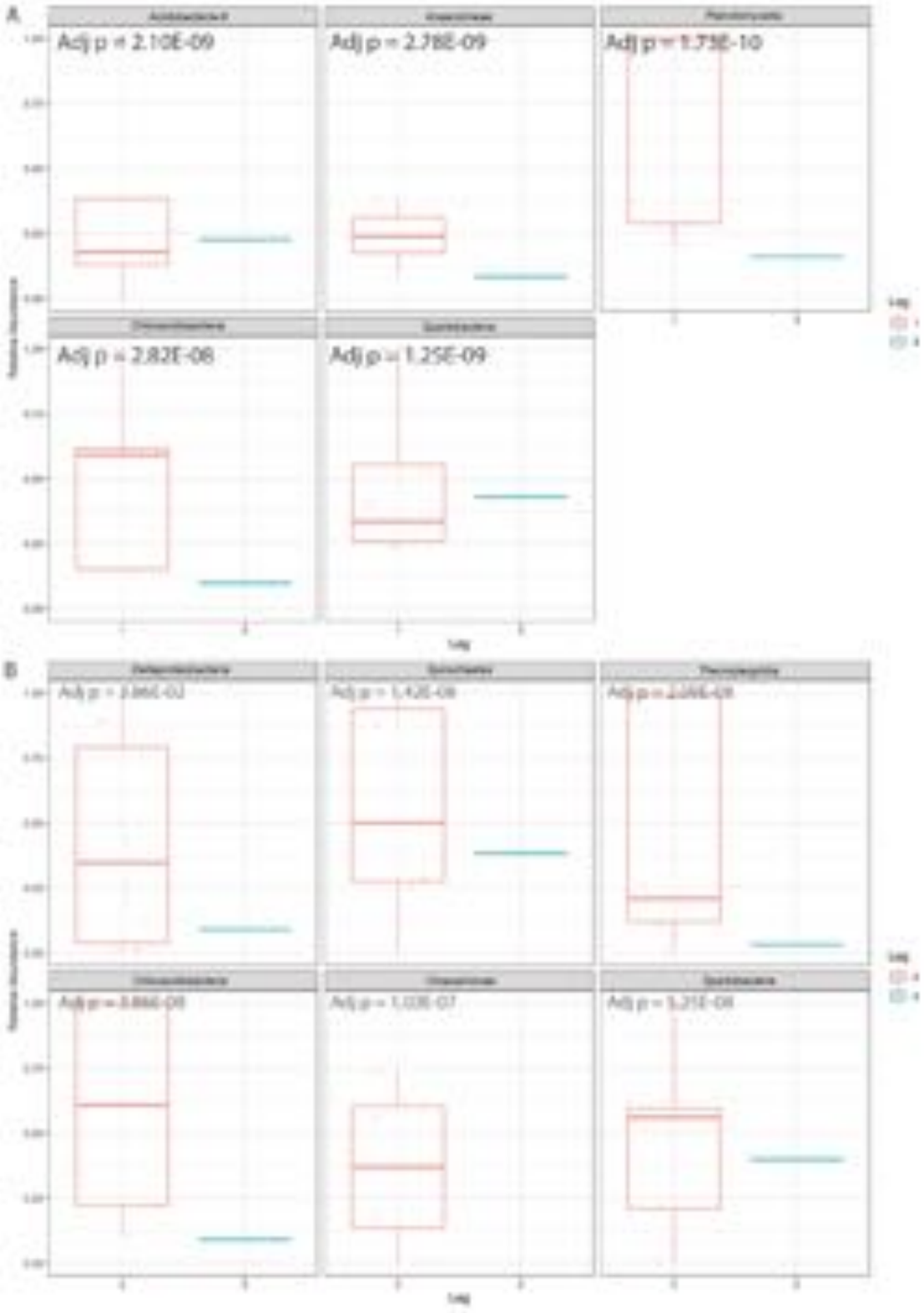


Figure 1.34. Boxplots showing the differentially abundant taxa as identified by DeSeq2 between A) Legs 1 and 3 and B) Legs 2 and 3

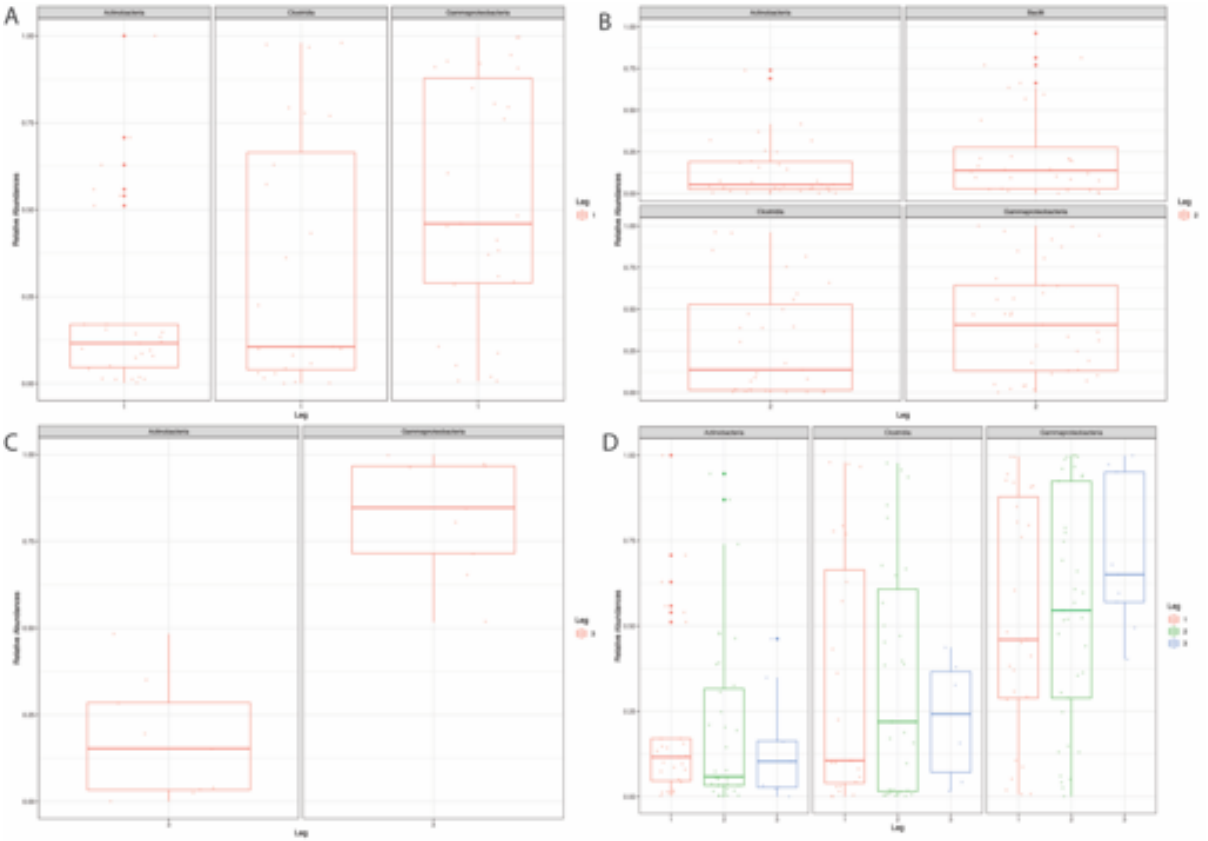


Figure 1.35. Box plots showing the relative abundance of the core microbiome for each of the 3 expedition legs A) Leg 1, B) Leg 2, C) Leg 3, and D) the entire expedition

4.3.7 Antarctic circumpolar current

Samples taken above -60°S (pre-ACC) and below -60°S (post-ACC) showed no significant difference in Observed OTUs (Figure 1.36A) where p = 0.47 and median values were 8 (post-ACC) and 9 (pre-ACC) and Shannon Index (Figure 1.36B) where p = 0.28 and median values were 1.3 (post-ACC) and 1.55 (pre-ACC). Additionally, there was no significant difference in the Bray-Curtis beta diversity metric between pre and post-ACC (PERMANOVA p = 0.83).

There were 3 significantly differentially abundant taxa when comparing within the Antarctic vortex (post-ACC) and outside the Antarctic vortex (pre-ACC), which were Anaerolineae, Phycisphaerae, and Synergistia, all appearing at significantly higher relative abundances pre-ACC (Fig 1.37).

The core microbiomes of both groups contained 3 classes. Both pre-ACC and post-ACC contained *Actinobacteria* and *Gammaproteobacteria* as core members; within the Antarctic vortex (post-ACC), *Bacilli* was also a core member (fig 1.38A), whilst outside of the ACC (pre-ACC) *Bacilli* were replaced by *Clostridia* (fig 1.38B). No genera were identified as core community members at either condition. This remained the case when the core microbiome threshold was dropped from 80% to 50%.

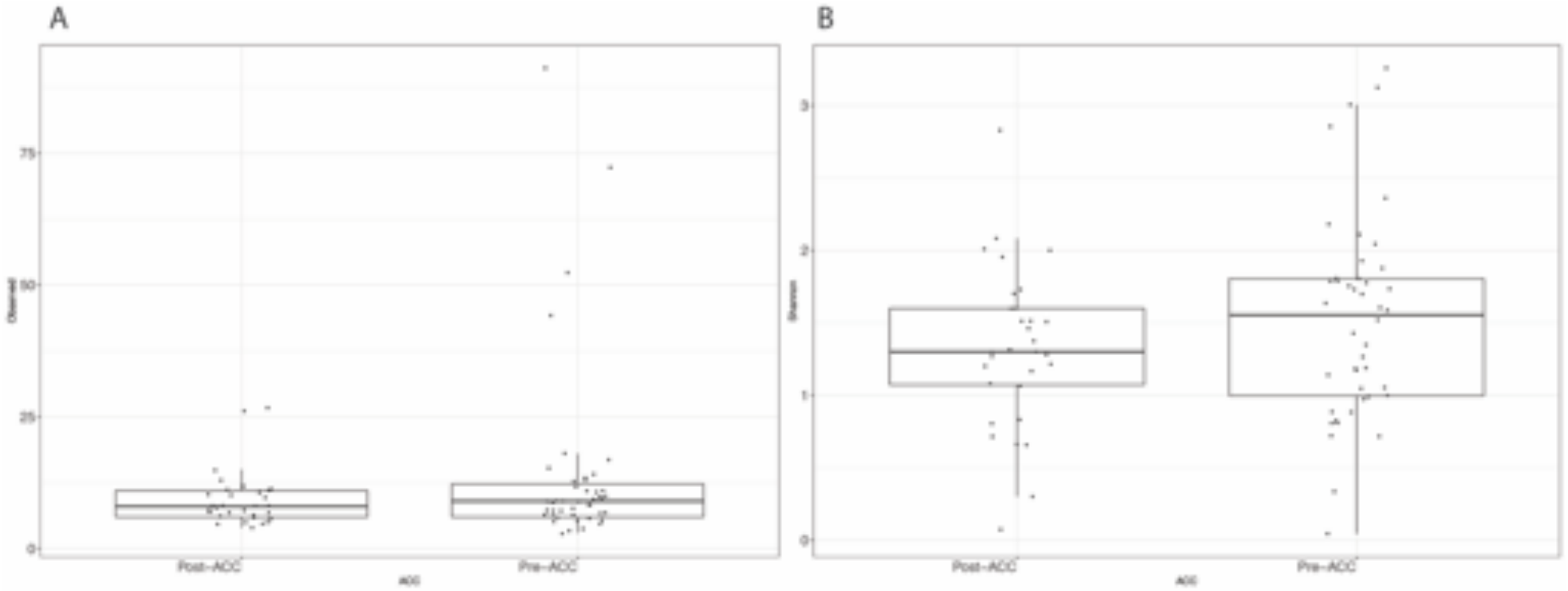


Figure 1.36. Boxplots showing alpha diversity metrics. A) Observed OTUs and B) Shannon Index for pre-acc and post-acc

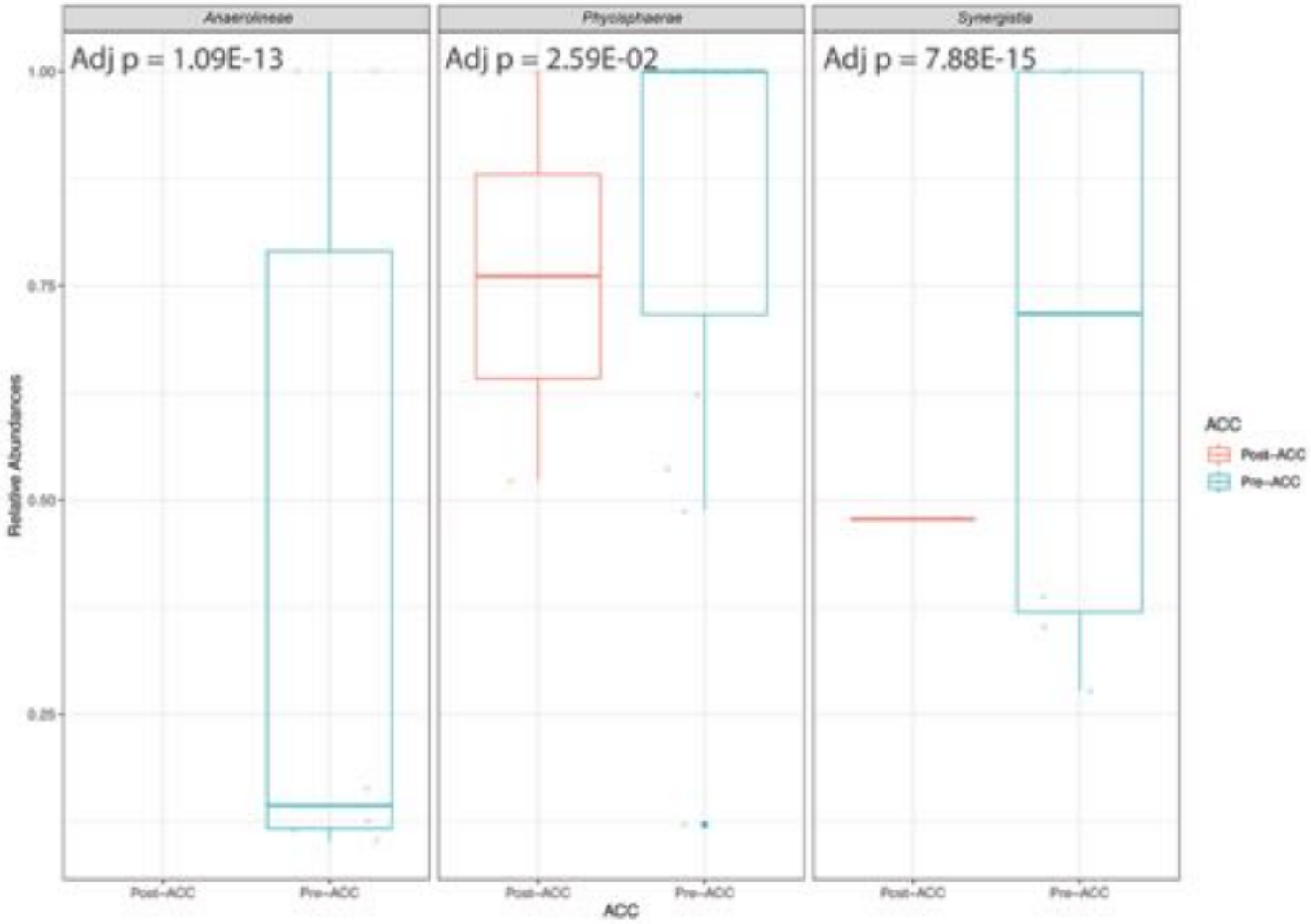


Figure 1.37. Boxplot showing the differentially abundant taxa as identified by DeSeq2 between post and pre acc

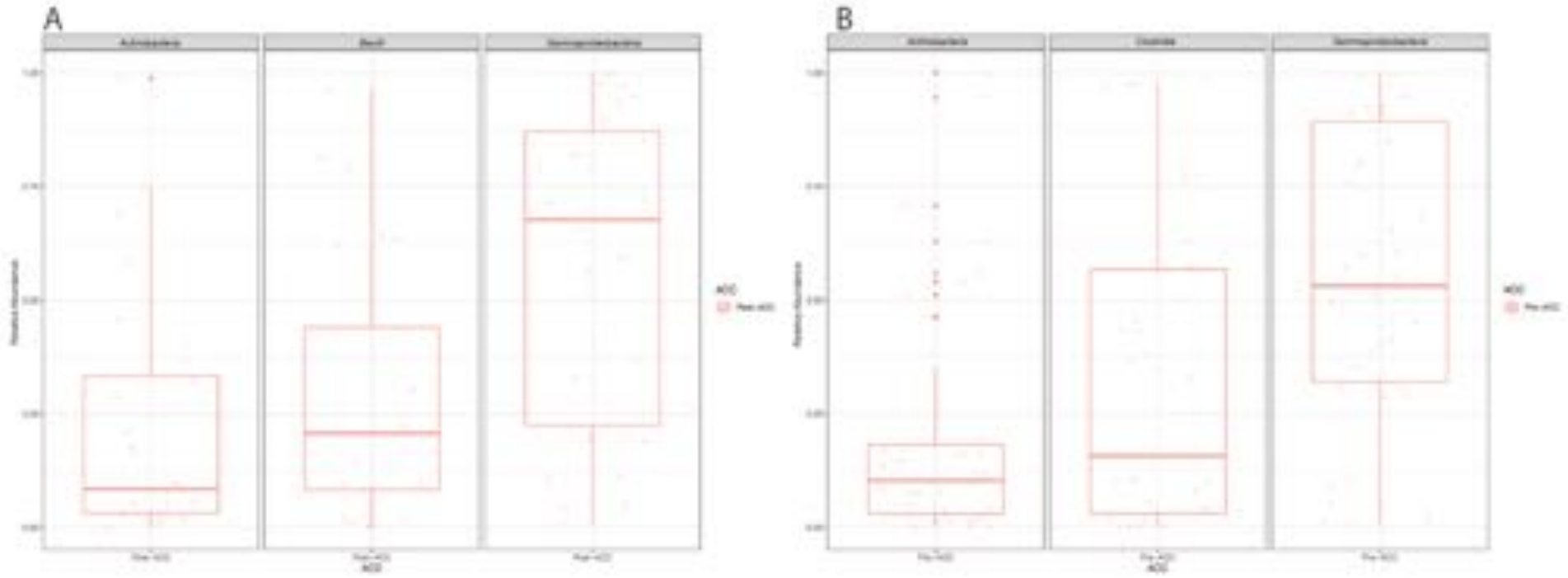


Figure 1.38. Box plots showing the relative abundance of the core microbiome for A) post acc and B) pre acc

4.3.8 Precipitation samples

Four rain samples contained a total of 2,054,649 reads, with the smallest sample containing 322,973 and the largest 718,468. Sufficient sampling depth was achieved for rain amplicon libraries, as shown by each curve reaching asymptote when samples were rarefied to 300,000 reads (lower than the smallest sample) (figure 1.39). Observed OTUs were 1266, 133, 277, and 217 for rain samples 1-4 respectively, whilst the Shannon Index was 2.86, 0.85, 1.29, and 1.04. 99.7% of the inter sample diversity was explained over two principle components when exploring beta diversity emphasising the similarity of the samples (figure 1.40). The composition of the 10 most abundant phyla, showed all 4 rain samples to be dominated by *Proteobacteria* (1.41). At class level, the *Proteobacteria* were comprised primarily of *Betaproteobacteria*, followed by *Gamma-* and *Alpha- Proteobacteria* (1.42).

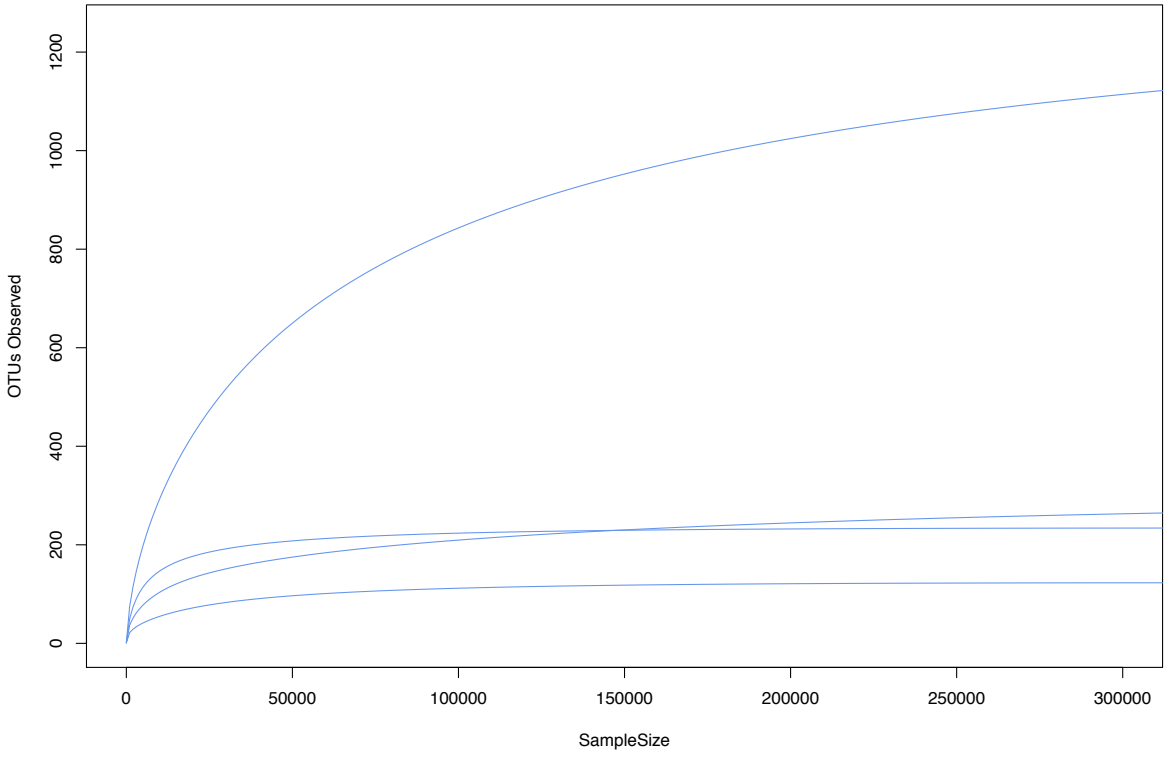


Figure 1.39. Rarefaction curve for rain samples

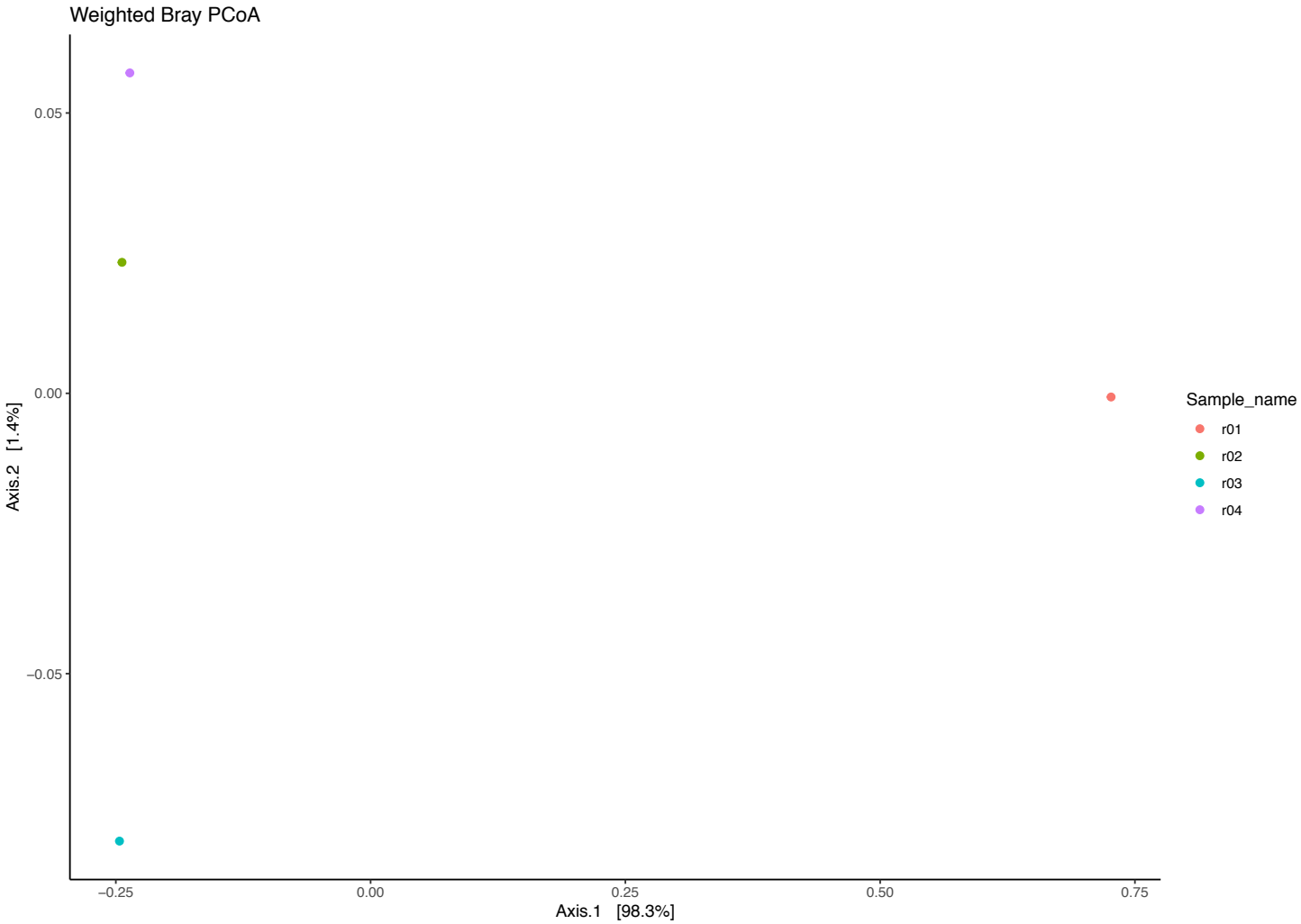


Figure 1.40. Principle Coordinate Analysis displaying the Bray-Curtis dissimilarity between rain samples

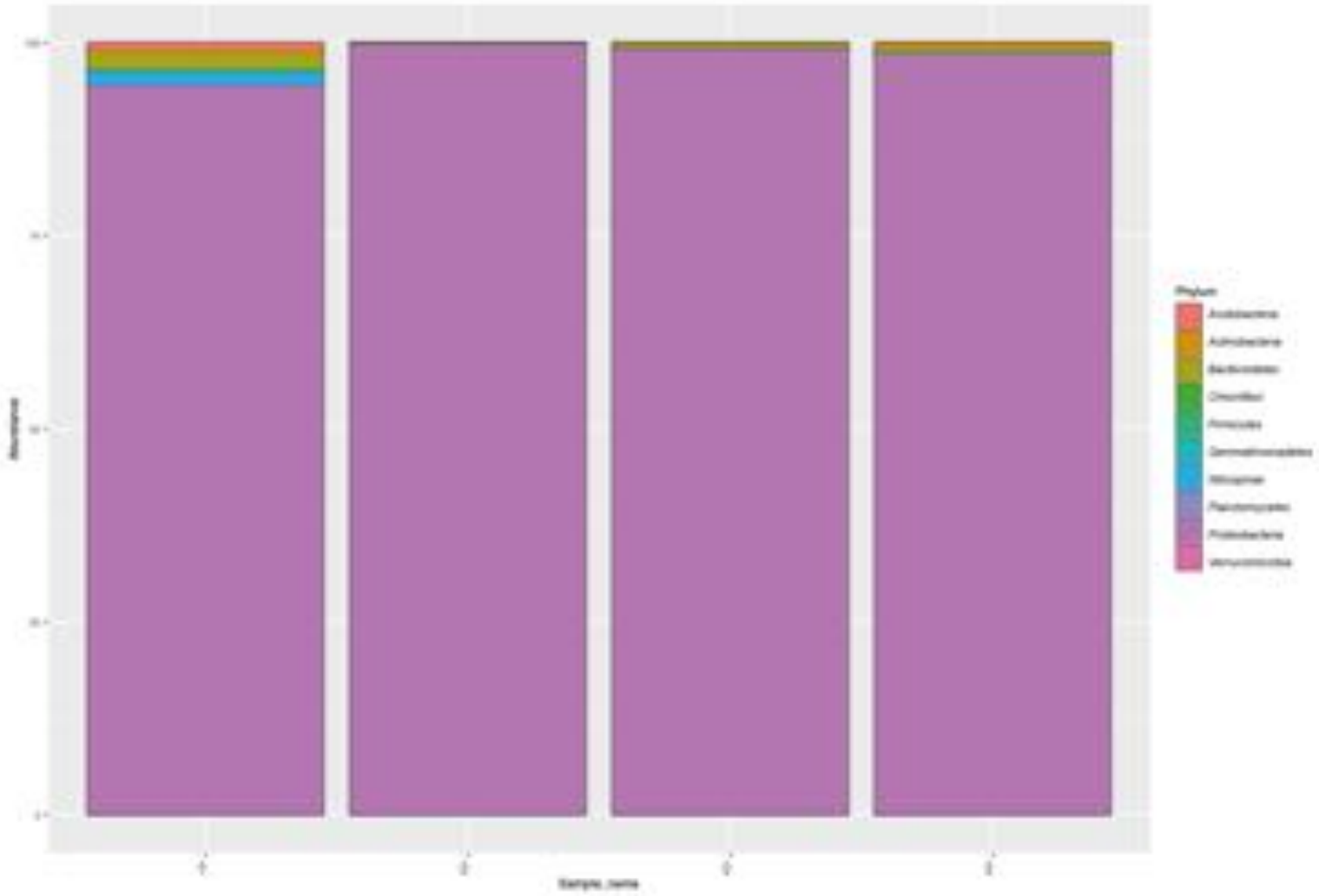


Figure 1.41. Stacked bar showing the relative abundance of the top 10 most abundant phyla for rain samples

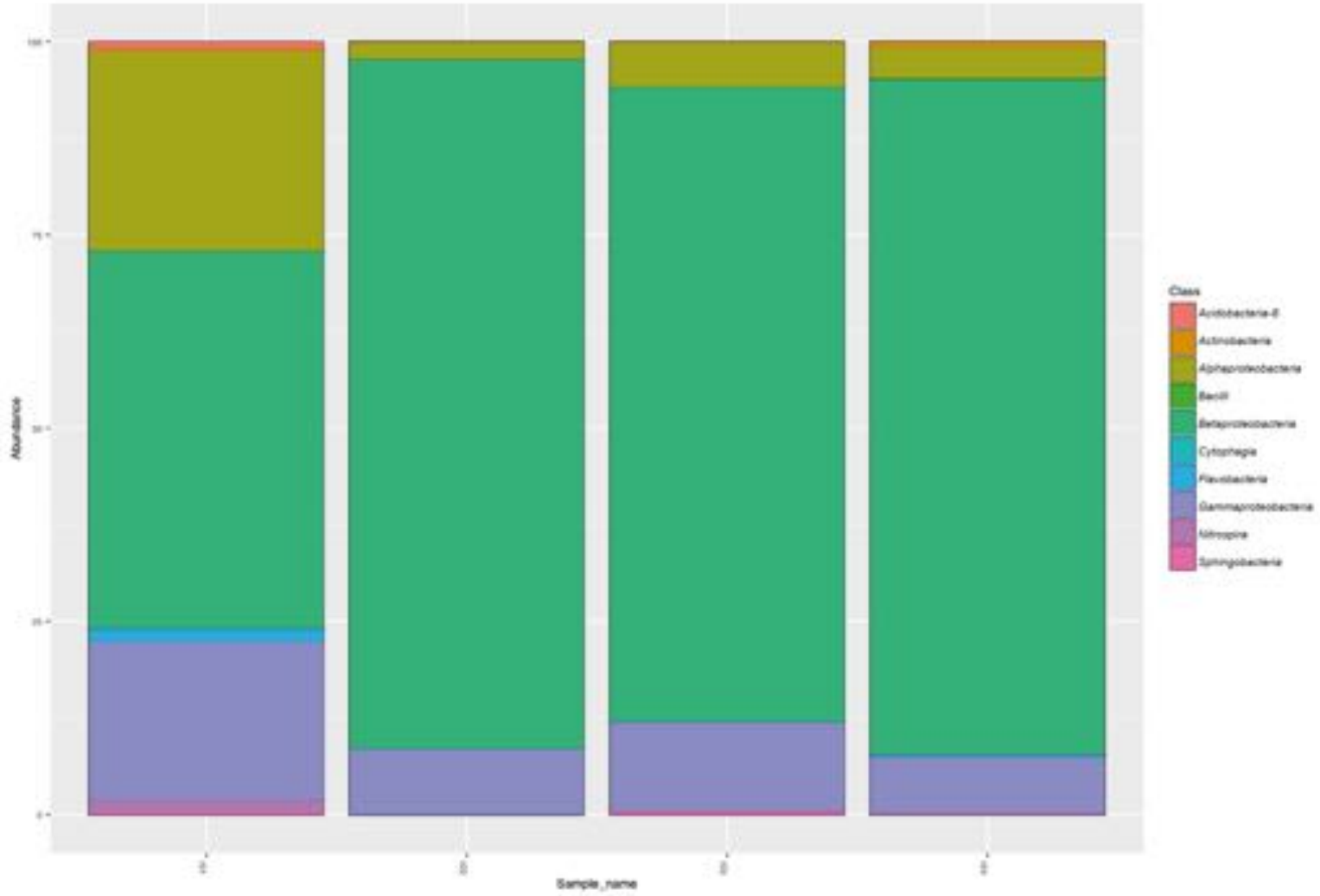


Figure 1.42. Stacked bar showing the relative abundance of the top 10 most abundant classes for rain samples

4.4 Discussion

Understanding the macroecological patterns within the Antarctic has long been of scientific importance, due to the ability of the region to act as a baseline for environmental studies because of a lack of human activity in the region and the remoteness of the continent (151). Terrestrial studies in the region have shown a clear relationship between local sources such as penguin colonies or dry valley soils and the bacterial communities residing in the adjacent air (214, 221).

Bioaerosols are known to influence local bacterial communities within the region. One study of lake Untersee revealed that aerosol deposits influenced microbial community development in glacier ice and cryoconite holes, therefore influencing the microbial mats of the subglacial lake (222). The relationship between local communities and bioaerosols highlights the importance of understanding the transfer of bacteria into and out of the Antarctic continent, and therefore their large scale spatial distribution around the region, if we are to truly understand the bacterial community dynamics.

All 75 air samples amplified and contained reads following quality filtering, showing bacteria to be ubiquitous in the atmosphere surrounding the Antarctic at both marine and terrestrial sites. All sample rarefaction curves reached asymptote showing the full class level diversity was collected at all sample sites. When comparing alpha diversity by leg of expedition, the number of observed OTUs and Shannon diversity showed no significant difference; in fact, median values for both alpha diversity metrics were similar for all legs of the expedition. Temperature difference is often touted one of the key limiting factors preventing colonisation by long range transported bacteria (151).

Previous studies of other Antarctic environments have found air temperature to be a significant influence on the alpha diversity of bacterial communities (212); however, despite a

significantly lower temperature during leg 2 of the expedition, no significant difference in alpha diversity was noted, additionally, when not stratifying samples by expedition leg, there was no significantly positive relationship between alpha diversity and temperature; therefore suggesting that temperature is not an influencing factor on the class level alpha diversity of bacterial communities residing in the atmosphere surrounding the Antarctic. None of the other environmental parameters tested showed a positive relationship with sample alpha diversity either suggesting that none of the variables are singularly responsible for community biodiversity.

Samples did not cluster by leg of expedition, and there was no significant difference in Bray-Curtis beta diversity between expedition legs. There was also no clustering of marine or terrestrial sites; clustering of terrestrial samples was not expected due to their geographic separation, however marine samples collected near to the sites of terrestrial samples also did not cluster tightly, suggesting that there is little interchange of bacteria between the sub-Antarctic islands of French Southern and Antarctic Lands, Siple, and South Georgia and their surrounding maritime environment.

45 phyla were recorded in total across all samples, a similar number of phyla to studies found on the continent (222), however a number considerably higher than that of bioaerosol studies in the Arctic (220). *Proteobacteria*, *Firmicutes*, *Actinobacteria* and *Bacteroidetes* were the dominant phyla, appearing on most sampling days. The majority of phyla appeared sporadically and at low relative abundances. This pattern is shared with a range of both Antarctic environments (211, 219, 222) and environments across the globe (24, 139, 217, 219, 220).

At class level, *Gammaproteobacteria* were the dominant taxa. *Gammaproteobacteria* are known to be ubiquitous it global waters (217), and have been shown to be dominant in the

Southern Ocean (218), which could provide a source for the *Gammaproteobacteria* in marine samples. The presence of *Gammaproteobacteria* in bioaerosol samples in the Antarctic region has also been shown to be directly associated with vicinity to penguin colonies at both up and down wind locations, with penguin faeces shown to be the primary source of the bacteria (221). Sample 33, collected on January 27th 2017 during Leg 2 of the expedition close to the Adelie coast of continental Antarctica, was dominated by *Gammaproteobacteria* (figure 1.32), >99% of the *Gammaproteobacteria* representation was of the genus *Psychrobacter*, which has previously been described in penguin faeces (223).

Clostridia and *Bacilli* were the next most prevalent bacterial classes, appearing less frequently and less abundantly than *Gammaproteobacteria*, but still in the majority of samples. *Bacilli* and *Clostridia* are both spore-forming classes of the *Firmicute* phylum (193), making them hardier and more suited to life in the challenging environment that is the atmosphere, explaining their frequent presence in samples. *Clostridia* and *Bacilli* were the most abundant members of the *Firmicute* phylum in all samples, which matches findings on continental Antarctica at the McMurdo Dry Valleys (152).

Spartobacterium were more relatively abundant in samples leg 2 than in leg 3, and in leg 3 than in leg 1. Aquatic *Spartobacteria* have been shown to be negatively correlated with water salinity (224), which could be a contributing factor to this pattern, as the class was most differentially abundant in samples collected whilst sailing through the lowest saline waters off the coast of Antarctica. Chlor*acidobacteria*, a class previously described in maritime Antarctic soils (212), were significantly more relatively abundant during legs 1 and 2 than during leg 3. *Thermoleophilia* were significantly more abundant during leg 2 than they were during leg 3; members of this class have the ability to form endospores making them suitable to life in the atmosphere, and have previously been described in Eastern Antarctic soils explaining their prevalence during leg 2 of the expedition (221).

The core microbiome, that is the classes of bacteria which were present in at least 80% of all samples, of the entire expedition, contained *Actinobacteria*, *Gammaproteobacteria*, and *Clostridia*. These classes of bacteria are almost ubiquitous across Antarctic environments, found in snow, soil and subglacial lakes (209, 211, 219). *Actinobacteria* were at their lowest during leg 2, whilst *Clostridia* and *Gammaproteobacteria* relative abundance increased throughout the entire expedition longitudinally. Leg 3 contained the simplest core community containing only *Gammaproteobacteria* and *Actinobacteria*, whilst leg 2 was the most complex with *Actinobacteria*, *Bacilli*, *Clostridia*, and *Gammaproteobacteria*. The increased number of core microbiome classes during leg 2 are likely due to the fact the closer vicinity to continental Antarctica and as such an increased number of source environments such as soils and Penguin faeces. *Actinobacteria*, *Bacilli*, and *Clostridia* all possess the ability to spore-form (225-227), making them ideally suited to the extreme UV and desiccating challenges of atmospheric life. Despite the apparent core microbiome at class level, there were no genus level taxa which were present in more than 40% of all samples, showing extreme diversity in the consistent classes.

Previous findings suggest that the Antarctic Circumpolar Current/Vortex acts as a major biogeographical boundary (228), however there was no evidence of this when separating samples by pre and post polar vortex, with no significant difference in either alpha or beta diversity. There were 3 taxa which were significantly more abundant outside of the polar vortex. *Anaerolineae*, which are widespread typically thermophilic, anaerobic, wastewater associated taxa, *Phycisphaerae* a marine bacterial class and the highly diverse *Synergistia* class which contains anaerobic wastewater genera along with a plethora or healthy and pathogenic human associated taxa. These 3 classes provide evidence for a small degree of selectivity by the polar vortex, however the equal dispersal of the remaining 111 classes suggests limited dispersal limitation. The core microbiome of both inside and outside of the polar vortex contains only 3 of the 114 classes, namely *Actinobacteria*, *Gammaproteobacteria* which were

part of both communities with *Bacilli* and *Clostridia* completing inside and outside of the vortex respectively. Whilst *Bacilli* and *Clostridia* were part of both inside and outside of the vortex communities, the difference in prevalence's suggest some degree of selectivity between the two regions.

Rainwater samples were dominated by *Betaproteobacteria*, similar to rain samples of marine and European origin collected in the Austrian alps, however the number of observed OTUs and Shannon index values were considerably higher showing the bacterial communities in rainwater collected above the Indian and Pacific Ocean to be more rich and even, however the Austrian study may have under sampled the environment due to a lack of saturation on their rarefaction curves (229). *Alpha-*, *Beta-*, and *Gamma- Proteobacteria*, were present at high abundance in all rain samples, and present frequently in air samples too, and as such rain may provide one general source for these bacteria into the aerial environment, however there was no distinct pattern when comparing the rain and air samples from the same day, or the air samples for the day following a rain event. The third most relatively abundant ASV identified in all rain samples was an uncultured member of the *Pseudomonas* genera, a National Centre for Biotechnology Information (NCBI) BLAST (basic local alignment search tool) search showed the sequence for this taxa matched closest to *Pseudomonas fluorescens* strain ESR7, a known IN active bacteria (230). The relative abundance of this sequence was highest on day 12 (r01), and cloud cover was measured at 7 Octants, the relative abundance decreased on day 15 (r02), where cloud cover reduced to 5 Octants, whilst the relative abundance of the sequence was lowest on day 27 (r04), where cloud cover was lowest at 3 Octants; this suggests that *Pseudomonas fluorescens* has an intimiate role with cloud cover over the oceans surrounding Antarctica.

4.5 Concluding remarks

i) This bioaerosol sampling regime, covering the largest longitudinal sampling range to date in the region, has revealed the great extent to which bacteria inhabit the atmosphere surrounding the Antarctic continent, with bacterial DNA present in all samples, meaning the hypothesis that bacteria are ubiquitous in the air surrounding the Antarctic can be accepted.

ii) Class level analysis appeared to show homogeneous bacterial communities around the Antarctic, however the lack of core microbiome at genus level showed there to be unprecedented levels of diversity in the atmosphere surround Antarctica. Based upon these findings, the hypothesis that bacterial communities surrounding the Antarctic are homogeneous must be rejected.

iii) The dominant phyla were similar to those previously identified at a wide range of Antarctic locations. The core microbiome in the atmosphere above the oceans surrounding the Antarctic was made up of two hardy, spore forming classes *Actinobacteria, Clostridia, and Gammaproteobacteria* with the latter likely attributed in part to local animal faeces. The links between local sources and bioaerosol studies allow the hypothesis that local sources contribute considerably to bioaerosol communities around the Antarctic to be partially accepted, however the lack of core community at genus level inhibits the full acceptance of this hypothesis.

iv) There was no significant difference when comparing the biodiversity of marine samples against terrestrial samples, therefore the hypothesis that marine bioaerosol communities differ from terrestrial communities must be rejected.

v) The biodiversity of samples collected before and after crossing the Antarctic Circumpolar Current were not significantly different, suggesting the extent to which

the current had previously been touted as a limiter to dispersal could have been overestimated. Therefore, the hypothesis that bacterial communities collected above or below -60°S would harbour distinct biodiversity patterns must be rejected.

Chapter 5 - Detection limits of low biomass bioaerosol samples

5.1 Introduction

To date, there are no studies informing best practice when carrying out next generation sequencing (NGS) based studies of low biomass samples stored on membrane filters. The affordability and turnaround times of NGS technologies has allowed the bacterial community profiles of many previously understudied low biomass environments to be explored, providing more ecological insight than has previously been available for both environmental (152, 211, 231) and clinical studies (232-234). New findings facilitated by NGS technologies can prove to be integral to our understanding of ecosystem succession or microbiome-host interactions, even acting as a guide for further research. Therefore, it is imperative to validate NGS data in order to give an accurate characterisation of the target microbial community.

PCR is extremely sensitive and has the ability to amplify from a single or very small number of molecules (235), for that reason it is challenging to attain a completely blank negative control. Contaminating bacterial DNA can be introduced to samples from a range of exogenous sources during sample collection, DNA extraction, and library preparation. Whilst precautions can be taken during sample collection, it is unlikely that a sample is completely independent of bacterial DNA from its surroundings. The environment specific streamlined extraction protocol of commercially available DNA extraction kits allows for the high throughput of samples, which when combined with their relative affordability, means they are by far the most commonly used method in the literature. It has also been shown the choice of DNA extraction kit can have a significant impact on DNA yield and bacterial DNA composition, therefore if the study is to be cross-compared to previous work, choice of DNA extraction kit is extremely important (236). Whilst there are many positive aspects of commercially available DNA extraction kits, contaminating bacterial DNA is known to be present in all commonly used kits (94). During library preparation, contaminating bacterial DNA can occur in PCR reagents such

as lab grade molecular water (237) or Taq polymerase (238). Exogenous contamination prior to library loading is not the only method by which new taxa can be introduced into samples. Cross talk between samples during sequencing is known to account for up to 2% of all reads in a sample (57), and chimeric sequences could also account for misleading taxa, with taxonomic databases potentially made up of as much as 46% chimeric sequences (239, 240). Variability in the accuracy of NGS results can be attributed to differences in library preparation protocol. Choice of Taq polymerase, the number of rounds of PCR amplification, choice of primer set, and 16S target region can all be attributed to differences in the community profile of 16S sequencing results (241, 242).

Contamination provides the biggest challenge to the validity of low biomass studies, in extreme cases, findings where studies have not properly controlled for contamination have been questioned (232, 243). Guidelines have been suggested in order to attempt to address the issue. One example of minimum standard guidelines put forward is the recent 'RIDE' guidelines which suggest reporting methodology (R), including negative controls (I), determine the level of contamination prior to analysis via controls (D), and exploring the impact of potential contaminants in downstream analysis (E) (244).

There is currently no consensus on how to effectively remove as much contaminating DNA from a sample set as possible. Prior to sequencing, extracting numerous negative controls and strenuous sterile technique are core to the reduction of contaminant DNA in low biomass samples; beyond this, additional steps may be taken such as the addition of a dsDNAse treatment to lab reagents prior to the addition of sample DNA, which has been shown to reduce contaminating reads generated during PCR by as much as 99% (245). Downstream, the use of prevalence patterns of taxa in true and negative samples, and DNA concentration correlations between taxa can be used to screen samples for potential contaminants (246), however these patterns are not common to all sample types and methods of identifying the contaminating taxa

rely on a minimum of 5-6 negative controls (as defined by Benjamin Callahan during communication). Source tracker is another potential way to remove contaminating sequences based on negative controls which represent potential sources for contamination (247), however this approach performs poorly when the experimental environment is poorly understood (248). Lists of contaminant genera are gradually being compiled within the literature (94, 244), however these lists must be treated with caution, as what may be a contaminant for one study may well be a true feature of a community in another, meaning these lists can be used to make an informed decision with regards to potential contaminants on a study by study basis. The biomass of a sample is known to impact the proportion of the true community which can be observed, due to the increased prevalence of contaminants as samples biomass decreases (94, 248). The number of observable taxa is has also been shown to increase with decreasing sample concentration (248).

Here, the efficiency of the Qiagen Powersoil kit (Hilden, Germany) at extracting bacterial DNA from membrane filters, as well as how the input biomass, number of rounds of PCR, and the addition of Arcticzyme impacted on the variability of community profiles obtained by MiSeq sequencing was investigated. Along with this, the reproducibility of reagent negatives was explored, in order to ascertain whether their profiles were as consistent as had previously been suggested, as well as how using kit negatives to screen for contaminants impacts community profile. The overarching aim of these investigations was to inform a best practice protocol for the sequencing of low biomass samples stored on membrane filters, by addressing the following hypotheses:

i) The Qiagen Powersoil kit efficiently and reproducibly extracts bacterial DNA stored on membrane filters within the concentration range of typical bioaerosol samples

ii) Sequenced kit negative controls provide reproducible and homogeneous community profiles

iii) Illumina MiSeq is a suitable instrument for the detection of low biomass airborne bacterial communities

iv) The unprecedented sequence level biodiversity of Antarctic air samples was due in part to technical variation as a result of their low biomass

5.2 Methodologies

5.2.1 Preparation of samples

Bacillus subtilis (NCTC 8236) was spread onto Nutrient Agar and incubated for 24 hours at 37°C. Colonies were then added to sterile 1X phosphate buffered saline (PBS) (Thermo Fisher Scientific, MA, USA). The suspension was adjusted to 0.5 McFarland units containing an approximate cell suspension of 1–1.5 × 10^8 CFU/mL^{-1} when measuring an absorbance of 0.132 at OD_{600nm}. A serial dilution was then performed down to 10^2 CFU/mL^{-1}. 4mL of each dilution was filtered onto a 47 mm × 0.2 μm pore size cellulose nitrate membrane filter (GE Healthcare Life Sciences, Chicago, IL, USA). DAPI counts were performed as described in chapter 2.2.11 in order to give enumerate the total bacterial load in each dilution. DNA extraction was carried out on ¼ of each filter in triplicate for all dilutions using a Qiagen Powersoil kit (Qiagen, Hilden, Germany) in an Envair Bio2+ class II microbiological safety cabinet (Lancashire, GB) as described in chapter 2.2.1, in order to generate a dilution series of *B.Subtillis* samples.

5.2.2 Library preparation

DNA extracts were quantified by Qubit (Invitrogen, Thermo Fisher Scientific, MA, USA) amplified at both 30 and 40 cycles as described in chapter 2.2.8. PCR product was cleaned up using Ampure XP beads at a ratio of 0.8, then quantified by Picogreen (Invitrogen, Thermo Fisher Scientific, MA, USA) and normalised into pools as described in sections 2.2.9. Library

preparation was then repeated with the additional step taken of cleaning reagents pre-PCR using ArcticZyme dsDNase (Tromsø, Norway) as follows; 0.5µL of enzyme, 0.5µL of 1mM DTT, 1µL forward primer, 1µL reverse primer, and 17µL of DNA polymerase mastermix per sample were pipette mixed and incubated at 37°C for 15 minutes to allow digestion of double stranded DNA, before the mixture was incubated at 60°C for 15 minutes to fully inactivate the enzyme. Libraries were loaded and sequenced on an Illumina MiSeq as described in chapter 2.2.8.

5.2.3 Sequence processing and analysis

Fastq files generated by 16S Illumina MiSeq were processed into an OTU and taxonomy table in QIIME2 (61), then screened for contaminants using Microsoft Excel (2013) as described in chapter 2.3.1. The OTU table, taxonomy table and metadata files were then read into R studio (R_Core_Team, 2014) and converted into a Phyloseq object (171) for statistical analyses. Diversity analyses were carried out using Vegan (249). Graphics were produced using Microsoft Excel (2013) and ggplot2 (250).

5.2.4 Antarctic air sample biomass

The biomass of a subset of samples from the ACE cruise (see chapter 4) was investigated using qPCR as described in chapter 2.2.7 in order to consider the validity of Antarctic samples in comparison with the known *B.Subtillis* samples processed during this experiment.

5.3 Results

5.3.1 Standard preparation and expected DNA extraction yields

A running mean of 104 cells per field of view was calculated for the third serial dilution of the 0.5 McFarland unit stock (figure 1.43), by DAPI counts taken from florescence microscopy imaging (fig 1.44). This value was then used to calculate the CFU per mL^{-1} of the dilution to

be 7.28 x 10^7. This dilution was used as the high standard for the study as it was considerably higher biomass than the environmental samples being investigated.

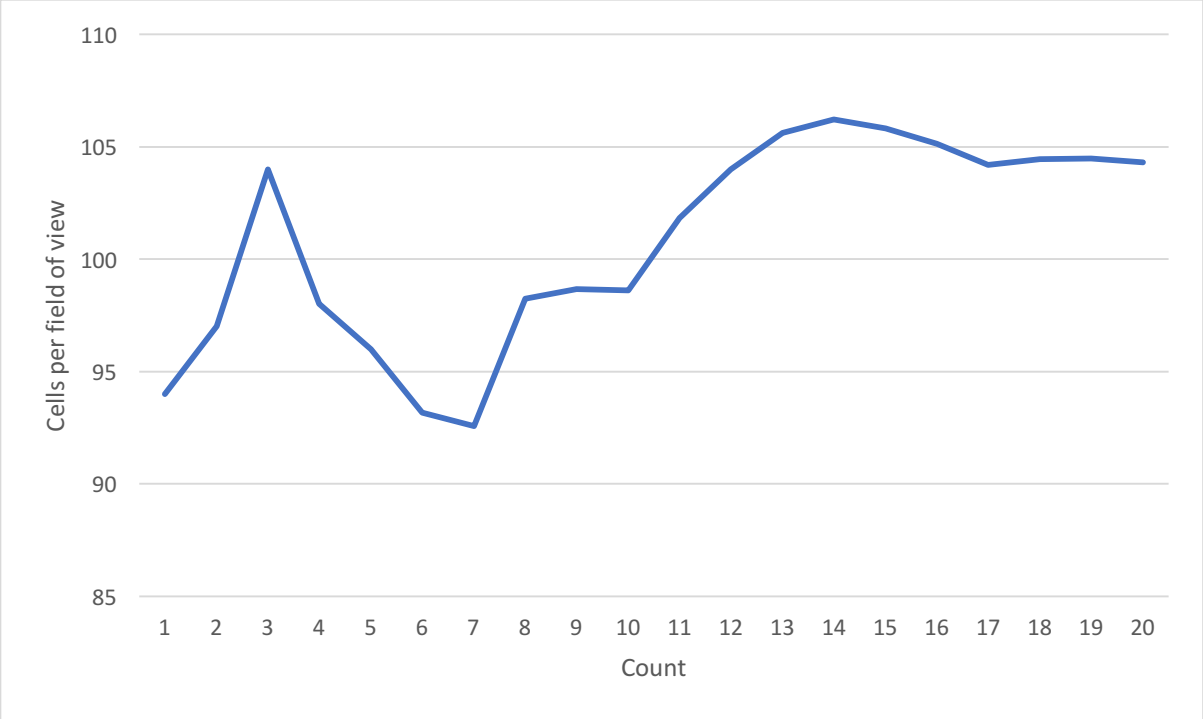


Figure 1.43. Running mean line plot showing counts taken from dilution 3

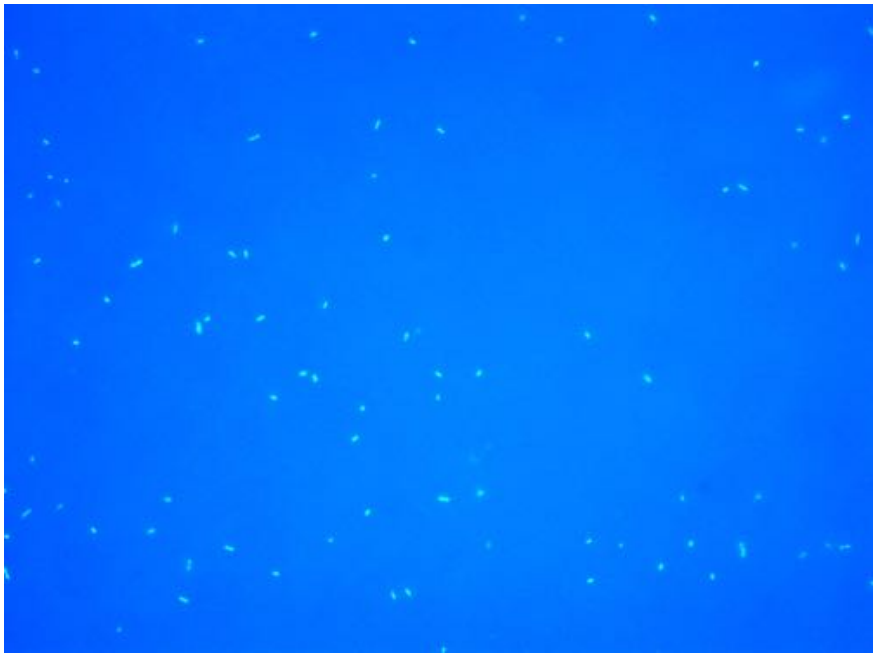


Figure 1.44. 630X magnification image of DAPI stained *B. subtilis* cells

The mass of DNA in a single *B. subtilis* cell was calculated as 4.3E-15g as follows:

$$6710g = \frac{B.subtilis \text{ genome size (bp (4146839))}}{\text{Average molecular weight of DNA basepair } \left(\frac{618\text{g}}{\text{mol}}\right)}$$

$$4.3E-15g = 6710g \; x \; \text{Avogadro's Constant } (6.02214086 \times 1023 \text{ mol} - 1)$$

The mass of a single *B. subtilis* cell was then multiplied by the total number of *B. subtilis* cells in each extraction to give the total DNA input (ng) per extraction. This value was then divided by 4 to give the expected DNA per ¼ filter, assuming the equal distribution of *B. subtilis* cells on each filter (table 1.12).

Cells CFU per mL^{-1}	Total DNA input (ng)	Expected DNA per 1/4 filter (ng)
7.3E+07	1.2E+03	3.1E+02
7.3E+06	1.2E+02	3.1E+01
7.3E+05	1.2E+01	3.1E+00
7.3E+04	1.2E+00	3.1E-01
7.3E+03	1.2E-01	3.1E-02

Table 1.12. CFU per mL^{-1}, total calculated DNA input (ng), and DNA per ¼ filter based on the assumption of equal dispersal across each filter for all 5 dilutions

5.3.2 Assessment of the DNA extraction efficiency based upon percentage recovery

The total extracted DNA yield per ¼ of each filter was then measured by qubit, and this value was compared to the expected maximum yield to calculate the DNA percentage recovery, based upon the assumption that the cells were equally distributed across each filter ¼ (table 1.13). Average DNA percentage recovery was 4% for samples at a starting concentration of 7.3E+07 and 12% at a starting concentration of 7.3E+06, % recovery could not be calculated for lower starting concentrations as the DNA extracts were below the qubit limit of detection

of 0.2ng. Percentage DNA recovery was then used to estimate the true representative number of *B. subtilis* genomes for each extraction (table 1.13).

DNA per 1/4 filter (ng)	% recovery	Estimated extract CFU per mL^{-1}
1.6E+01	5	3.7E+06
8.8E+00	3	2.1E+06
9.6E+00	3	2.3E+06
5.0E+00	16	1.2E+06
3.3E+00	11	7.8E+05
2.8E+00	9	6.6E+05

Table 1.13. DNA concentration values per quarter filter as measured by qubit, total percentage recovery of DNA per ¼ filter, and representative CFU per mL^{-1} based upon extraction efficiency and starting CFU per mL^{-1}. Filter quarters with an expected DNA concentration 3.1ng or below were below the qubit limit of detection.

5.3.3 Comparison of the proportion of samples representing target and non-target read

At an estimated post extraction yield of 2.7E+06 CFU per mL^{-1}, 98% and 97% of all reads for 30 and 40 PCR cycle MiSeq runs respectively were comprised of the target sequence, with a standard deviation of 1% (figure 1.45). For the following dilution, which represented an estimated post extraction yield of 8.8E+05 CFU per mL^{-1}, the proportion of total reads comprised of the target sequence dropped to 58% for 30 cycle PCR and 33% for 40 cycle PCR. The estimated post extraction yield for each following dilution was non-calculable due to DNA concentrations below the qubits limit of detection, however the average percentage of target community represented dropped to 34%, 9%, an 8% for sequential dilutions when amplified at 30 cycles and 35%, 9%, 7% when amplified for 40.

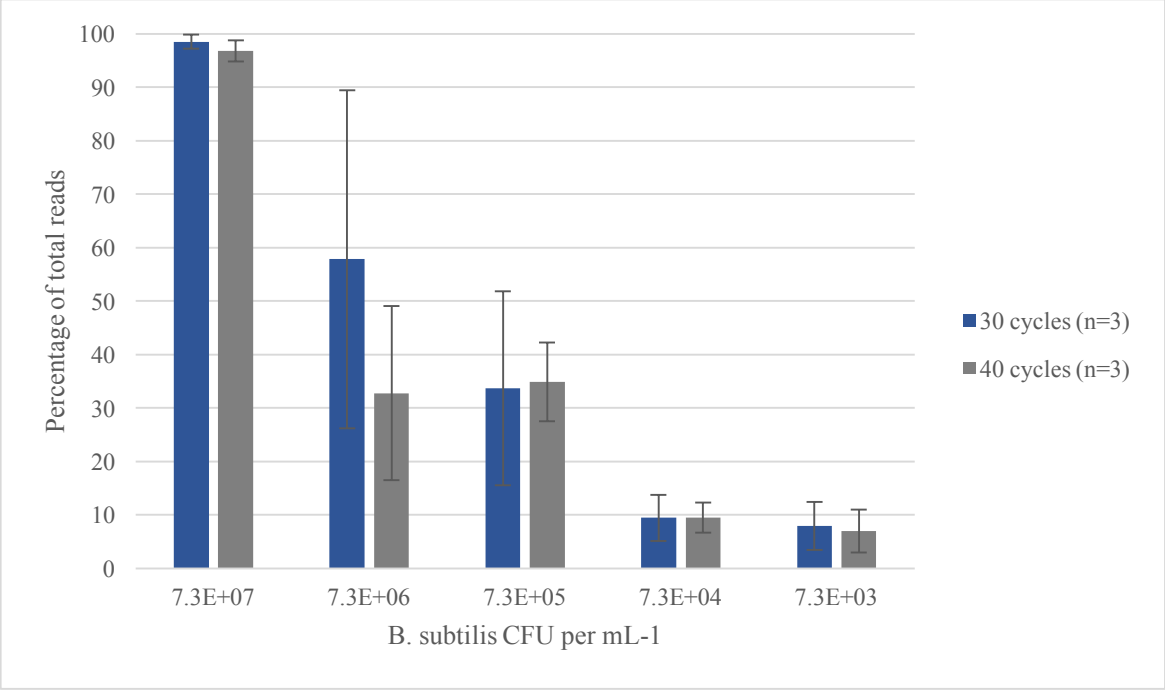


Figure 1.45. Clustered column chart showing the total % reads assigned to *B. subtilis* for each starting concentration.

When splitting the data into individual replicates, a higher degree of variability of the percentage at which target sequence is represented can be seen when dropping below an estimated average input of 2.7E+06 CFU per mL^{-1}. The highest concentration sample from the second highest batch of dilutions, contained a similar estimated input of cells to the highest dilution which was represented by a target sequence representation of 87%, however, for the other two samples at this dilution, which contained 43-82% less estimated input cells, the percentage of target reads dropped to 62% and 24% for samples amplified at 30 cycles and 24% and 23% for 40 cycle samples (figure 1.46).

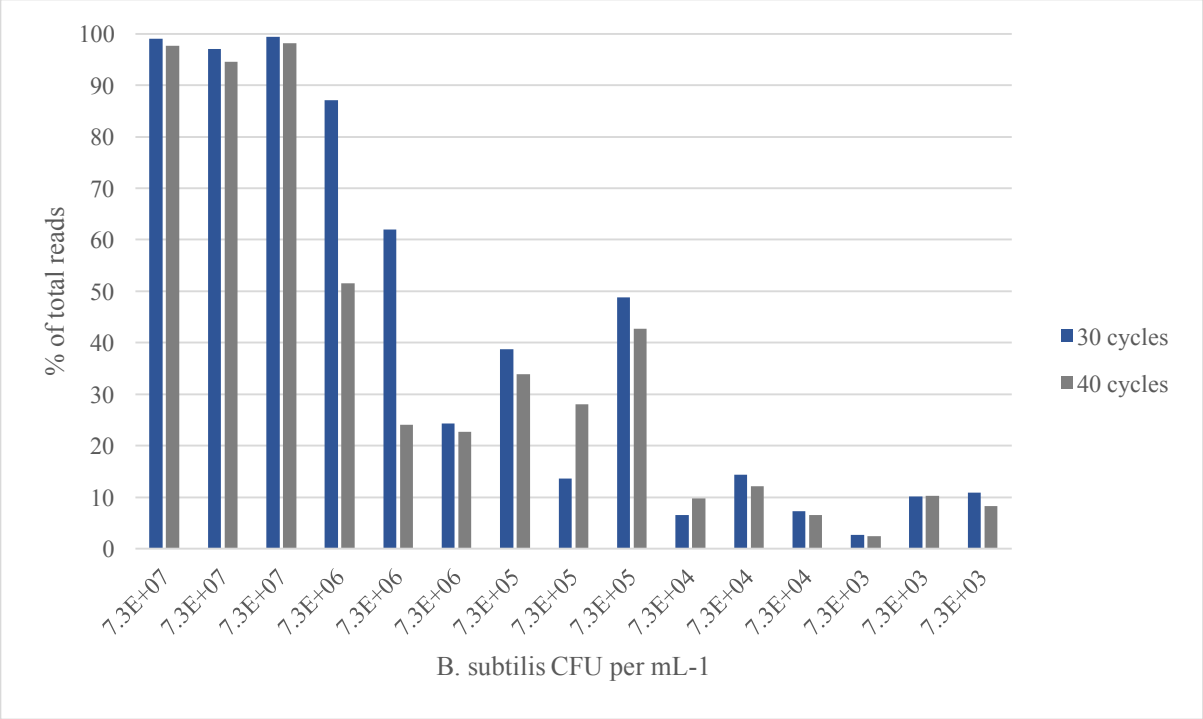


Figure 1.46. Clustered column chart showing the total percentage of reads assigned to the target taxa for each individual replicate.

5.3.4 Target and non-target community composition and diversity metrics

38 non-target bacterial classes (reads belonging to any bacteria other than *B.subtilis*) were represented in samples amplified for 30 cycles, whilst there were 88 for 40 PCR cycle samples (figure 1.47). In total, 98.51% of all reads in the sample with the highest concentration of input *B. Subtilis* DNA belonged to target reads. Non-target contaminant reads were dominated by *Gammaproteobacteria* for all samples, which was the only non-target class at a relative abundance of >1% in the highest concentration sample, this class was present at varying relative abundances in all samples and controls for both 30 and 40 cycle sample sets.

Bacilli was also a prominent non-target class member of the top 20 most relatively abundant classes when sample replicates were combined, however when replicates were viewed independently, *Bacilli* were less prevalent in both 30 and 40 cycle sample sets. *Clostridia* and *Deltaproteobacteria* were also consistently present in joined replicate samples, however again

were less prevalent when samples were viewed independently. The relative abundance of the target sequence was higher for the 30 cycle PCR library than 40 cycle in samples at a starting concentration of 7.3E+07 and 7.3E+06, however there was no clear difference for the lower concentration samples. The lower the concentration of starting bacteria, the higher the relative abundance of non-target reads.

Whilst there was only a single target ASV input into each sample (*B. subtilis*), a total of 362 unique amplicon sequence variants (ASVs) were present when 30 cycle PCR was undertaken, whilst 1268 unique ASVs were present for 40 cycle PCR (figure 1.48). The dominant non-target ASV was a member of the *Enterobacteriaceae* family unclassified at genus level; the relative abundance of this ASV almost mirrored the relative abundance of non-target *Gammaproteobacteria* (figure 1.48). When viewing samples independently, the proportion of *Enterobacteriaceae* ASVs reduced slightly at 40 cycles, this reduction was mirrored by an increase in the remaining 1248 low relative abundance ASVs (figure 1.48). The ASVs which did not make the top 20 most relatively abundant, comprising the other category, were present at a higher abundance in all samples when PCR was carried out at 40 cycles.

When investigating sample alpha diversity (figure 1.49) at 30 cycles of PCR, observed ASVs were lowest in the sample beginning at a starting concentration of 7.3E+07 with a median of 6, and highest in the following dilution (7.3E+06) with a median number of observed ASVs of 21. The number of ASVs observed in negative controls was varied, but remained near or below 25 for all. At 40 PCR cycles, the highest concentration sample again saw the lowest number of observed ASVs, with a median of 21, and again the second dilution saw the highest number of observed ASVs with a median of 132. In negative controls, the value was again variable but higher than at 30 cycles, and whilst the general pattern of samples with higher or lower numbers of ASVs observed remained consistent, kit negative 7 and 8, as well as the facility sequencing

negative differed considerably from their 30 cycle counterparts, with the facility negative containing more than 300 observable ASV.

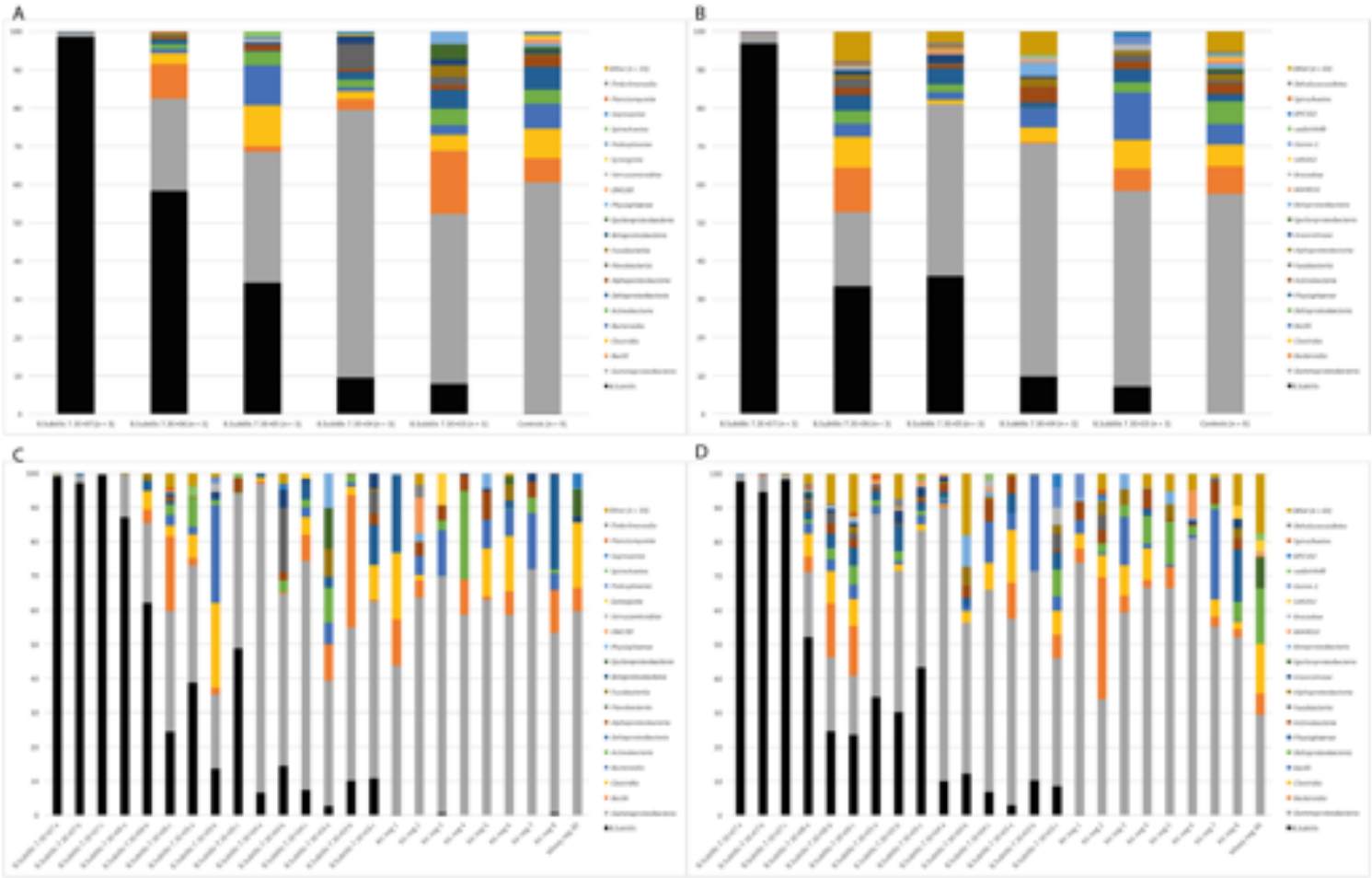


Figure 1.47. Stacked bar charts showing the relative abundances of the top 20 most abundant bacterial classes for *B. subtilis* dilutions and negative controls. A) 30 cycle PCR with sample replicates combined. B) 40 cycle PCR with sample replicates combined. C) 30 cycle PCR with individual replicates. D) 40 cycle PCR with individual replicates

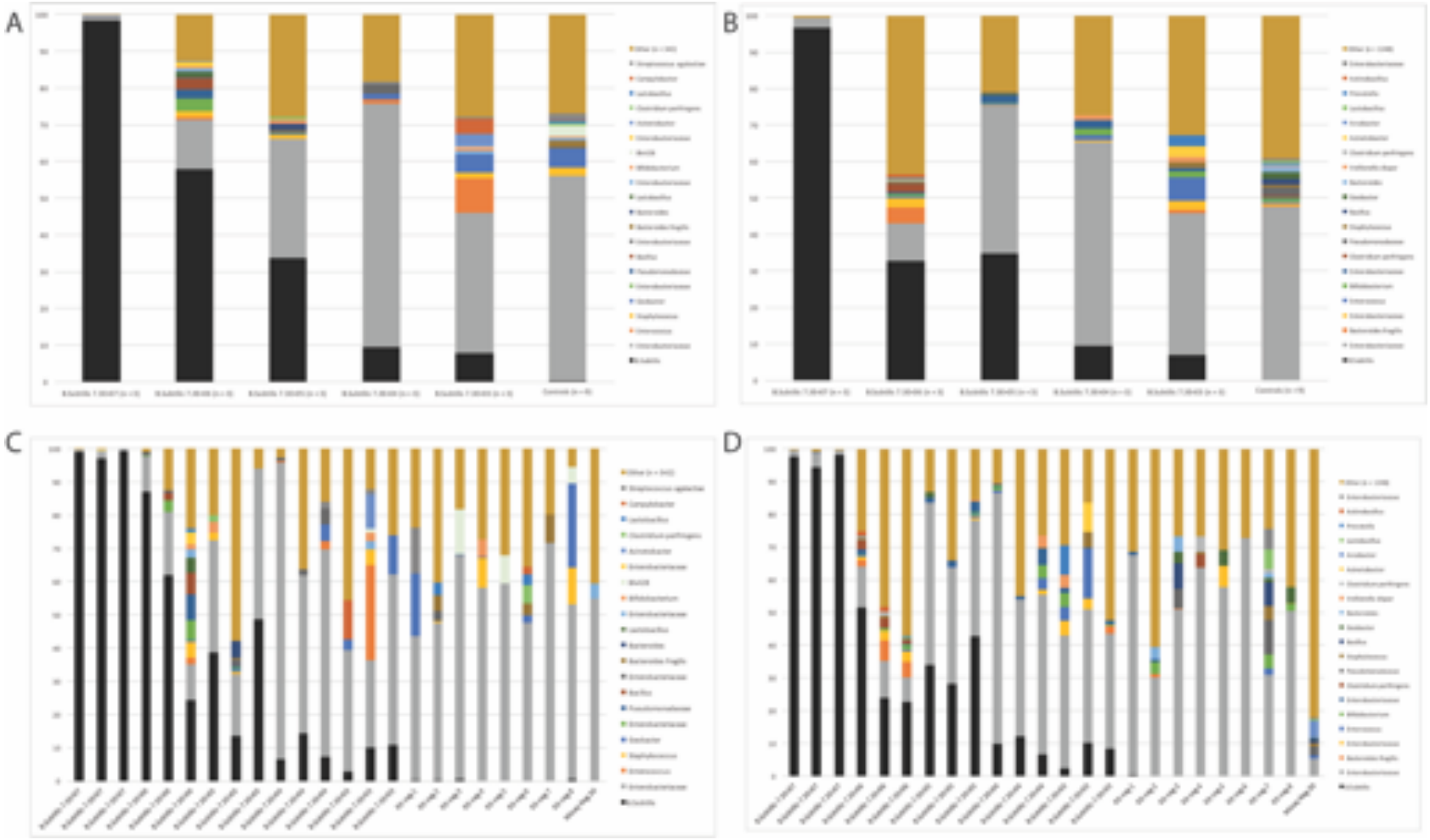


Figure 1.48. Stacked bar charts showing the relative abundances of the top 20 most abundant bacterial ASVs for *B. subtilis* dilutions and negative controls. A) 30 cycle PCR with sample replicates combined. B) 40 cycle PCR with sample replicates combined. C) 30 cycle PCR with individual replicates. D) 40 cycle PCR with individual replicates

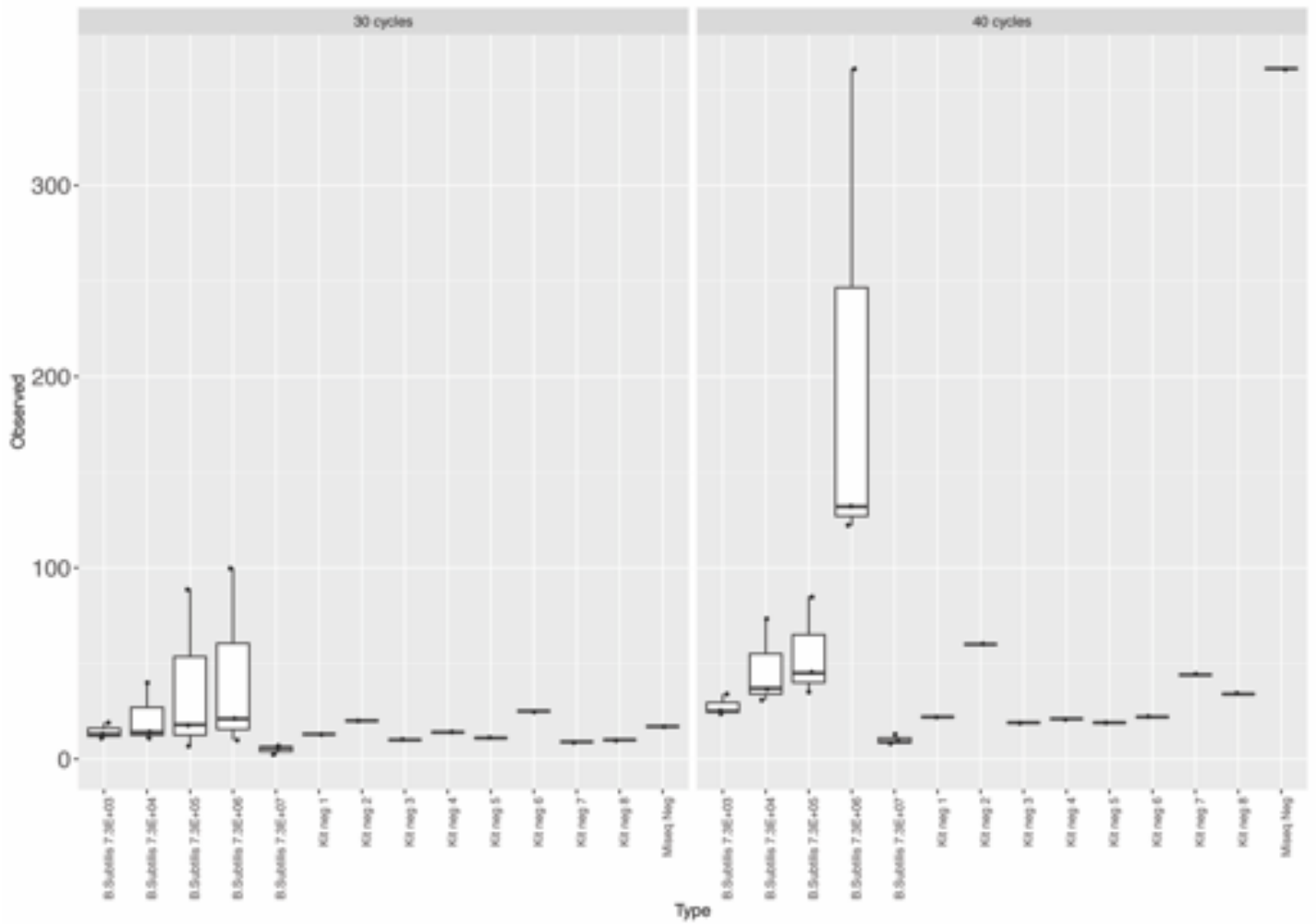


Figure 1.49. Box and whisker plot showing the observed ASV alpha diversity metric for samples and negative controls

30 cycle samples which had a *B. subtilis* relative abundance of ~50% and above clustered tightly, whilst the two samples which contained 20-40% *B. subtilis* clustered tightly away from the high abundance samples. The low abundance samples and kit negatives clustered from left to right, with a noticeable shift in the beta diversity of the communities in relation to the *Enterobacteriaceae* ASV, along with a number of lower abundance community members (figure 1.50). The pattern was generally similar for 40 cycle samples, however the second dilution clustered away from the highest concentration sample.

The addition of Arcticzyme to reagents prior to PCR amplification did not remove any contaminants, and instead appeared to reduce the relative abundance of the target sequence, suggesting technical inefficiencies both in the catalysis of the breakdown of contaminant DNA in reagents and in the inactivation of the enzyme.

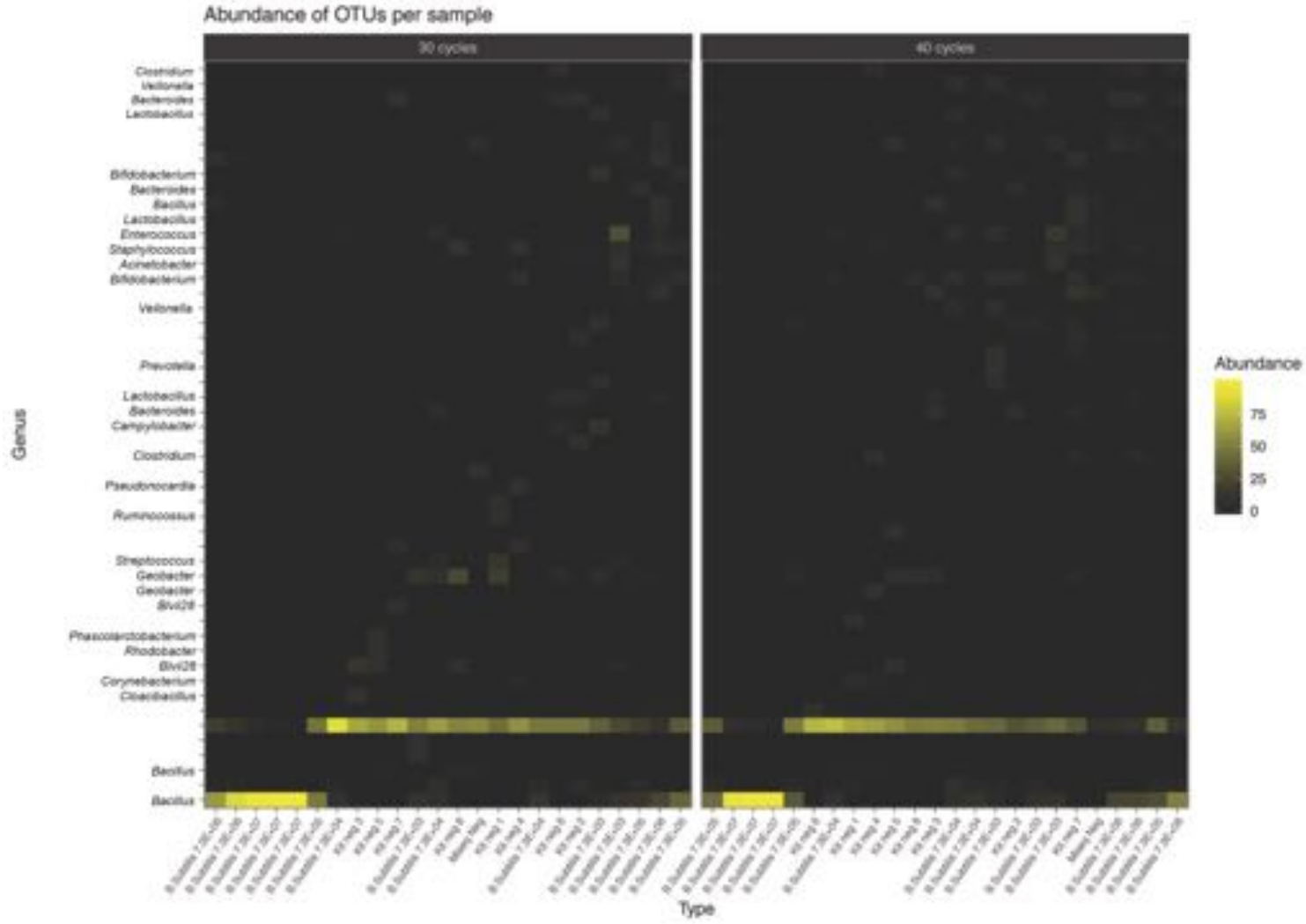


Figure 1.50. Heat map, organised by Bray-Curtis PCoA ordination, displaying the 50 most relatively abundant genera for each sample, faceted by number of PCR cycles

5.3.5 Impact of de-contaminating dataset on community profile

Following contaminant removal as per the method described in chapter 2.3.1, 30 cycle *B. subtilis* relative abundance in the highest concentration sample increased from 98.49% to 99.85%. Relative abundance in the second dilution increased from 57.81% 68.07%. The highest concentration 40 cycle sample target reads % rose from 96.79% to 99.58%. The largest increase following contaminant removal was from 34.90% to 63.70% in samples which had a starting concentration of *B. subtilis* 7.3E+05, which equated to an estimated CFU per mL^{-1} of <6.6E+05.

Looking at samples individually, the relative abundance of target reads increased from 48.78% to 89.93% for one of the 30 cycles 7.3E+05 samples (figure 1.51). The highest % increase in target sequence reads was seen in the lowest abundance samples (figure 1.52), however following this increase samples with 7.3E+04 and 7.3E+03 inputs had <30% target reads for both 30 and 40 cycle datasets.

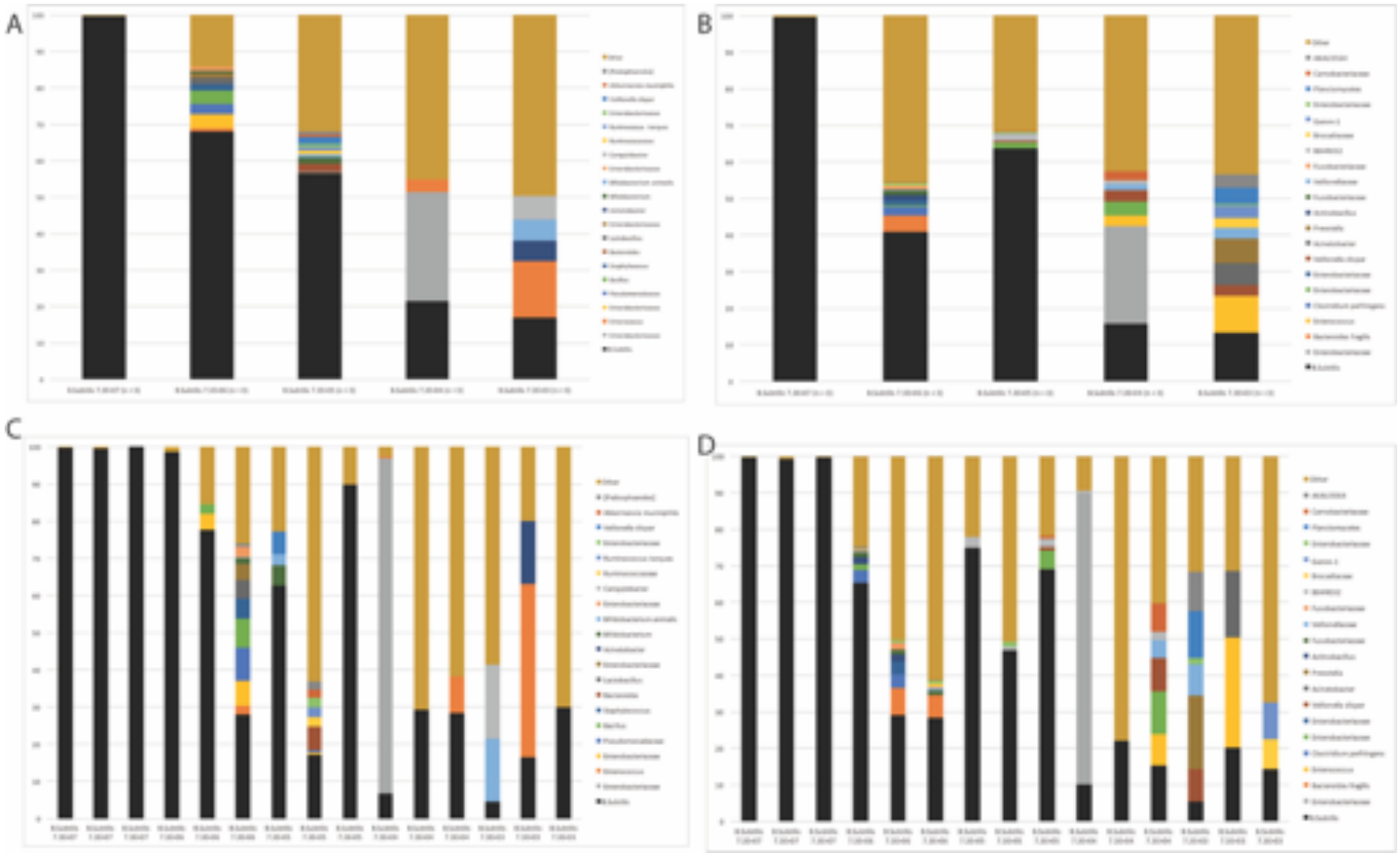


Figure 1.51. Stacked bar charts showing the relative abundances of the top 20 most abundant bacterial ASVs for *B. subtilis* dilutions following removal of ASVs present in relative abundances less than 50% higher in samples than in negative controls. A) 30 cycle PCR with sample replicates combined. B) 40 cycle PCR with sample replicates combined. C) 30 cycle PCR with individual replicates. D) 40 cycle PCR with individual replicates

5.3.6 Biomass of Antarctic air samples from chapter 4

Average 16S copy number for kit negative, marine and terrestrial samples collected during the ACE cruise (see chapter 4), was generated by 16S qPCR on 5 µls of DNA extract. Raw copy number was transformed to copy number per mL^{-1} by multiplying the values by 200. This value was then divided by the average number of 16S copies per bacterial cell of 4.2 (46), to give an estimated average CFU per mL^{-1}, or bacteria per quarter filter, for each environment (figure 1.53). There was no significant difference in copy number when performing pairwise Wilcoxon rank sum tests with Bonferroni correction between kit negative controls and marine samples (adj p = 1) or terrestrial samples (adj p = 1). There was also no significant difference between marine and terrestrial samples (adj p = 0.45).

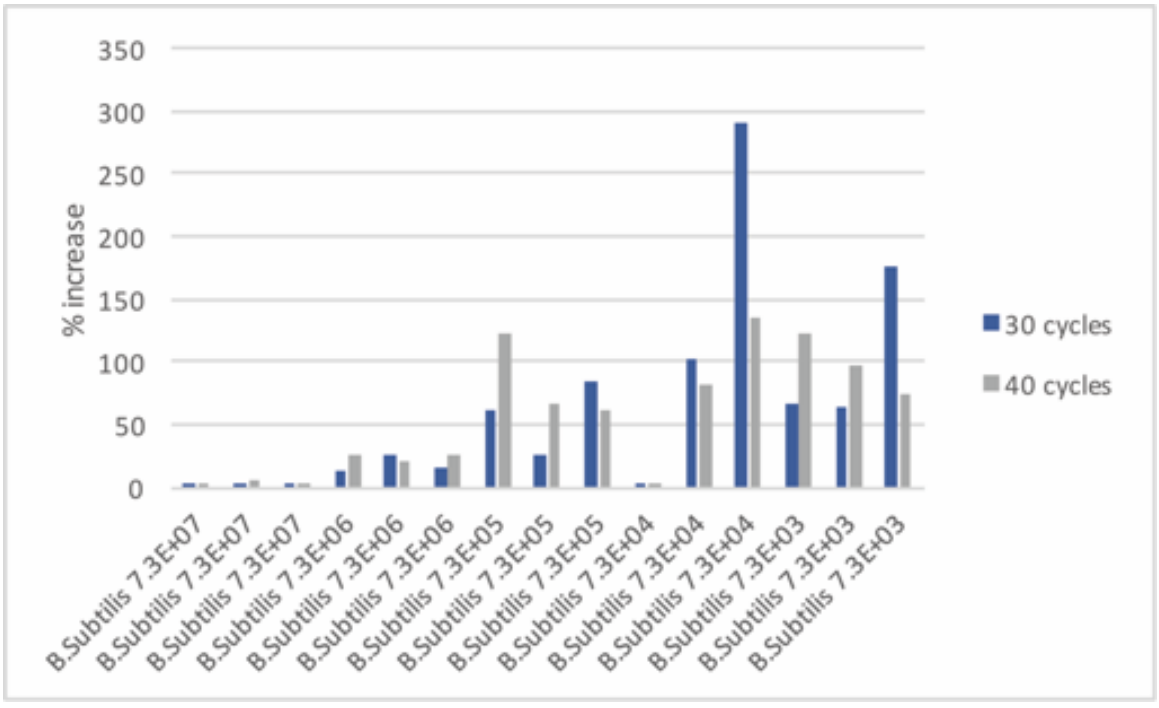


Figure 1.52. Clustered column chart showing % increase in target sequence reads per starting concentration following the removal of ASVs present in relative abundances less than 50% higher in samples than in negative controls

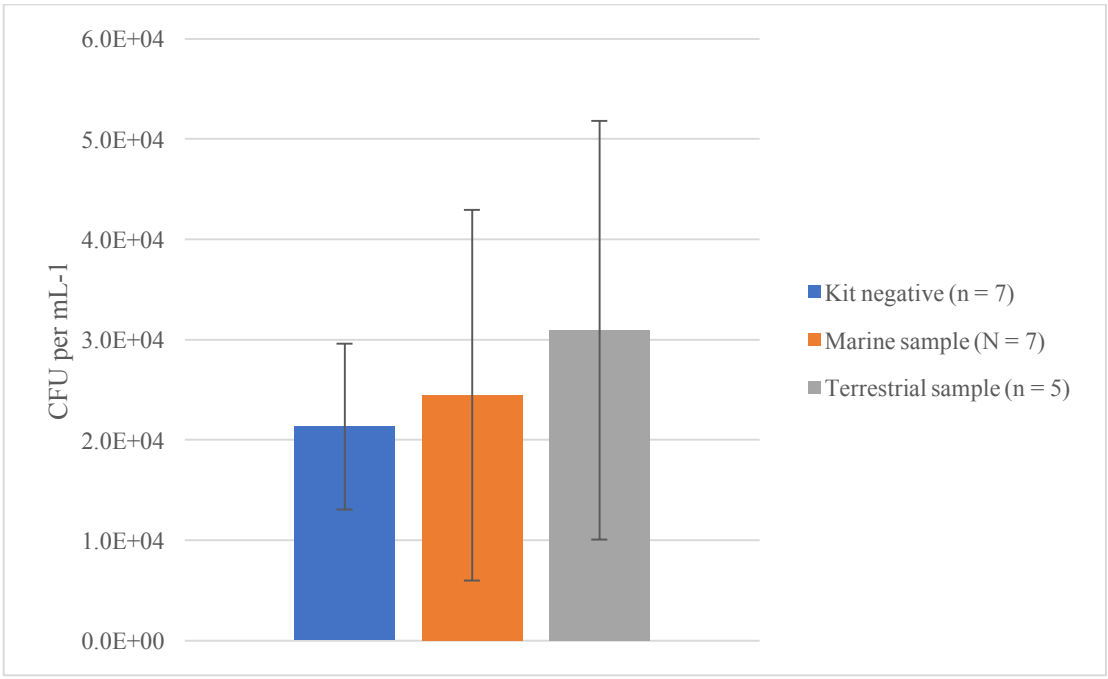


Figure 1.53. Bar plot showing the estimated mean CFU per mL^{-1} of samples collected during the ACE cruise

5.4 Discussion

It is known that choice of DNA extraction method influences community composition (236), and consequently it is important to consider the DNA extraction method of cross-comparable studies prior to undertaking an investigation. Whilst there are commercially available kits designed for a range of different sampling environments such as soil, water or stool, lesser investigated environments do not have specific kits. Qiagen kits are relatively similar, save their sample loading bead tubes. For example, the Qiagen Powerwater kit contains a large loading tube capable of containing a whole 47mm membrane filter, whilst the Qiagen Powersoil kit contains a small loading tube in order to fit 0.2g of soil inside.

Numerous aerobiological studies cite the Qiagen/MoBio Powersoil kit as their extraction method (25, 196, 251-256), despite the fact the bead tube is not designed to have a filter paper loaded. This is likely due to the ability to compare to previous studies, rather than the kits suitability to the sample form, as most aerobiological samples are stored either as liquid or on membrane filters. The original studies which chose to use this kit may have done so opportunistically, as they may have already had the kit available due to previous soil studies. Regardless, this kit remains the most consistently used in the literature for this type of sample. The majority of these studies dissect and load a $1cm^2$ punch of filter into the bead tube. This method of DNA extraction from membrane filters was not optimal, with measurable extraction efficiencies ranging from 3 to 16% of total input cells, this was suitable for samples with a starting bacterial load of 7.3E+07, as the concentration of DNA retained for extracts at this cell concentration was enough to provide a true community when sequenced, however below this input concentration, the method was not efficient enough to extract a high enough DNA concentration to reveal the true community using MiSeq sequencing.

Of the two concentration samples which could be quantified, DNA extraction efficiency appeared to increase with decreasing starting DNA concentration, indicating potential

saturation of the spin column. The amount of DNA recovered from each replicate was also different, which suggests the bacteria were not evenly distributed on the membrane filter, not taking this into consideration could cause an underestimation of biodiversity in a sample if a quarter of a filter was chosen which contained considerably less DNA that the rest of the filter.

Previous studies provide different information on the minimum input biomass required to reveal the true community of a sample. Our input DNA ranged from 7.3E+07, however due to extraction efficiency the true concentration of cells going into each PCR reaction, and therefore being sequenced, were around ten-fold less than this. This means that whilst we found that 7.3E+07 cells were required on a membrane filter to reveal a true community following DNA extraction, the number of cells required to be loaded into a PCR during library preparation to reveal the majority of the true community by MiSeq was actually in the region of 2-4E+06.

When the number of cells dropped an order of magnitude below this, each time the percentage of reads matching the true community dropped roughly 40-60%. Below E+03, less than 10% of the sequenced community were represented by target reads, with 90%+ of the sample constituents being exogenous contaminants. These numbers are consistent with the findings from samples sequenced at Imperial College London (ICL) by Salter, Cox (94). Found that all samples above an estimated cell input of 1E+04 CFU per mL^{-1} returned community profiles above 70% accuracy, whilst samples below this input threshold all represented <50% target sequences. Brandt and Albertsen (257) reported that 92% of target reads were represented in a sample containing a 1E+01 sample of S*higella*, however as *E. coli* was present at varying abundances in the controls for this study (9.9-76.7%). Whilst the authors used low read number and sample to control contamination to explain the prevalence of the taxa in negative controls, however, it is known that *Escherichia* and *Shigella* are hard to resolve using standard NGS technologies (258), over the short conserved fragment of 16S gene sequenced, could mean that their reported target sequence % is over inflated due to misinterpretation of data, however their

data may also be a true representation of their sequencing facility, as facility to facility variation is known (94).

Low concentration DNA may not be amplified until later cycles of PCR, as it is less likely to come into contact with a Taq polymerase enzyme. Increasing the number of PCR cycles is one method which could increase the yield of low abundance target DNA, however this can also increase the number of spurious reads through polymerase errors and further amplify contaminant DNA (241). Despite this, some studies have shown that increasing PCR cycle number does not compromise the integrity of NGS data (259). Reducing the number of PCR cycles is one potential option to reduce the amount of spurious sequences in a sample and increase the proportion of target DNA, however as shown by Salter, Cox (94), sequencing low biomass samples at 20 cycles means samples have very few reads, less than 50 in this case, which nullifies the biological relevance of the data.

Due to previously experiencing low read counts in low biomass samples sequenced at the facility, the difference in community profiles at different input biomasses using both 30 and 40 cycle PCR were compared. Both cycle numbers gave sufficient numbers of reads for analysis, and the % of reads hitting the target sequence was comparable at all input levels excluding the 3.7E+06 sample, where more than 20% reduction in sequence reads was seen at 40 cycles when compared with 30, which could possibly be attributed to pipetting error or inefficient amplification. Standard deviations were smaller for all 40 cycle samples as oppose to 30 cycle samples, this reduction in data variation is likely to all representative members of the sample community undergoing sufficient amplification due to the extended number of cycles.

Of the 38 non-target bacterial classes which were present, six bacterial classes were present in relative abundances >1% in the control samples and all but the highest *B. subtilis* sample (figure 1.48A); these classes were *Gammaproteobacteria*, *Bacilli*, *Clostridia*, Bacteroidia,

Actinobacteria, and Delta*proteobacteria*. These classes of bacteria were consistent with those reported by Salter, Cox (94), and as with the target sequence %, the class level patterns matched most closely with the contaminant profiles from samples processed at the ICL, despite use of a different DNA extraction kit. Despite similar bacterial classes between the studies, the relative abundance of the classes is noticeably different, with *Gammaproteobacteria* being by far the most relatively abundant taxa in our study, whilst *Actinobacteria* was the most abundant contaminant in that study, further emphasising the variation in contaminant profiles for samples sequenced at different facilities. This difference in contaminant profile at class level could be due to either differences in the wider lab contaminants at the different sequencing faculties or kit based contaminants between the Qiagen Powersoil kit used here and the FastDNA SPIN kit for soil used in that study, as choice of extraction kit is known to influence community structure (236).

Previous studies have shown a decrease in biomass to correlate directly with the observed OTU alpha diversity metric, showing the number of OTUs to increase as the biomass of a sample decreases (248), however whilst the lowest number of observed ASVs was in the highest concentration sample, this was not the pattern observed, as the number of ASVs remained consistently low in low biomass samples, including negative controls. An increased number of ASVs based on number of PCR cycles was observed, with 362 present at 30 cycles and 1268 at 40, showing that increasing the number of PCR cycles can inflate sample diversity. At ASV level, around 12-30% of the samples were made up of 'other' low abundance ASVs, a pattern again comparable with previous findings (94), however the profiles of the dominant contaminating ASVs were not as similar to previous studies, with an unclassified member of the Enterobactereiaceae family constituting a considerable proportion of non-target reads in all samples, with the remaining taxa much more sporadic than has previously been shown.

This pattern becomes only more complex when the sample replicates are viewed individually as oppose to stacked, at ASV level, only the unclassified member of the Enterobactereiaceae family was consistent in all samples, including the facility PCR negative, suggesting that the origin of this contaminating sequence was not during extraction but library preparation. The remaining top 20 most abundant ASVs were present sporadically, with *Enterococci, Staphylococci*, and *Geobacter* appearing the next most frequently observed. Enterococcus and Staphylococcus are both common human associated bacteria and so it is possible that this was their source, however this could also be an artefact of the fact that taxonomic databases were predominantly curated based on human associated taxa, whilst *Geobacter*, an anaerobic environmental genus, is difficult to pinpoint, and to the best of our knowledge has not previously been described in contaminant studies, suggesting this may be lab specific and related to work previously carried out in the lab where extraction was carried out.

The lack of consistency at ASV level in negative controls, as well as sample replicates, renders using standard contaminant removal methods like decontam (246), which relies on increased prevalence of a taxa in control samples when compared to true samples to identify contaminants, sub optimal for low biomass samples. As decontam was not appropriate for this dataset, nor was source tracker, a stringent *ad-hoc* contaminant screening was performed comparing taxa found in control samples to those found in true samples as described in chapter 2.3.1, in order to see if contaminant removal using negatives could improve the proportion of target sequences in each sample. This method proved effective for the high biomass samples, eradicating all but 0.15% of non-target reads from the samples, and whilst in general the proportion of non-target reads was reduced at all dilutions, by as much as 175% for one replicate at a starting concentration of 3.7E+04, the total % of the sample which was comprised of non-target reads still remained too high, meaning sample integrity and therefore data validity was still low.

Whilst contaminant screening was not effective enough to improve the majority of samples to a valid standard, one 3.7E+06 replicate went from ~75% to >98% target reads and one 3.7E+05 sample went from <70% to >90% target reads, showing that if the community profiles of the negative controls match closely to that of non-target taxa in a sample, sample integrity can greatly be increased, for this reason considering % change in community following contaminant removal could provide insight into how well de-contaminated a sample has been, however this must be viewed with regards to total read number.

As shown here and in previous studies (94, 248, 257), the amount of cells added to the PCR reaction during library preparation influences the proportion of sequencing reads which match the true community structure of the sample which has been extracted. Therefore, when working with samples expected to contain a low bacterial biomass, it is imperative that an estimate of bacterial concentration is attained prior to data analysis in order to ascertain the validity of the community structure. Previously, the microbiome of the Antarctic atmosphere has been investigated, however found considerable variation at ASV level in the data, in fact no genera were present in more than 40% of samples showing an unprecedented level of inter sample variability in the Antarctic atmosphere (152, 260), leading us to investigate whether the variation could be artificial due to technical issues (see chapter 4).

Using qPCR data, the CFU per mL^{-1} of the marine samples was estimated to be between the region of 2-3E+04 and terrestrial samples to be in the region of 3E+04; taking into account the data from this study for the same concentration range, this would suggest that this input biomass is insufficient to generate data showing a true community profile, and whilst negative controls were sequenced and decontamination using these controls was undertaken, the proportion of the decontaminated samples which could be counted as a true representation of the sampling environment could be between 18-90%. Whilst the biomass of Antarctic air is currently unknown, airborne bacterial concentrations are typically known to be in the range of

10^4 to 10^6 cells per m^3 (141); the higher range is typically found within the built environment where humans contribute considerably to the airborne biomass, how remote and harsh the Antarctic environment is likely puts the aerial biomass at the lower end of this scale. Therefore, by sampling for 12 hours using a membrane filtration apparatus as suggested by Pearce, Alekhina (151), a collecting an estimated volume of 14400L, equating to 14.4m^3 of air, an estimated total of 1.44E+05 to 1.44E+07 bacteria would be collected onto a membrane filter, and so ¼ of each filter may contain as little as 3.6E+03 to 3.6E+05 bacteria, considerably low the required input found here to derive a true community profile by MiSeq.

The Coriolis Micro (Bertin, Montigny-le-Bretonneux, France) (see chapter 2.1), which collects at a maximum flow rate of 300L/min for up to 1 hour, would comparatively collect 1.8E+05 to 1.8E+07 bacteria in liquid, dependent on the biomass of the environment. Whilst the biomass collected with this device would be similar to a sample taken over a longer period of time at a lower flow rate, the short sampling duration relies upon the assumption of a consistent bacterial load in the atmosphere at all times. The estimated bacterial yields for both sampling methods being below the required biomass input into a successful MiSeq run, explains the similar variability of the resultant datasets (Coriolis data not shown, see appendix IV).

5.5 Concluding remarks

The most commonly used, and therefore most cross-comparable method of DNA extraction bioaerosol samples stored on membrane filters was not optimal for low biomass samples, and as a result this method of DNA extraction for samples with a biomass lower than 3.7E+07 is not recommended. Furthermore, there was considerably more variation between kit negatives than has been previously described in studies, variable kit negatives reduce the resolving power between contaminating sequences and true sequences; this variability highlights the importance of sequencing multiple negative controls in order to reduce as much of the contaminating

signature in sample as possible, and the inappropriateness of particular contaminant removal procedures for low biomass data.

The findings of this study highlight the difficulty of using modern NGS technologies to investigate low biomass samples stored on membrane filters. Our results reiterate those of previous studies which found that 1E+06 was the lower limit at which Illumina MiSeq produced a reliable community profile independent of contaminant screening. The results of this investigation mean the first three hypotheses laid out within the introduction of the chapter must be rejected, whilst the hypothesis that the unprecedented sequence level biodiversity of Antarctic air samples was due in part to technical variation as a result of their low biomass, can be accepted.

When working with low biomass datasets, the following recommendations should be considered:

i) When using the Qiagen Powersoil kit (Hilden, Germany), for DNA extraction from low biomass samples on membrane filters, quantify as many samples as possible prior to sequencing

ii) Running a dilution series of a known control community within the expected sample concentration range alongside samples to assess the accuracy of the sequencing data

iii) Running as many sequencing negative controls as possible alongside samples in order to identify as many contaminating DNA sequences as possible

iv) Assessing the data validity of the data and choose an appropriate decontamination method for the dataset

Chapter 6 - Bioinformatics based variation in low biomass air sample analysis

6.1 Introduction

Culture-independent microbiome studies typically involve the use of the 16S rRNA marker gene to identify the community structure of a target environment, the analysis and interpretation of this datatype requires both molecular and computational tools (261). These are necessary to accurately characterise large datasets, however the most widely used of these profiling tools show a large degree of variability (262). Two of the most widely used bioinformatics tools are Mothur (263) and Quantitative Insights Into Microbial Ecology (QIIME) 2 (61). Mothur is a manually curated pipeline updated, optimised and validated by the Schloss lab whilst QIIME 2 represents a curation of the best and most current microbiome analysis tools. QIIME 2 places greater emphasis on data reproducibility than the previously version QIIME 1, and has now been shown to provide better sequence annotation (264). Few studies have compared the differences in datasets between the two frequently used pipelines (265), finding higher variability in the less common taxa, which impacted on beta diversity. To the best of our knowledge, no studies have compared the variability when working with low biomass samples.

The most common OTU picking strategy in recent years has been to combine sequences *de novo* into bins at a 97% pairwise similarity threshold (266); however, this method is inherently variable, as the representative OTU to which samples are binned varies each time data is put through the analysis pipeline. More recently, due to improvements in sequencing technologies and the improved ability of bioinformatics pipelines to identify and remove low quality reads, the use of exact sequence variants, referred to as amplicon sequence variants (ASVs) or sub-OTUs, has been suggested as the gold standard (67, 68); the main advantage constructing exact sequence variants over OTUs, is that the resultant datasets are reproducible, as sequences do

not have to be randomly binned. Naturally, assigning reads as exact sequence variants will increase the number of unique taxa which are represented in a dataset. Closed reference OTU picking is the least frequently reported OTU picking method, likely due to its limited application; closed reference OTU picking involves assigning OTUs based on an already curated reference taxonomy database and discarding unmatched sequences; this method is a necessity when comparing multiple hypervariable regions, and can also be useful if the study environment is well described and you trust the reference database to be comprehensive, however in poorly described environments, this method of OTU picking has the potential to considerably waste valuable data. Choice of taxonomic database is also known to impart a level of variation on datasets, and can impact results to some extent (75). Greengenes (71) and the Ribosomal Database Project (RDP) (70) are two of the more commonly used reference databases. QIIME 2 uses Greengenes as its default classifier, however this database has not been updated since 2013, whilst Mothur recommends RDP, with the most recent version included in the Mothur MiSeq sop being version 16 which was curated in 2016.

The aim of this study was to compare the imparted variance in bacterial community composition of low biomass Antarctic air samples when analysed using Mothur and QIIME 2. Using these two pipelines, the difference in the resultant microbiome based on choice of taxonomic database and OTU picking strategy was also investigated. The following hypotheses were considered:

i) Choice of analysis pipeline does not impact on the biodiversity of a dataset

ii) Choice of OTU picking strategy does not impact on the biodiversity of a dataset

iii) Choice of taxonomic database does not impact on the biodiversity of a dataset

6.2 Methodologies

6.2.1 Sample processing and analysis

Bacterial DNA was extracted from samples collected during the ACE cruise (chapter 4). Fastq files were generated by sequencing the V4 region of the bacterial 16S rRNA gene by Illumina MiSeq (chapter 2.2.8). Fastq files were processed into an OTU and taxonomy table using one of 4 variations of the standard QIIME 2 (61) or Mothur (263) pipelines (see below), then screened for contaminants in Microsoft excel (2013) as described in chapter 2.3.1. The Phyloseq (171), Ape (267), Vegan (249), and GGplot2 (250) packages in R, as well as MS Excel (2013) were used to carry out diversity and statistical analyses and to produce graphics.

The following 4 pipeline alterations were compared:

A) Samples were processed as per the Mothur MiSeq SOP (https://Mothur.Mothur.org/wiki/MiSeq_SOP), binning OTUs *de novo*, and assigning taxonomy using the RDP 2013 taxonomic database. The SOP first clusters sequences, and then assigns taxonomy to clusters.

B) Samples were processed as per the Mothur MiSeq SOP (https://Mothur.Mothur.org/wiki/MiSeq_SOP), binning OTUs *de novo*, and assigning taxonomy using the RDP 2016 taxonomic database. The SOP first clusters sequences, and then assigns taxonomy to clusters.

C) Samples were processed in Mothur, using a phylotypic (closed reference) OTU assignment approach. Sequences were classified using the RDP 2016 taxonomic database first and then clustered into OTUs based on taxonomic classification.

D) Samples were processed in QIIME 2, as described in chapter ?. Unique sequences are stored as amplicon sequence variants (ASVs) as oppose to OTUs, and ASVs were classified using the Greengenes 2013 taxonomic database.

6.3 Results

6.3.1 Raw reads and sampling depth

Of the 84 total samples collected by membrane filtration, 75 were retained for samples processed using the QIIME 2 analysis pipeline with the Greengenes 2013 taxonomic database following denoising and decontamination, equating to 89% of samples (figure 1.55C), all of these samples reached asymptote by 1,000 reads (figure 1.54D). This pipeline retained the highest number of reads at 1,061,145 (figure 1.55A), as well as the highest mean reads per sample with 14,148 (figure 1.55B). The closed reference pipeline in Mothur using the RDP 2016 taxonomic database provided the lowest sample retention, with 56 samples (figure 1.55C), along with the lowest total reads and mean reads, which were 628,890 and 11,230 respectively (Figure 1.55A & 1.55B). Comparing the RDP 2013 and RDP 2016 database for taxonomic assignment, the number of samples retained was identical with 60, whilst total and mean reads were near identical (figure 1.55). All amplified samples processed using the RDP taxonomic database and the Mothur pipeline reached asymptote on their relevant rarefaction curve (figure 1.54).

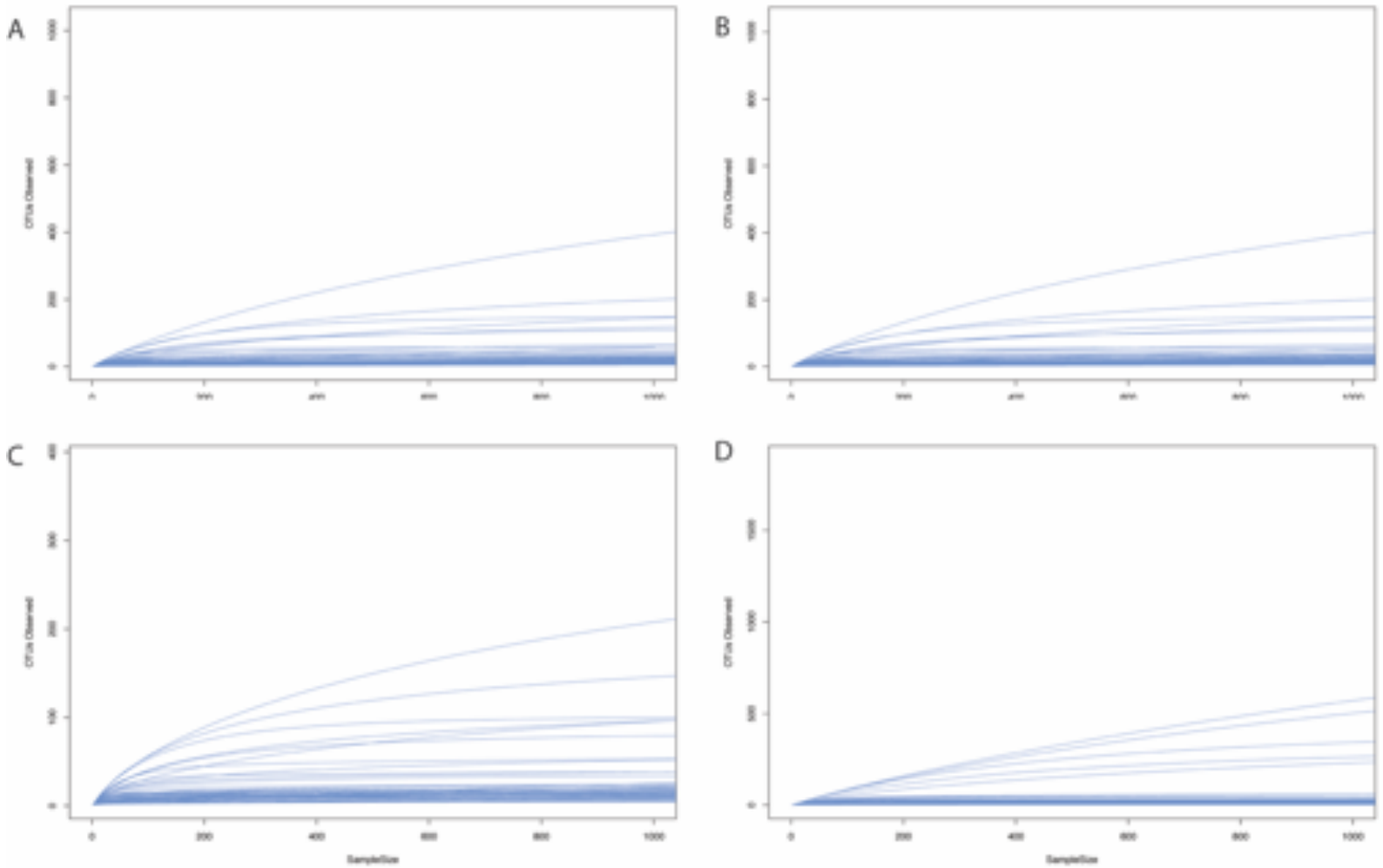


Figure 1.54. Rarefaction curves for samples processed using A) Mothur RDP 2013 B) Mothur RDP 2016 C) Mothur RDP 2016 closed reference OTU picking D) QIIME 2 ASV Greengenes 2013

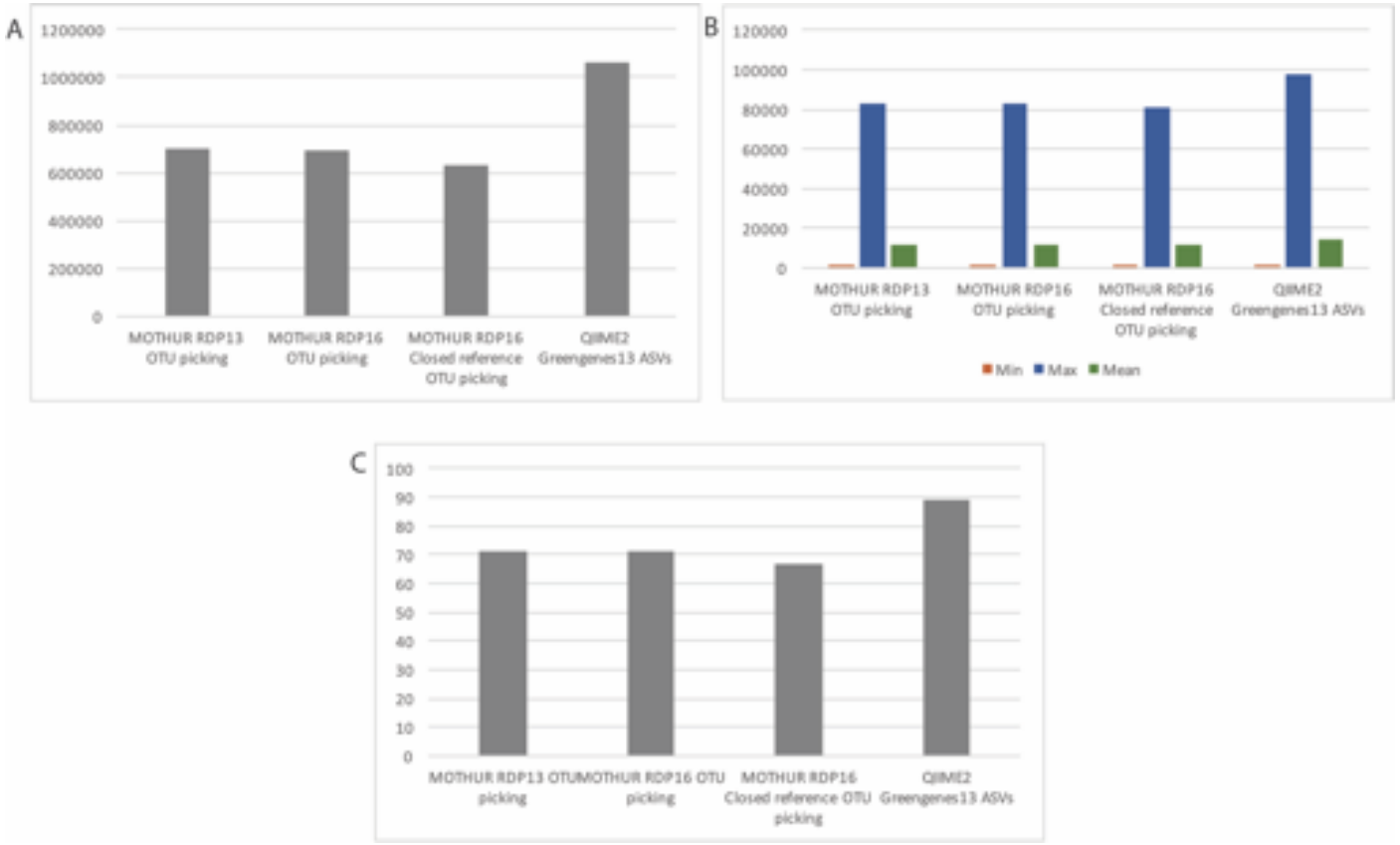


Figure 1.55. Bar charts showing A) Total sequence reads retained, B) mean, minimum, maximum total sequence reads retained for samples, and C) the % of samples retained

6.3.2 Alpha and beta diversity

Alpha diversity measures were markedly similar independent of analysis pipeline. Using Mothur with the RDP 2016 taxonomic database provided the highest number of observable OTUs (figure 1.56) with a median value of 28. This was closely followed by samples processed in Mothur using the RDP 2013 taxonomic database which had a median observed OTUs value of 27. Using the RDP 2016 database but a closed reference OTU assigning strategy reduced the median observable OTUs down to 21. Whilst the QIIME 2 pipeline assigning ASVs instead of OTUs, whilst using the Greengenes 2013 taxonomy provided the lowest median number of observable OTUs with 20. There was no significant difference between the mean number of observable OTUs based on analysis pipeline when tested using the non-parametric Kruskall-Wallis test at a significance level of 0.05 ($p=0.75$). There was a more pronounced difference in alpha diversity when viewing the Shannon index values which is more influenced by sample evenness (figure 1.57). The QIIME 2 pipeline had the highest median Shannon diversity value of 2.13, suggesting that this pipeline provides the most even community. The standard Mothur pipelines using the RDP 2013 and RDP 2016 taxonomies both had the same median Shannon diversity value of 1.97. The closed reference Mothur pipeline had the lowest Shannon diversity index value of 1.84 showing it to have the least evenly distributed communities. There was no significant difference between the mean Shannon diversity index values based on analysis pipeline when tested using the non-parametric Kruskall-Wallis test at a significance level of 0.05 ($p=0.17$). The amount of variation in community dissimilarity varied based upon analysis pipeline (figure 1.58). The first two principle components described very little of the variation in each dataset for all 4 pipelines with 10.2% of the variance described for the Mothur RDP 2013 and Mothur RDP 2016 pipelines, 11% for the Mothur closed reference RDP 2016 pipeline, and only 7.7% for the QIIME 2 Greengenes 2013 pipeline. Sample clustering was almost identical for the RDP 2013 and 2016 pipelines, whilst samples clustered considerably

less tightly when processed using the closed reference pipeline. Samples processed in QIIME 2 clustered in a similar pattern to those processed in Mothur using RDP 2013 and 2016, however the clustering was less tight

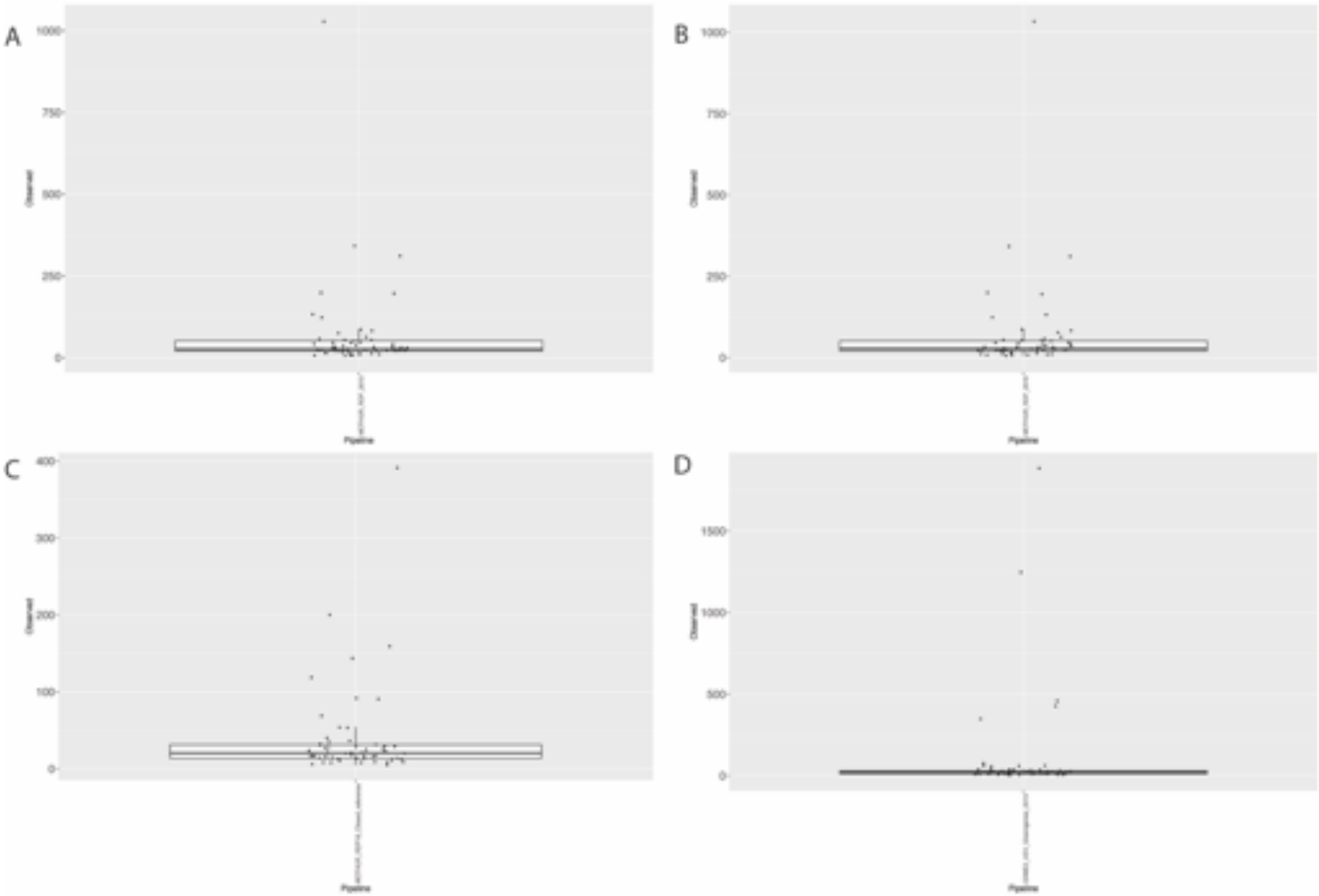


Figure 1.56. Observed OTUs alpha diversity metric for membrane filtration samples processed using A) Mothur RDP 2013 B) Mothur RDP 2016 C) Mothur RDP 2016 closed reference OTU picking D) QIIME 2 ASV Greengenes 2013

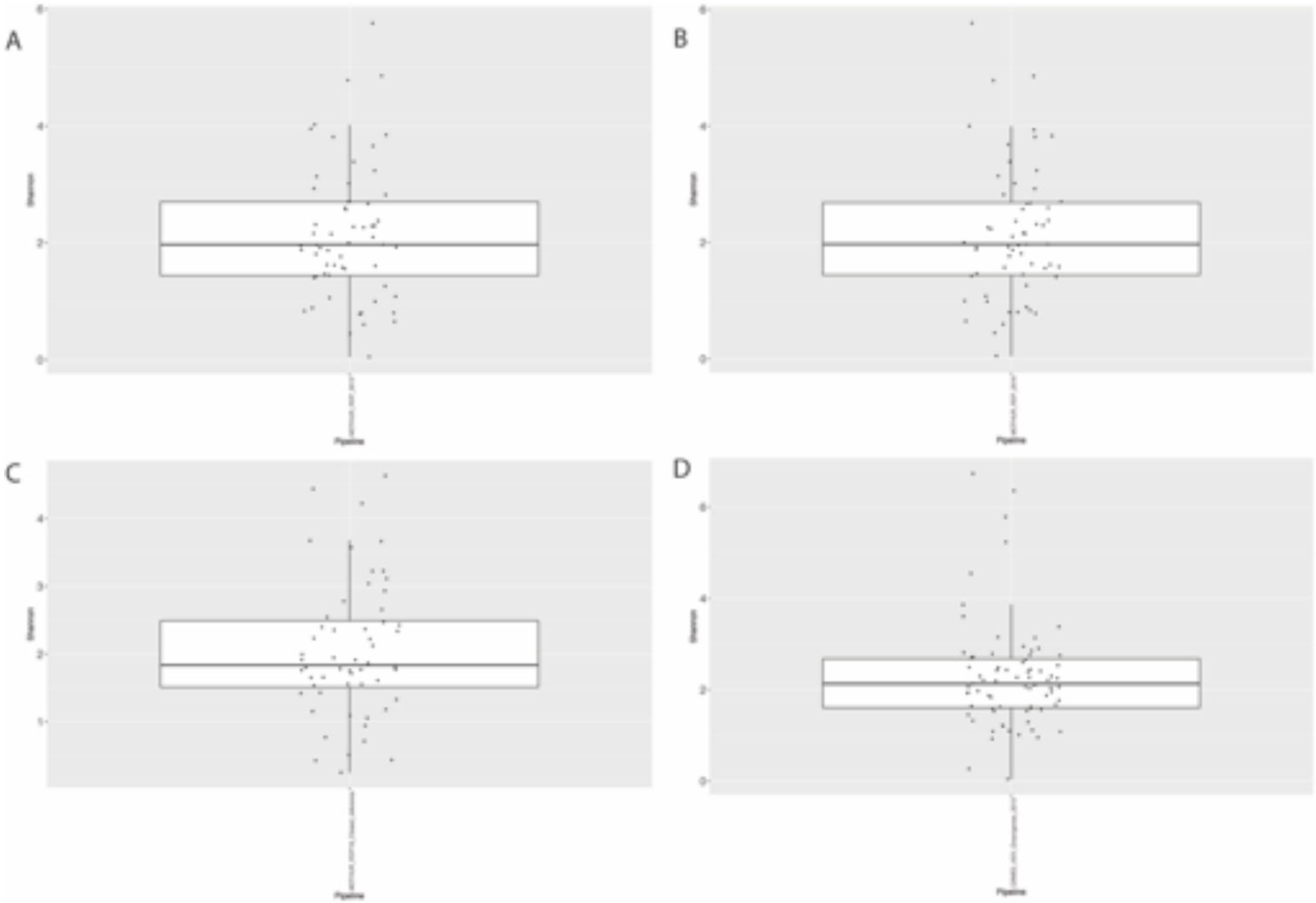


Figure 1.57. Shannon index alpha diversity metric for membrane filtration samples processed using A) Mothur RDP 2013 B) Mothur RDP 2016 C) Mothur RDP 2016 closed reference OTU picking D) QIIME 2 ASV Greengenes 2013

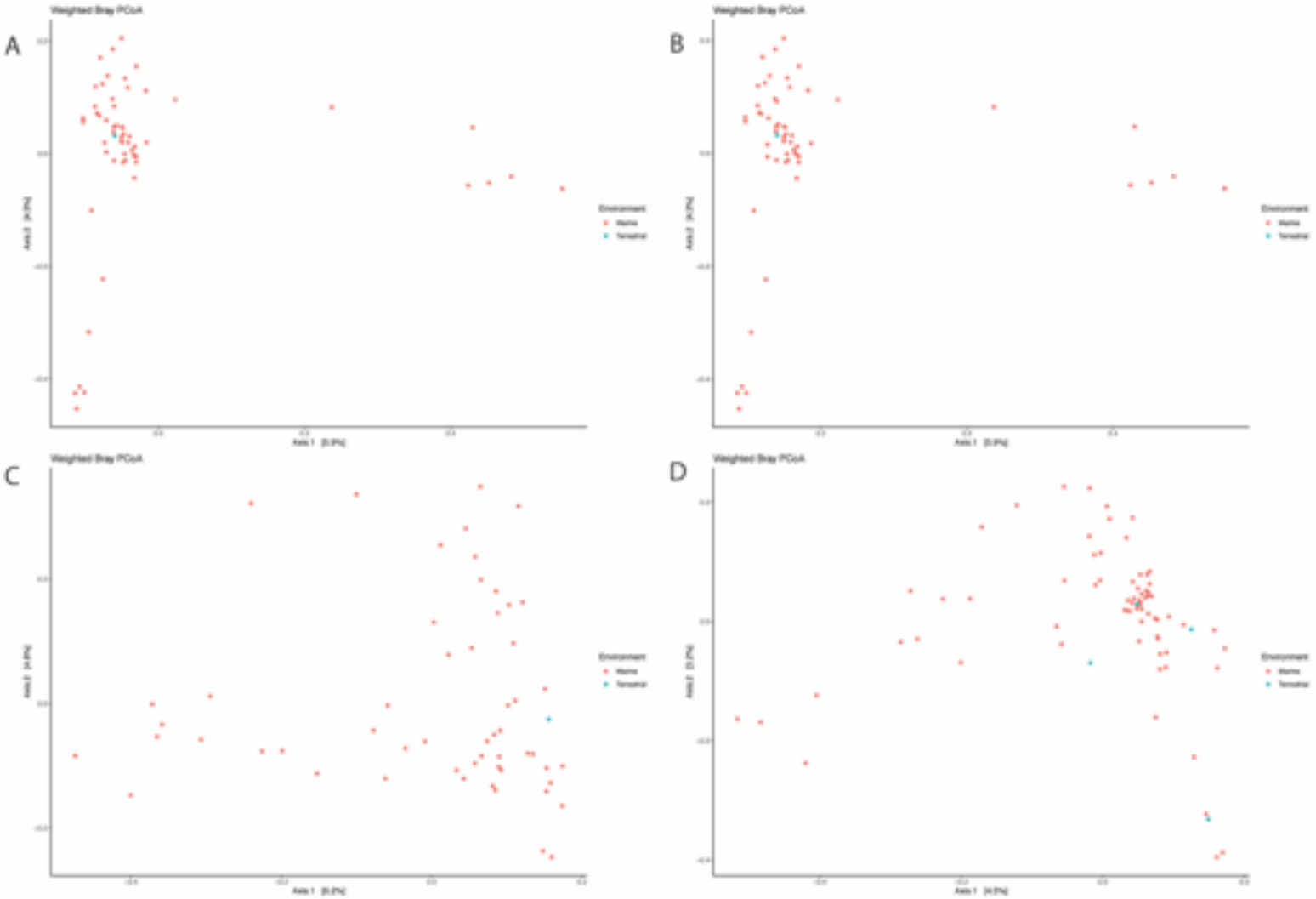


Figure 1.58. PCoA showing Bray-Curtis beta diversity for membrane filtration samples processed using A) Mothur RDP 2013 B) Mothur RDP 2016 C) Mothur RDP 2016 closed reference OTU picking D) QIIME 2 ASV Greengenes 2013. Samples coloured by marine or terrestria

6.3.3 Taxonomy assignment

The number of taxa produced by each method varied. The QIIME 2 pipeline, which grouped reads as ASVs at a 100% similarity threshold produced the most individual taxa as a result with 4456, whilst the closed reference 97% OTU picking pipeline processed in Mothur which discarded any reads not matched to a previously existing sequence in the RDP 2016 database produced the fewest unique taxa with 816. The standard Mothur pipeline which combined open reference and *de novo* OTU picking strategies, produced a very similar number of unique taxa when using either the older RDP 2013 taxonomy or more recent RDP 2016 taxonomy, with only 5 more taxa identified using the more recent database. Despite producing the most unique taxa, Greengenes 2013 identified the lowest % of taxa at genus level of the compared methods with 59.61% (table 1.14). Despite revealing a lower proportion of the total taxa at genus level, this pipeline still identified 2656 unique genera, the most of any of the pipelines. The RDP 2016 closed reference pipeline identified the highest % of samples to genus level. At class level, the QIIME 2 pipeline performed best, identifying 99.78% of all unique taxa, considerably higher than the two standard RDP based pipelines, whilst the closed reference RDP pipeline performed a close second best identifying 98.87% of retained reads at class level.

The top 10 most abundant taxa identified for the RDP 2013 and RDP 2016 Mothur pipelines shared similar taxonomic profiles, with the only difference that an unclassified member of the *Firmicute* phylum was present in samples classified using RDP 2013 and not in samples classified using RDP 2016, where it had been replaced by *Prevotella* (figure 1.59). When using closed reference OTU picking, the profile of the 10 most abundant taxa was markedly different in samples when compared to the RDP 2016 pipeline, despite the same taxonomic database being used to classify OTUs, with only *Acinetobacter, Aeromonas, Arcobacter,* and *Pseudoalteromonas* shared between the two sample sets. The different naming strategies of the RDP and Greengenes taxonomies makes the taxa difficult to compare between the two,

however some taxa are clearly present using both methods, as an ASV belonging to the *Aeromonadaceae* family which was not identified at Genus level was present at a relative abundance of near 10%, whilst *Aeromonas,* a member of the *Aeromonadaceae* family were present in all 3 RDP based datasets*. Despite the differences in taxonomy, half of the top 10 most abundant taxa produced in QIIME 2 using the Greengenes 2013 taxonomy match those produced in Mothur using the RDP 2013 and RDP 2016 taxonomies, with *Arcobacter, Acinetobacter, Pseudoalteromonas, Aeromonas**, and an unclassified member of the *Bacteroidales* order shared between the taxonomic profiles of the 3 analysis pipelines. The unclassified member of the *Bacteroidales* order was the most relatively abundant member of the top 10 most abundant taxa for the Mothur RDP 2013 and RDP 2016 pipelines, *Bacteroides* was the most relatively abundant taxa for the closed reference pipeline, whilst *Psychrobacter* was the most relatively abundant taxa for the QIIME 2 pipeline.

Level	Mothur RDP13 OTU picking	Mothur RDP16 OTU picking	Mothur RDP16 Closed reference OTU picking	QIIME 2 Greengenes 2013 ASVs
Unique taxa	2493	2488	816	4456
Classes identified	86.93%	86.89%	98.87%	99.78%
Genera identified	63.85%	63.74%	88.89%	59.61%

Table 1.14. Total number of unique taxa per analysis pipeline, and the proportion of those taxa which were identified at class and/or genus level

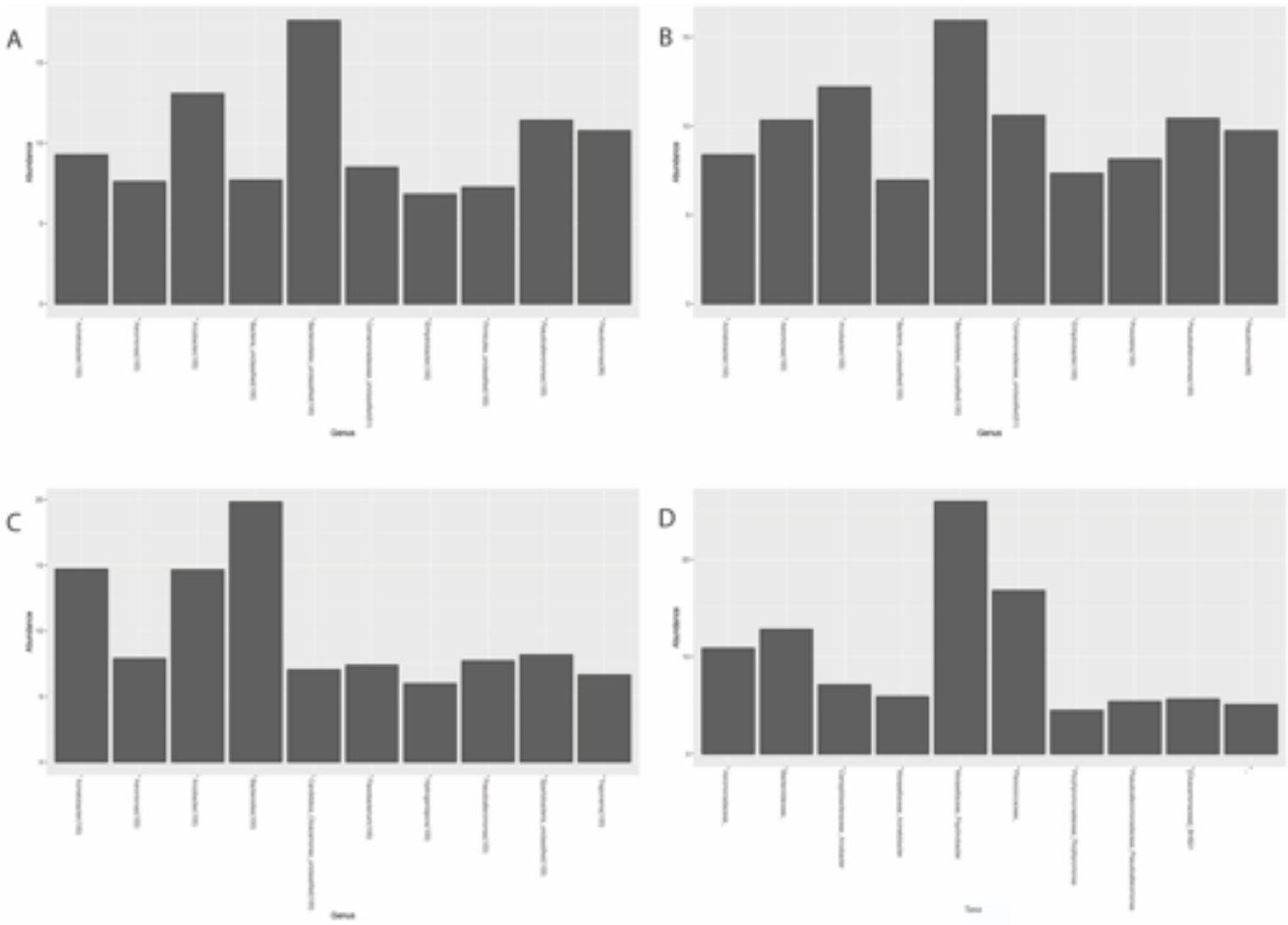


Figure 1.59. Bar plots showing the relative abundance of the top 10 most abundant taxa in all samples for membrane filtration samples processed using A) Mothur RDP 2013 B) Mothur RDP 2016 C) Mothur RDP 2016 closed reference OTU picking D) QIIME 2 ASV Greengenes 2013

6.4 Discussion

Choice of bioinformatics pipeline, OTU/ASV assignment strategy, and taxonomic database had an impact on the interpretation of a low biomass dataset. This supports previous assertions that different bioinformatics tools impart variability when analysing 16S NGS data (262). The primary aim of all microbiome studies is to collect and analyse a full dataset containing as little error as possible. The number of samples retained following the correction of sequencing errors, removal of chimeric sequences, removal of non-target sequences and removal of samples below 1000 reads, was highest for the QIIME 2 pipeline, with 75 of 84 samples retained, whilst the Mothur based pipelines retained 56-60 samples.

The difficulties amplifying low biomass samples, mean some degree of sample loss is expected due to low read number. As sample loss is already expected, it is important that the method by which samples are denoised or quality filtered, is highly efficient, in order to retain as many successfully sequenced samples as possible, therefore, when working with low biomass air samples, QIIME 2, which wraps DADA2 for sample denoising, was the most appropriate pipeline.

There was no significant difference in the richness of samples between the 4 compared pipelines. A significantly larger sample richness (total number of genera) has previously been reported when comparing QIIME to Mothur, however this study compared QIIME version 1, which still processed samples into OTUs and also compared preterm gut samples (268), as oppose to QIIME 2 and low biomass air samples. There was a lower richness when comparing QIIME 2 to Mothur, with the Mothur pipeline having a median value of 28 when compared to that of 20 when using QIIME 2. Another study, comparing bacterial communities in dairy cow rumen found no significant difference in richness when comparing the two pipelines, however this study used the SILVA database as oppose to Greengenes, citing the fact SILVA had been more recently updated than Greengenes as a potential reason for the lack of a significant

difference (265). Here, the richness of samples processed using the same pipeline (Mothur), but a 2013 vs 2016 RDP taxonomic database, was highly similar, with median observable OTUs being 27 and 28 respectively, suggesting that how recently a taxonomic database has been updated has a small impact on community richness.

Despite having the lowest observed OTUs value, the samples processed through the QIIME 2 pipeline had the highest Shannon diversity index, which acts as a proxy for sample evenness. There was no significant difference between the community evenness of samples processed using any of the four pipelines suggesting choice of pipeline, including choice of taxonomic database and OTU assigning method does not impact this community feature for low biomass air samples. The choice of analysis pipeline had an impact on how tightly samples clustered and the amount of variation between communities; samples were noticeably more dissimilar when processed in QIIME 2 than when processed in Mothur using *de novo* clustering, with less of the total sample dissimilarity described by two axis and less tight clusters, potentially due to the fact sequences were clustered at 100% as oppose to 97%. Despite some difference in how closely samples clustered, the core cluster of the *de novo* Mothur samples and QIIME 2 samples were similar sized. The closed reference pipeline had the least tight clustering, and no obvious core cluster, suggesting the taxa responsible for the similarities of the QIIME 2 and *de novo* Mothur samples were not part of the RDP 2016 taxonomic database, emphasising the limitations of using closed reference OTU picking for poorly described environments.

Plummer, Twin (268) found 10.27% and 28.92% of reads were unclassified at genus level when comparing QIIME and Mothur respectively, for preterm gut samples. 36% of sequences processed using Mothur and 40% processed using QIIME 2 were unclassified at genus level. This reduction in genus level taxonomic identification could be due in part to the disparity in research volume between the two environments, with many more studies focusing on the preterm gut microbiota meaning the taxonomy databases are biased towards bacteria with

clinical relevance than environmental. The high discrepancy in taxa identified at genus level between samples processed by closed reference OTU picking and *de novo* OTU picking in Mothur, followed by taxonomic assignment with the RDP 2016 database, as well as the lower number of unique taxa and total reads, suggests that a high proportion of diversity is being discarded by closed reference OTU picking due to the lack of representative taxa from the sampled environment in the RDP 2016 database, emphasising the unsuitability of closed reference OTU picking for low biomass air samples collected in the Antarctic; this data also explains the differences in sample beta diversity between the two pipelines.

Class level analyses are often carried out in poorly annotated environments, due to the lower number of genus level calls making it difficult to make meaningful biological statements at that taxonomic rank. QIIME 2 and Greengenes performed best at identifying taxa at class level, identifying 99.78% off all unique taxa at this rank. Comparatively, the *de novo* methods in Mothur using either RDP taxonomic database did not surpass 87% identification at class level, showing QIIME 2 to perform better when investigating class level patterns in low biomass Antarctic air samples. Whilst there were some consistent taxa present independent of pipeline, the most relatively abundant taxa were different for ASV, *de novo* OTU, and closed reference OTU pipelines. Overall there was a noticeable difference in the top 10 most abundant taxa for all air samples when using different bioinformatics pipelines.

6.5 Concluding remarks

When working with low biomass air samples, choice of bioinformatics pipeline can considerably impact the number of samples discarded prior to analysis following quality filtering/denoising, with QIIME 2 the least severe, meaning the hypothesis that choice of pipeline does not influence the biodiversity of a dataset must be rejected. All sample amplicon libraries were sampled at sufficient depth independent of pipeline. There was no significant difference in the alpha diversity of samples when processed using any of the four pipelines,

specifically with regards to sample richness and evenness; notably, this remained the case when comparing a previous version of the RDP taxonomic database to the most recently updated, the hypothesis that choice of taxonomic database does not impact the biodiversity of a dataset can therefore be rejected. Sample beta diversity was most impacted by choice of OTU picking/ASV assignment strategy, with closed reference OTU picking increasing the dissimilarity between samples, due to the removal taxa unclassified in the RDP 2016 database; as a result, the hypothesis that OTU picking strategy does not impact the biodiversity of a dataset is rejected. The QIIME 2 pipeline performed best when annotating samples at class level. The large discrepancy between the % of taxa identified at genus level by closed reference OTU picking when compared with *de novo* or ASVs, emphasises the lack of representative sequences in the RDP and Greengenes taxonomic databases for the sampled environment.

Chapter 7 - Discussion and recommendations

Aerial bacterial communities are understudied in both the Arctic and Antarctic. This study began by characterising the biodiversity of Arctic bacterial bioaerosols on Svalbard. Bacteria were found to be ubiquitous and their communities to be homogeneous. Prior to this study, Arctic air bacterial communities had only been investigated by a single molecular study, in which samples were collected in the Canadian high Arctic (143); that study found fewer phyla than described here on Svalbard (chapter 3), but shared the *Cytophagales*, *Lactobacillus*, *Staphylococcus*, *Janthinobacterium*, *Pseudomonas* and *Polaromonas* genera. Following the publication of the Svalbard work, one further study of bioaerosols in the Arctic has been undertaken in Greenland (269), widening the range at which bioaerosol communities have been described in the Arctic. Bacterial communities above Greenland were dominated by the same phyla as on Svalbard and in the Canadian high Arctic; these are *Proteobacteria*, *Firmicutes*, *Bacteroidetes*, *Actinobacteria*, *Acidobacteria* and *Cyanobacteria*. The results of this work point towards the potential existence of an indigenous community of bioaerosols above the Arctic, however future studies over a larger spatial and temporal range are required to confirm whether this is truly the case.

Bacterial DNA was found to be ubiquitous in the air surrounding the Antarctic, as in the Arctic. *Acidovorax, Acinetobacter, Cloacibacterium, Pseudomonas* and *Sphingomonas* were present in samples collected on Svalbard, and have also been described in the Antarctic at Halley station (153), suggesting the potential for long range atmospheric transport and region specific biogeography. *Acinetobacter, Cloacibacterium, Pseudomonas,* and *Spingomonas* were present in at least one of the samples collected above the oceans surrounding the Antarctic, adding weight to unique bi-polar airborne bacterial communities, however *Acidovorax* were not present in any of the samples collected during the ACE cruise (chapter 5). There were 45 phyla detected across the 75 samples collected around Antarctica, considerably more than the 12

which were detected in the Arctic; *Proteobacteria*, *Firmicutes*, *Bacteroidetes*, *Actinobacteria*, and *Acidobacteria*, the dominant phyla present in the Arctic samples were present in overall high abundance in the Antarctic samples, however at much more variable relative abundances. Additionally, the classes which primarily constitute the *Proteobacteria* phyla in Arctic air are *Alpha-*, *Beta-*, and *Gamma- Proteobacteria*, which was reflected in the class level constitution of Antarctic *Proteobacteria* too.

A core microbiome was present at class level around the Antarctic, consisting of *Actinobacteria*, *Gammaproteobacteria*, and *Clostridia*, classes which are ubiquitous in Antarctic environments, however when investigated at sequence level, a core microbiome was not present, and no taxa were present in more than 40% of samples, a more extreme level of diversity than has previously been described in the Antarctic atmosphere (152, 260). There was no significant difference in the alpha or beta diversity of communities above the oceans surrounding the Antarctic, independent of whether samples were collected above or below the latitude of the polar vortex, if samples were collected at sea or on land, or when comparing samples based on environmental variables.

The results of the investigation show bacterial DNA to be ubiquitous in the atmosphere in both the Arctic and Antarctic, however whilst bacterial communities were fairly homogeneous in the Arctic, extreme sequence level diversity was observed in the air surrounding the Antarctic. Therefore, the hypothesis that bacterial communities residing in the air in the polar regions are ubiquitous can be accepted, however the hypothesis that these communities harbour homogeneous patterns of biodiversity due to the extreme selectivity and remoteness of the environment must be rejected.

Precipitation samples were highly homogenous, and dominated by *Proteobacteria*, largely of the *Betaproteobacteria* class, a pattern shared with rainwater collected from temperate regions

(229). Rainwater samples were higher biomass than air samples and as such could provide a significant input of bacteria into the Antarctic via long range atmospheric dispersal. This sampling regime was the first of its kind at this scale, and the first study to sample bioaerosols on certain sub-Antarctic islands. As such, further studies investigating the bacterial communities within the air surrounding the Antarctic must be carried out to confirm the validity of the findings of this study. Additionally, comparing the bioaerosol sample profiles collected during the cruise with seawater and soil samples attained by other research groups could provide a greater insight into the relationship between the aerial environment and the adjoining medium.

The unprecedented sequence level variability of datasets belonging samples collected whilst investigating the biodiversity of the air surrounding the Antarctic, directed the study towards potential technical variation, in order to assess the hypothesis that the unprecedented sequence level biodiversity of Antarctic air samples was due in part to technical variation as a result of their low biomass. Multiple efforts were made to improve library preparation, with sample concentration by speed-vac (chapter 2.2.3), the addition of multiple rounds of per sample Ampure XP bead clean up (chapter 2.2.10), per sample Picogreen normalisation (chapter 2.2.9), per sample concentration of DNA, and 20, 30, and 40 cycle PCR preparation steps to the standard Schloss wet lab sequencing protocol employed by NU-OMICS (chapter 2.2.8), as well as a final pre-sequencing run using a MiSeq Nanorun kit to maximise sample normalisation based on expected sample reads. Despite these protocol alterations, continued variation in datasets with regard to normalisation and taxonomic profile suggested that the issue was with sample DNA, not sequencing protocol. This was assessed by preparing a dilution series of *B. subtilis*, which acted as a representative hardy bacterium with spore forming capabilities, making it suitable for atmospheric life, onto membrane filters at concentrations similar to and below those found in the atmosphere. This simple model sample was then used

to assess the efficiency of DNA extraction using the most common method in the literature, and then to assess the lower true sample detection limit of the Illumina MiSeq protocol previously optimised for low biomass samples. It was found that despite its wide use in bioaerosol studies (25, 196, 251-256, 270), the Qiagen Powersoil kit was sub-optimal at extracting DNA from membrane filters, extracting between 3-16% of measurable DNA, and also the efficiency of the extraction appeared to increase for decreasing sample biomass where DNA was measurable. Furthermore, sample negative controls were not consistent as was previously thought (94), making the identification of contaminating taxa more difficult for low biomass samples. The sequencing of these samples, showed an estimated DNA yield representative of $1x10^6$ CFU per mL^{-1} to be the lower limit at which Illumina MiSeq can produce a reliably true signal of a sample when using the library preparation method previously described, meaning an initial biomass in the range of $1x10^7$ CFU per mL^{-1} must be present to perform DNA extraction with the Qiagen Powersoil kit and attain a true community profile. Following these findings, the biomass of samples collected during the ACE cruise was explored by qPCR, and found to be considerably lower than the limit of detection for Illumina MiSeq, and as a result the hypothesis that the unprecedented sequence level biodiversity of Antarctic air samples was due in part to technical variation as a result of their low biomass must be accepted. Therefore, for future low biomass air studies, with an expected CFU per mL^{-1} below $1x10^7$, a more efficient, cleaner, method of DNA extraction should be developed and utilised. Gene specific investigation of biodiversity using qPCR, or culture dependent studies could be utilised to provide a less detailed, but more valid description of bioaerosol biodiversity in low biomass environments.

In order to further understand the causes of variation in low biomass datasets, the impact bioinformatics pipeline can have on the analysis of low biomass air samples was investigated by comparing two commonly used pipelines (QIIME 2 and Mothur), as well as different OTU

picking strategies and different taxonomic databases. This study showed that choice of pipeline can influence the number of retained samples and taxonomic profile of a low biomass dataset, however choice of pipeline did not have a significant impact on sample alpha diversity. Closed reference OTU picking increased the difference in beta diversity, greatly reducing the number of closely clustered samples when compared to *de novo* OTU picking or ASV assignment, showing OTU picking strategy can considerably impact beta diversity when investigating understudied environments. By comparing the results generated when using the latest 2016 RDP database to an older version of the same database curated in 2013, this study also emphasised that how recently a taxonomic database has been updated, does not greatly influence the results of a study, despite this being one of the main reasons cited in online bioinformatics discussions for choosing one database over another.

The impact of sample biomass and bioinformatics approach on aerobiological datasets presented above suggests some caution must be taken with regards to the validity of the patterns observed in the Antarctic dataset, as it is likely that due to the observed biomass within Antarctic air samples, a significant and unidentifiable proportion of the data could potentially be a contaminant, despite considerable efforts taken both when in the lab processing and whilst undertaking analysis of the samples to reduce the impact. Despite this, whilst this dataset remains novel and until disproven by further experimental fieldwork in the Antarctic region, it remains the best to date insight into bacterial communities residing in the lower atmosphere above the Southern Ocean and above terrestrial sub-Antarctic island sites.

Data generated from low biomass air samples using modern molecular techniques can be subject to considerable technical variation, both in the lab and during analysis. Therefore, based upon the findings of this research, it is suggested that when undertaking a low biomass study of air samples, researchers consider undertaking the following steps:

i) Only use the Qiagen Powersoil kit for DNA extraction when sample biomass is expected to be above $1x10^7$ CFU per mL^{-1} and provide some sample quantification data for verification

ii) Only us Illumina MiSeq on sample which have an expected estimated DNA yield representative of $1x10^6$ CFU per mL^{-1}

iii) Run a dilution series of a mock community at a concentration range similar to that expected in samples, from DNA extraction through to sequencing, alongside samples, in order to assess the validity of NGS results

iv) Extract as many kit negative controls as possible alongside samples, and carry these through sequencing in order to identify as high a number of contaminating DNA sequences as possible

v) Use negative controls to remove as many contaminating sequences as possible, using the most appropriate decontamination method for the dataset, and then assess the validity of the dataset following decontamination

vi) Carefully consider the most appropriate bioinformatics strategy for the dataset, with regards to how well described an environment is, what studies are most useful to compare with, and what aspect of biodiversity is of interest

Bibliography

1. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nature Reviews Microbiology. 2014;12:635.

2. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. Nature Microbiology. 2016;1:16048.

3. Seaby RM, Henderson PA. Species Diversity and Richness. Lymington, England: Pisces Conservation Ltd.; 2006.

4. Hooke R. Micrographia : or some physiological descriptions of minute bodies made by magnifying glasses. With observations and inquiries thereupon: London : Printed by Jo. Martyn, and Ja. Allestry ... and are to be sold at their shop ..., 1665.; 1665.

5. Bardell D. The roles of the sense of taste and clean teeth in the discovery of bacteria by Antoni van Leeuwenhoek. Microbiological Reviews. 1983;47(1):121-6.

6. Mancini R, Nigro M, Ippolito G. [Lazzaro Spallanzani and his refutation of the theory of spontaneous generation]. Le infezioni in medicina : rivista periodica di eziologia, epidemiologia, diagnostica, clinica e terapia delle patologie infettive. 2007;15(3):199-206.

7. Pasteur L. Nouvelles expériences relatives aux générations dites spontanées. Paris :: Mallet-Bachelier; 1860.

8. Koch R. The etiology of anthrax, based on the life history of. Bacillus; 1876.

9. Airy H. Pollen-grains in the air [2]. Nature. 1874;10(253):355.

10. Dyar HG. XII.—On Certain Bacteria from the Air of New York City. Annals of the New York Academy of Sciences1894. p. 322-80.

11. Haeckel E. Generelle Morphologie der Organismen: Allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begrundet durch die von Charles Darwin

reformirte Descendenz-Theorie: II Allgemeine Entwickelungsgeschichte der Organismen1866.

12. Dworkin M. Sergei Winogradsky: a founder of modern microbiology and the first microbial ecologist. FEMS microbiology reviews. 2012;36(2):364-79.

13. Checinska A, Probst AJ, Vaishampayan P, White JR, Kumar D, Stepanov VG, et al. Microbiomes of the dust particles collected from the International Space Station and Spacecraft Assembly Facilities. Microbiome. 2015;3(1):50.

14. Ichijo T, Yamaguchi N, Tanigaki F, Shirakawa M, Nasu M. Four-year bacterial monitoring in the International Space Station—Japanese Experiment Module "Kibo" with culture-independent approach. Npj Microgravity. 2016;2:16007.

15. Luongo JC, Barberán A, Hacker-Cary R, Morgan EE, Miller SL, Fierer N. Microbial analyses of airborne dust collected from dormitory rooms predict the sex of occupants. Indoor Air. 2017;27(2):338-44.

16. Pearce DA, Bridge PD, Hughes KA, Sattler B, Psenner R, Russell NJ. Microorganisms in the atmosphere over Antarctica. FEMS Microbiol Ecol. 2009;69(2):143-57.

17. Morris CE, Bardin M, Berge O, Frey-Klett P, Fromin N, Girardin H, et al. Microbial Biodiversity: Approaches to Experimental Design and Hypothesis Testing in Primary Scientific Literature from 1975 to 1999. Microbiology and Molecular Biology Reviews. 2002;66(4):592-616.

18. Hairston NG, Allan JD, Colwell RK, Futuyma DJ, Howell J, Lubin MD, et al. The Relationship between Species Diversity and Stability: An Experimental Approach with Protozoa and Bacteria. Ecology. 1968;49(6):1091-101.

19. Li H, Yang Q, Li J, Gao H, Li P, Zhou H. The impact of temperature on microbial diversity and AOA activity in the Tengchong Geothermal Field, China. Scientific Reports. 2015;5:17056.

20. Zhalnina K, Dias R, de Quadros PD, Davis-Richardson A, Camargo FA, Clark IM, et al. Soil pH determines microbial diversity and composition in the park grass experiment. Microbial ecology. 2015;69(2):395-406.

21. Stomeo F, Makhalanyane TP, Valverde A, Pointing SB, Stevens MI, Cary CS, et al. Abiotic factors influence microbial diversity in permanently cold soil horizons of a maritime-associated Antarctic Dry Valley. FEMS Microbiology Ecology. 2012;82(2):326-40.

22. Hunting E. UV radiation and organic matter composition shape bacterial functional diversity in sediments. Frontiers in Microbiology. 2013;4(317).

23. Fierer N, Liu Z, Rodríguez-Hernández M, Knight R, Henn M, Hernandez MT. Short-Term Temporal Variability in Airborne Bacterial and Fungal Populations. Applied and Environmental Microbiology. 2008;74(1):200-7.

24. Seifried JS, Wichels A, Gerdts G. Spatial distribution of marine airborne bacterial communities. MicrobiologyOpen. 2015;4(3):475-90.

25. Bowers RM, McLetchie S, Knight R, Fierer N. Spatial variability in airborne bacterial communities across land-use types and their relationship to the bacterial communities of potential source environments. ISME J. 2011;5(4):601-12.

26. Pankhurst LJ, Whitby C, Pawlett M, Larcombe LD, McKew B, Deacon LJ, et al. Temporal and spatial changes in the microbial bioaerosol communities in green-waste composting. FEMS Microbiology Ecology. 2012;79(1):229.

27. Hill R, Saetnan ER, Scullion J, Gwynn-Jones D, Ostle N, Edwards A. Temporal and spatial influences incur reconfiguration of Arctic heathland soil bacterial community structure. Environmental Microbiology. 2016;18(6):1942-53.

28. Womack AM, Bohannan BJM, Green JL. Biodiversity and biogeography of the atmosphere. Philosophical Transactions of the Royal Society of London B: Biological Sciences. 2010;365(1558):3645-53.

29. Ross AA, Neufeld JD. Microbial biogeography of a university campus. Microbiome. 2015;3(1):66.

30. Roeselers G, Coolen J, van der Wielen PW, Jaspers MC, Atsma A, de Graaf B, et al. Microbial biogeography of drinking water: patterns in phylogenetic diversity across space and time. Environ Microbiol. 2015;17(7):2505-14.

31. Flores GE, Bates ST, Knights D, Lauber CL, Stombaugh J, Knight R, et al. Microbial Biogeography of Public Restroom Surfaces. PLOS ONE. 2011;6(11):e28132.

32. Bokulich NA, Thorngate JH, Richardson PM, Mills DA. Microbial biogeography of wine grapes is conditioned by cultivar, vintage, and climate. Proceedings of the National Academy of Sciences. 2014;111(1):E139.

33. Fierer N. Microbial Biogeography: Patterns in Microbial Diversity across Space and Time. Accessing Uncultivated Microorganisms: American Society of Microbiology; 2008.

34. Martiny JBH, Bohannan BJM, Brown JH, Colwell RK, Fuhrman JA, Green JL, et al. Microbial biogeography: putting microorganisms on the map. Nat Rev Micro. 2006;4(2):102-12.

35. Amann RI, Ludwig W, Schleifer KH. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev. 1995;59(1):143-69.

36. Relman DA, Falkow S. Identification of uncultured microorganisms: expanding the spectrum of characterized microbial pathogens. Infectious agents and disease. 1992;1(5):245-53.

37. Dahm R. Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. Human genetics. 2008;122(6):565-81.

38. Avery OT, MacLeod CM, McCarty M. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES. The Journal of Experimental Medicine. 1944;79(2):137.

39. Chargaff E. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. Experientia. 1950;6(6):201-9.

40. Watson JD, Crick FH. Molecular structure of nucleic acids. Nature. 1953;171(4356):737-8.

41. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proceedings of the national academy of sciences. 1977;74(12):5463-7.

42. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proceedings of the National Academy of Sciences. 1977;74(11):5088.

43. Woese CR. Bacterial evolution. Microbiological Reviews. 1987;51(2):221-71.

44. Stern S, Powers T, Changchien L-M, Noller HF. RNA-protein interactions in 30S ribosomal subunits: folding and function of 16S rRNA. Science. 1989;244(4906):783-90.

45. Stevenson BS, Schmidt TM. Life History Implications of rRNA Gene Copy Number in Escherichia coli. Applied and Environmental Microbiology. 2004;70(11):6670-7.

46. Větrovský T, Baldrian P. The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. PLOS ONE. 2013;8(2):e57923.

47. Rappe MS, Giovannoni SJ. The uncultured microbial majority. Annual review of microbiology. 2003;57:369-94.

48. Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, et al. Insights from 20 years of bacterial genome sequencing. Functional & integrative genomics. 2015;15(2):141-61.

49. Pulschen AA, Bendia AG, Fricker AD, Pellizari VH, Galante D, Rodrigues F. Isolation of Uncultured Bacteria from Antarctica Using Long Incubation Periods and Low Nutritional Media. Frontiers in Microbiology. 2017;8:1346.

50. Weber CF, Werth JT. Is the lower atmosphere a readily accessible reservoir of culturable, antimicrobial compound-producing Actinomycetales? Frontiers in Microbiology. 2015;6:802.

51. Tucker T, Marra M, Friedman JM. Massively Parallel Sequencing: The Next Big Thing in Genetic Medicine. American Journal of Human Genetics. 2009;85(2):142-54.

52. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. The Isme Journal. 2012;6:1621.

53. Claesson MJ, O'Sullivan O, Wang Q, Nikkilä J, Marchesi JR, Smidt H, et al. Comparative Analysis of Pyrosequencing and a Phylogenetic Microarray for Exploring Microbial Community Structures in the Human Distal Intestine. PLOS ONE. 2009;4(8):e6669.

54. Allali I, Arnold JW, Roach J, Cadenas MB, Butz N, Hassan HM, et al. A comparison of sequencing platforms and bioinformatics pipelines for compositional analysis of the gut microbiome. BMC Microbiology. 2017;17:194.

55. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, et al. Performance comparison of benchtop high-throughput sequencing platforms. Nature Biotechnology. 2012;30:434.

56. Salipante SJ, Kawashima T, Rosenthal C, Hoogestraat DR, Cummings LA, Sengupta DJ, et al. Performance comparison of Illumina and ion torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling. Appl Environ Microbiol. 2014;80(24):7583-91.

57. Edgar RC. UNCROSS: Filtering of high-frequency cross-talk in 16S amplicon reads. bioRxiv. 2016.

58. Emerson JB, Keady PB, Brewer TE, Clements N, Morgan EE, Awerbuch J, et al. Impacts of flood damage on airborne bacteria and fungi in homes after the 2013 Colorado Front Range flood. Environ Sci Technol. 2015;49(5):2675-84.

59. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nature methods. 2010;7(5):335-6.

60. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform. Applied and Environmental Microbiology. 2013;79(17):5112-20.

61. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet C, Al-Ghalith GA, et al. QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. PeerJ Preprints. 2018;6:e27295v2.

62. Rognes T, Flouri T, Nichols B, Quince C, Mahe F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4:e2584.

63. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics (Oxford, England). 2010;26(19):2460-1.

64. Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G, et al. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome research. 2011;21(3):494-504.

65. Edgar RC. Accuracy of microbial community diversity estimated by closed- and open-reference OTUs. PeerJ. 2017;5:e3889.

66. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High resolution sample inference from Illumina amplicon data. Nature methods. 2016;13(7):581-3.

67. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, et al. Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. mSystems. 2017;2(2):e00191-16.

68. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The Isme Journal. 2017;11:2639.

69. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, et al. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nature methods. 2012;10:57.

70. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Research. 2014;42(Database issue):D633-D42.

71. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006;72(7):5069-72.

72. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, et al. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Research. 2014;42(Database issue):D643-D8.

73. Federhen S. The NCBI Taxonomy database. Nucleic Acids Res. 2012;40(Database issue):D136-43.

74. Beiko RG. Microbial Malaise: How Can We Classify the Microbiome? Trends in Microbiology. 2015;23(11):671-9.

75. Balvočiūtė M, Huson DH. SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare? BMC Genomics. 2017;18(2):114.

76. Werner JJ, Koren O, Hugenholtz P, DeSantis TZ, Walters WA, Caporaso JG, et al. Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys. The ISME Journal. 2012;6(1):94-103.

77. Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R. PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics (Oxford, England). 2010;26(2):266-7.

78. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution. 2013;30(4):772-80.

79. Rajan V. A method of alignment masking for refining the phylogenetic signal of multiple sequence alignments. Molecular biology and evolution. 2013;30(3):689-712.

80. Sheneman L, Evans J, Foster JA. Clearcut: a fast implementation of relaxed neighbor joining. Bioinformatics (Oxford, England). 2006;22(22):2823-4.

81. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Molecular biology and evolution. 2009;26(7):1641-50.

82. Magurran AE. Measuring biological diversity. Measuring Biological Diversity. 2 ed: Wiley-Blackwell; 2004.

83. Lande R. Statistics and Partitioning of Species Diversity, and Similarity among Multiple Communities. Oikos. 1996;76(1):5-13.

84. May RM. Patterns of species abundance and diversity. Ecology and evolution of communities. 1975:81-120.

85. Simpson EH. Measurement of diversity. nature. 1949;163(4148):688.

86. Faith DP. Conservation evaluation and phylogenetic diversity. Biological Conservation. 1992;61(1):1-10.

87. Morris EK, Caruso T, Buscot F, Fischer M, Hancock C, Maier TS, et al. Choosing and using diversity indices: insights for ecological applications from the German Biodiversity Exploratories. Ecology and Evolution. 2014;4(18):3514-24.

88. McMurdie PJ, Holmes S. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. PLOS Computational Biology. 2014;10(4):e1003531.

89. Hugerth LW, Andersson AF. Analysing Microbial Community Composition through Amplicon Sequencing: From Sampling to Hypothesis Testing. Frontiers in Microbiology. 2017;8(1561).

90. Sinha R, Abnet CC, White O, Knight R, Huttenhower C. The microbiome quality control project: baseline study design and future directions. Genome Biology. 2015;16(1):276.

91. Glassing A, Dowd SE, Galandiuk S, Davis B, Chiodini RJ. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. Gut Pathogens. 2016;8:24.

92. Jervis-Bardy J, Leong LEX, Marri S, Smith RJ, Choo JM, Smith-Vaughan HC, et al. Deriving accurate microbiota profiles from human samples with low bacterial content through post-sequencing processing of Illumina MiSeq data. Microbiome. 2015;3(1):19.

93. Lusk RW. Diverse and Widespread Contamination Evident in the Unmapped Depths of High Throughput Sequencing Data. PLOS ONE. 2014;9(10):e110808.

94. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biology. 2014;12(1):87.

95. Campbell AM, Fleisher J, Sinigalliano C, White JR, Lopez JV. Dynamics of marine bacterial community diversity of the coastal waters of the reefs, inlets, and wastewater outfalls of southeast Florida. MicrobiologyOpen. 2015;4(3):390-408.

96. Jenkins JR, Viger M, Arnold EC, Harris ZM, Ventura M, Miglietta F, et al. Biochar alters the soil microbiome and soil function: results of next-generation amplicon sequencing across Europe. GCB Bioenergy. 2017;9(3):591-612.

97. de Souza RSC, Okura VK, Armanhi JSL, Jorrín B, Lozano N, da Silva MJ, et al. Unlocking the bacterial and fungal communities assemblages of sugarcane microbiome. Scientific Reports. 2016;6:28774.

98. Galan M, Razzauti M, Bard E, Bernard M, Brouat C, Charbonnel N, et al. 16S rRNA Amplicon Sequencing for Epidemiological Surveys of Bacteria in Wildlife. mSystems. 2016;1(4).

99. Davis NM, Proctor D, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. bioRxiv. 2017.

100. Cuthbertson L, Pearce DA. Aeromicrobiology. Psychrophiles: From Biodiversity to Biotechnology: Springer; 2017. p. 41-55.

101. Meier FC, Lindbergh CA. Collecting micro-organisms from the Arctic atmosphere. The scientific monthly. 1935;40.

102. Lurie MB. Experimental epidemiology of Tuberculosis. The Journal of Experimental Medicine. 1930;51(5):743.

103. Brown WA, Allison VD. Infection of the air of scarlet-fever wards with Streptococcus pyogenes. Journal of Hygiene. 1937;37(1):1-13.

104. Andrewes CH, Glover RE. Spread of Infection from the Respiratory Tract of the Ferret. I. Transmission of Influenza A Virus. British Journal of Experimental Pathology. 1941;22(2):91-7.

105. Bourdillon RB, Lidwell OM, Thomas JC. A slit sampler for collecting and counting air-borne bacteria. Journal of Hygiene. 1941;41(2):197-224.

106. Wells WF. Bacteriologic procedures in sanitary air analysis: With special reference to air disinfection. Journal of Bacteriology. 1943;46(6):549-57.

107. Hirst JM. An automatic volumetric spore trap. Annals of Applied Biology. 1952;39(2):257-65.

108. Andersen AA. New sampler for the collection, sizing and enumeration of viable airborne particles. Journal of Bacteriology. 1958;76(5):471-84.

109. NASA. Earth's Atmospheric Layers NASA2013 [updated 31/07/2015. Available from: https://www.nasa.gov/mission_pages/sunearth/science/atmosphere-layers2.html.

110. Imshenetsky AA, Lysenko SV, Kasakov GA, Ramkova NV. Resistance of stratospheric and mesospheric micro-organisms to extreme factors. Life sciences and space research. 1977;15:37-9.

111. Sattler B, Puxbaum H, Psenner R. Bacterial growth in supercooled cloud droplets. Geophysical Research Letters. 2001;28(2):239-42.

112. Burrows SM, Elbert W, Lawrence MG, Pöschl U. Bacteria in the global atmosphere – Part 1: Review and synthesis of literature data for different ecosystems. Atmos Chem Phys. 2009;9(23):9263-80.

113. Jones AM, Harrison RM. The effects of meteorological factors on atmospheric bioaerosol concentrations—a review. Science of The Total Environment. 2004;326(1–3):151-80.

114. Lighthart B, Shaffer BT. Bacterial flux from chaparral into the atmosphere in mid-summer at a high desert location. Atmospheric Environment. 1994;28(7):1267-74.

115. Blanchard DC, Syzdek LD. Water-to-Air Transfer and Enrichment of Bacteria in Drops from Bursting Bubbles. Applied and Environmental Microbiology. 1982;43(5):1001-5.

116. Griffin DW. Atmospheric Movement of Microorganisms in Clouds of Desert Dust and Implications for Human Health. Clinical Microbiology Reviews. 2007;20(3):459-77.

117. Stewart FJ. Where the genes flow. Nature Geosci. 2013;6(9):688-90.

118. Burrows SM, Butler T, Jöckel P, Tost H, Kerkweg A, Pöschl U, et al. Bacteria in the global atmosphere – Part 2: Modeling of emissions and transport between different ecosystems. Atmos Chem Phys. 2009;9(23):9281-97.

119. Smith DJ, Jaffe DA, Birmele MN, Griffin DW, Schuerger AC, Hee J, et al. Free tropospheric transport of microorganisms from Asia to North America. Microbial ecology. 2012;64(4):973-85.

120. Deguillaume L, Leriche M, Amato P, Ariya PA, Delort AM, Pöschl U, et al. Microbiology and atmospheric processes: chemical interactions of primary biological aerosols. Biogeosciences. 2008;5(4):1073-84.

121. Dimmick RL, Straat PA, Wolochow H, Levin GV, Chatigny MA, Schrot JR. Evidence for metabolic activity of airborne bacteria. Journal of Aerosol Science. 1975;6(6):387-93.

122. Vali G. Quantitative Evaluation of Experimental Results an the Heterogeneous Freezing Nucleation of Supercooled Liquids. Journal of the Atmospheric Sciences. 1971;28(3):402-9.

123. Bauer H, Kasper-Giebl A, Löflund M, Giebl H, Hitzenberger R, Zibuschka F, et al. The contribution of bacteria and fungal spores to the organic carbon content of cloud water, precipitation and aerosols. Atmospheric Research. 2002;64(1–4):109-19.

124. Ariya PA, Nepotchatykh O, Ignatova O, Amyot M. Microbiological degradation of atmospheric organic compounds. Geophysical Research Letters. 2002;29(22):34-1--4.

125. Hill KA, Shepson PB, Galbavy ES, Anastasio C, Kourtev PS, Konopka A, et al. Processing of atmospheric nitrogen by clouds above a forest environment. Journal of Geophysical Research. 2007;112(D11).

126. Després VR, Huffman JA, Burrows SM, Hoose C, Safatov AS, Buryak G, et al. Primary biological aerosol particles in the atmosphere: a review. Tellus B; Vol 64 (2012). 2012.

127. Margesin R, Miteva V. Diversity and ecology of psychrophilic microorganisms. Research in microbiology. 2011;162(3):346-61.

128. Möhler O, DeMott PJ, Vali G, Levin Z. Microbiology and atmospheric processes: the role of biological particles in cloud physics. Biogeosciences. 2007;4(6):1059-71.

129. Morris CE, Conen F, Alex Huffman J, Phillips V, Pöschl U, Sands DC. Bioprecipitation: a feedback cycle linking Earth history, ecosystem dynamics and land use through biological ice nucleators in the atmosphere. Global Change Biology. 2014;20(2):341-51.

130. Fahlgren C, Hagstrom A, Nilsson D, Zweifel UL. Annual variations in the diversity, viability, and origin of airborne bacteria. Appl Environ Microbiol. 2010;76(9):3015-25.

131. Madsen AM, Zervas A, Tendal K, Nielsen JL. Microbial diversity in bioaerosol samples causing ODTS compared to reference bioaerosol samples as measured using Illumina sequencing and MALDI-TOF. Environmental Research. 2015;140:255-67.

132. Nonnenmann MW, Bextine B, Dowd SE, Gilmore K, Levin JL. Culture-Independent Characterization of Bacteria and Fungi in a Poultry Bioaerosol Using Pyrosequencing: A New Approach. Journal of Occupational and Environmental Hygiene. 2010;7(12):693-9.

133. Barberán A, Ladau J, Leff JW, Pollard KS, Menninger HL, Dunn RR, et al. Continental-scale distributions of dust-associated bacteria and fungi. Proceedings of the National Academy of Sciences. 2015;112(18):5756-61.

134. Griffin DW, Gonzalez C, Teigell N, Petrosky T, Northup DE, Lyles M. Observations on the use of membrane filtration and liquid impingement to collect airborne microorganisms in various atmospheric environments. Aerobiologia. 2011;27(1):25-35.

135. Moeller R, Setlow P, Reitz G, Nicholson WL. Roles of Small, Acid-Soluble Spore Proteins and Core Water Content in Survival of Bacillus subtilis Spores Exposed to Environmental Solar UV Radiation. Applied and Environmental Microbiology. 2009;75(16):5202-8.

136. Kochkina GA, Ivanushkina NE, Karasev SG, Gavrish EY, Gurina LV, Evtushenko LI, et al. Survival of Micromycetes and *Actinobacteria* under Conditions of Long-Term Natural Cryopreservation. Microbiology. 2001;70(3):356-64.

137. Johansson E, Adhikari A, Reponen T, Yermakov M, Grinshpun SA. Association Between Increased DNA Mutational Frequency and Thermal Inactivation of Aerosolized Bacillus Spores Exposed to Dry Heat. Aerosol Science and Technology. 2011;45(3):376-81.

138. Wainwright M, Wickramasinghe NC, Narlikar JV, Rajaratnam P. Microorganisms cultured from stratospheric air samples obtained at 41 km. FEMS Microbiology Letters. 2003;218(1):161.

139. Bowers RM, McCubbin IB, Hallar AG, Fierer N. Seasonal variability in airborne bacterial communities at a high-elevation site. Atmospheric Environment. 2012;50:41-9.

140. Dong L, Qi J, Shao C, Zhong X, Gao D, Cao W, et al. Concentration and size distribution of total airborne microbes in hazy and foggy weather. Science of The Total Environment. 2016;541:1011-8.

141. Lighthart B. Mini-review of the concentration variations found inthe alfresco atmospheric bacterial populations. Aerobiologia. 2000;16(1):7-16.

142. Polunin N, Pady SM, Kelly CD. Arctic aerobiology. Nature. 1947;160(4077):876-7.

143. Harding T, Jungblut AD, Lovejoy C, Vincent WF. Microbes in High Arctic Snow and Implications for the Cold Biosphere. Applied and Environmental Microbiology. 2011;77(10):3234-43.

144. Amato P, Hennebelle R, Magand O, Sancelme M, Delort A-M, Barbante C, et al. Bacterial characterization of the snow cover at Spitzberg, Svalbard. FEMS Microbiology Ecology. 2007;59(2):255-64.

145. Møller AK, Søborg DA, Al-Soud WA, Sørensen SJ, Kroer N. Bacterial community structure in High-Arctic snow and freshwater as revealed by pyrosequencing of 16S rRNA genes and cultivation. 2013. 2013.

146. Marshall WA. Aerial dispersal of lichen soredia in the maritime Antarctic. New Phytologist. 1996;134(3):523-30.

147. Marshall WA, Chalmers MO. Airborne dispersal of antarctic terrestrial algae and cyanobacteria. Ecography. 1997;20(6):585-94.

148. Marshall WA. Aerial Transport of Keratinaceous Substrate and Distribution of the Fungus Geomyces pannorum in Antarctic Soils. Microbial ecology. 1998;36(2):212-9.

149. Vincent WF. Evolutionary origins of Antarctic microbiota: invasion, selection and endemism. Antarctic Science. 2000;12(3):374-85.

150. Herbold CW, Lee CK, McDonald IR, Cary SC. Evidence of global-scale aeolian dispersal and endemism in isolated geothermal microbial communities of Antarctica. Nat Commun. 2014;5.

151. Pearce DA, Alekhina IA, Terauds A, Wilmotte A, Quesada A, Edwards A, et al. Aerobiology Over Antarctica – A New Initiative for Atmospheric Ecology. Frontiers in Microbiology. 2016;7:16.

152. Bottos EM, Woo AC, Zawar-Reza P, Pointing SB, Cary SC. Airborne bacterial populations above desert soils of the McMurdo Dry Valleys, Antarctica. Microbial ecology. 2014;67(1):120-8.

153. Pearce DA, Hughes K, Lachlan-Cope T, Harangozo S, Jones AE. Biodiversity of air-borne microorganisms at Halley station, Antarctica. Extremophiles : life under extreme conditions. 2010;14(2):145-59.

154. Hughes KA, McCartney HA, Lachlan-cope TA, Pearce DA. A Preliminary Study of Airborne Microbial Biodiversity Over Peninsular Antarctica2004.

155. Van Houdt R, De Boever P, Coninx I, Le Calvez C, Dicasillati R, Mahillon J, et al. Evaluation of the Airborne Bacterial Population in the Periodically Confined Antarctic Base Concordia. Microbial ecology. 2009;57(4):640-8.

156. Baas Becking LGM. Geobiologie of inleiding tot de milieukunde. Den Haag: W.P. Van Stockum & Zoon; 1934.

157. Finlay BJ, Clarke KJ. Ubiquitous dispersal of microbial species. Nature. 1999;400(6747):828-.

158. Alger AS. Diatoms of the McMurdo Dry Valleys, Antarctica: A taxonomic appraisal including a detailed study of the genus Hantzschia. Ann Arbor, Michigan1999.

159. Vyverman W, Verleyen E, Wilmotte A, Hodgson DA, Willems A, Peeters K, et al. Evidence for widespread endemism among Antarctic micro-organisms. Polar Science. 2010;4(2):103-13.

160. Lovejoy C, Vincent WF, Bonilla S, Roy S, Martineau M-J, Terrado R, et al. Distribution, Phylogeny, and Growth of Cold-Adapted Picoprasinophytes in Arctic Seas. Journal of Phycology. 2007;43(1):78-89.

161. Brodie EL, DeSantis TZ, Parker JP, Zubietta IX, Piceno YM, Andersen GL. Urban aerosols harbor diverse and dynamic bacterial populations. Proceedings of the National Academy of Sciences of the United States of America. 2007;104(1):299-304.

162. Gallisai R, Peters F, Volpe G, Basart S, Baldasano JM. Saharan Dust Deposition May Affect Phytoplankton Growth in the Mediterranean Sea at Ecological Time Scales. PLoS ONE. 2014;9(10):e110762.

163. Giddings SN, MacCready P, Hickey BM, Banas NS, Davis KA, Siedlecki SA, et al. Hindcasts of potential harmful algal bloom transport pathways on the Pacific Northwest coast. Journal of Geophysical Research: Oceans. 2014;119(4):2439-61.

164. Smith DJ, Griffin DW, Jaffe DA. The high life: Transport of microbes in the atmosphere. Eos, Transactions American Geophysical Union. 2011;92(30):249-50.

165. Molesworth AM, Cuevas LE, Connor SJ, Morse AP, Thomson MC. Environmental Risk and Meningitis Epidemics in Africa. Emerging Infectious Diseases. 2003;9(10):1287-93.

166. Chen PS, Tsai FT, Lin CK, Yang CY, Chan CC, Young CY, et al. Ambient influenza and avian influenza virus during dust storm days and background days. Environmental health perspectives. 2010;118(9):1211-6.

167. Fish KE, Collins R, Green NH, Sharpe RL, Douterelo I, Osborn AM, et al. Characterisation of the Physical Composition and Microbial Community Structure of Biofilms within a Model Full-Scale Drinking Water Distribution System. PLOS ONE. 2015;10(2):e0115824.

168. Suzuki MT, Taylor LT, DeLong EF. Quantitative analysis of small-subunit rRNA genes in mixed microbial populations via 5'-nuclease assays. Applied and environmental microbiology. 2000;66(11):4605-14.

169. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proceedings of the National Academy of Sciences. 2011;108(Supplement 1):4516.

170. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nature methods. 2012;9(7):671-5.

171. McMurdie PJ, Holmes S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLOS ONE. 2013;8(4):e61217.

172. Fisher RA. Student. Annals of Eugenics. 1939;9(1):1-9.

173. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. Journal of the American statistical Association. 1952;47(260):583-621.

174. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology. 2014;15(12):550-.

175. Vali G, Christensen M, Fresh RW, Galyan EL, Maki LR, Schnell RC. Biogenic Ice Nuclei. Part II: Bacterial Sources. Journal of the Atmospheric Sciences. 1976;33(8):1565-70.

176. Margesin R, Schinner F, Marx JC, Gerday C. Psychrophiles: From biodiversity to biotechnology2008. 1-462 p.

177. Shinn EA, Smith GW, Prospero JM, Betzer P, Hayes ML, Garrison V, et al. African dust and the demise of Caribbean Coral Reefs. Geophysical Research Letters. 2000;27(19):3029-32.

178. Litchman E. Invisible invaders: non-pathogenic invasive microbes in aquatic and terrestrial ecosystems. Ecology Letters. 2010;13(12):1560-72.

179. DeLeon-Rodriguez N, Lathem TL, Rodriguez-R LM, Barazesh JM, Anderson BE, Beyersdorf AJ, et al. Microbiome of the upper troposphere: Species composition and prevalence, effects of tropical storms, and atmospheric implications. Proceedings of the National Academy of Sciences. 2013;110(7):2575-80.

180. Lopatina A, Medvedeva S, Shmakov S, Logacheva MD, Krylenkov V, Severinov K. Metagenomic Analysis of Bacterial Communities of Antarctic Surface Snow. Frontiers in Microbiology. 2016;7(398).

181. Smith DJ, Timonen HJ, Jaffe DA, Griffin DW, Birmele MN, Perry KD, et al. Intercontinental Dispersal of Bacteria and Archaea by Transpacific Winds. Applied and Environmental Microbiology. 2013;79(4):1134-9.

182. Population of Svalbard, 1 January 2015: Statistics Norway; 2015 [updated 9/4/15. Available from: https://www.ssb.no/en/befolkning/statistikker/befsvalbard/halvaar/2015-04-09.

183. Stein AF D, RR, Rolph, GD, Stunder, BJ, Cohen, MD, Ngan, F. NOAA's HYSPLIT Atmospheric Transport and Dispersion Modeling System. Bulletin of the American Meteorological Society. 2015;96(12):2059-77.

184. Hammer Oyvind DH, Paul Ryan. past: paleontological statistics software package for education and data analysis. Palaeontologia Electronica. 2001;4(1).

185. Grzesiak J, Górniak D, Świątecki A, Aleksandrzak-Piekarczyk T, Szatraj K, Zdanowski MK. Microbial community development on the surface of Hans and Werenskiold Glaciers (Svalbard, Arctic): a comparison. Extremophiles : life under extreme conditions. 2015;19(5):885-97.

186. Zhang Sh, Hou Sg, Yang Gl, Wang Jh. Bacterial community in the East Rongbuk Glacier, Mt. Qomolangma (Everest) by culture and culture-independent methods. Microbiological Research. 2010;165(4):336-45.

187. Singh P, Singh SM, Roy U. Taxonomic characterization and the bio-potential of bacteria isolated from glacier ice cores in the High Arctic. Journal of Basic Microbiology. 2016;56(3):275-85.

188. Li M, Qi J, Zhang H, Huang S, Li L, Gao D. Concentration and size distribution of bioaerosols in an outdoor environment in the Qingdao coastal region. The Science of the total environment. 2011;409(19):3812-9.

189. Lewandowski R, Kozłowska K, Szpakowska M, Trafny EA. Evaluation of applicability of the Sartorius Airport MD8 sampler for detection of Bacillus endospores in indoor air. Environmental Monitoring and Assessment. 2013;185(4):3517-26.

190. Stewart SL, Grinshpun SA, Willeke K, Terzieva S, Ulevicius V, Donnelly J. Effect of impact stress on microbial recovery on an agar surface. Applied and Environmental Microbiology. 1995;61(4):1232-9.

191. Vartoukian SR, Palmer RM, Wade WG. Strategies for culture of 'unculturable' bacteria. FEMS Microbiol Lett. 2010;309(1):1-7.

192. Shin S-K, Kim J, Ha S-m, Oh H-S, Chun J, Sohn J, et al. Metagenomic Insights into the Bioaerosols in the Indoor and Outdoor Environments of Childcare Facilities. PLOS ONE. 2015;10(5):e0126960.

193. Galperin MY. Genome Diversity of Spore-Forming *Firmicutes*. Microbiology spectrum. 2013;1(2):TBS-0015-2012.

194. Lightfield J, Fram NR, Ely B. Across Bacterial Phyla, Distantly-Related Genomes with Similar Genomic GC Content Have Similar Patterns of Amino Acid Usage. PLOS ONE. 2011;6(3):e17677.

195. Gupta RS. The phylogeny of *proteobacteria*: relationships to other eubacterial phyla and eukaryotes. FEMS microbiology reviews. 2000;24(4):367-402.

196. Bowers RM, Clements N, Emerson JB, Wiedinmyer C, Hannigan MP, Fierer N. Seasonal Variability in Bacterial and Fungal Diversity of the Near-Surface Atmosphere. Environmental Science & Technology. 2013;47(21):12097-106.

197. Zeng YX, Yan M, Yu Y, Li HR, He JF, Sun K, et al. Diversity of bacteria in surface ice of Austre Lovenbreen glacier, Svalbard. Archives of microbiology. 2013;195(5):313-22.

198. Stibal M, Gözdereliler E, Cameron KA, Box JE, Stevens IT, Gokul JK, et al. Microbial abundance in surface ice on the Greenland Ice Sheet. Frontiers in Microbiology. 2015;6.

199. Zhang G, Cao T, Ying J, Yang Y, Ma L. Diversity and novelty of *actinobacteria* in Arctic marine sediments. Antonie van Leeuwenhoek. 2014;105(4):743-54.

200. Lysnes K, Thorseth IH, Steinsbu BO, Øvreås L, Torsvik T, Pedersen RB. Microbial community diversity in seafloor basalt from the Arctic spreading ridges. FEMS Microbiology Ecology. 2004;50(3):213.

201. Choidash B, Begum Z, Shivaji S. Bacterial diversity of Ny-Ålesund, Arctic Archipelago Svalbard. Mong J Biol Sci. 2012;10(1-2):67-72.

202. Abraham WP, Thomas S. Draft Genome Sequence of *Pseudomonas psychrophila* MTCC 12324, Isolated from the Arctic at 79 degrees N. Genome Announc. 2015;3(3).

203. Yergeau E, Sanschagrin S, Beaumier D, Greer CW. Metagenomic analysis of the bioremediation of diesel-contaminated Canadian high arctic soils. PLoS One. 2012;7(1):e30058.

204. Yang GL, Hou SG, Le Baoge R, Li ZG, Xu H, Liu YP, et al. Differences in Bacterial Diversity and Communities Between Glacial Snow and Glacial Soil on the Chongce Ice Cap, West Kunlun Mountains. Sci Rep. 2016;6:36548.

205. Oh HM, Lee K, Jang Y, Kang I, Kim HJ, Kang TW, et al. Genome Sequence of Strain IMCC9480, a Xanthorhodopsin-Bearing *Betap*roteobacterium Isolated from the Arctic Ocean. Journal of Bacteriology. 2011;193(13):3421-.

206. Schneiker S, Martins dos Santos VA, Bartels D, Bekel T, Brecht M, Buhrmester J, et al. Genome sequence of the ubiquitous hydrocarbon-degrading marine bacterium Alcanivorax borkumensis. Nat Biotechnol. 2006;24(8):997-1004.

207. Bell TH, Yergeau E, Maynard C, Juck D, Whyte LG, Greer CW. Predictable bacterial composition and hydrocarbon degradation in Arctic soils following diesel and nutrient disturbance. ISME J. 2013;7(6):1200-10.

208. Miller RV, Whyte LG. Polar Microbiology: Life in a Deep Freeze: American Society of Microbiology; 2012.

209. Achberger AM, Christner BC, Michaud AB, Priscu JC, Skidmore ML, Vick-Majors TJ, et al. Microbial Community Structure of Subglacial Lake Whillans, West Antarctica. Frontiers in microbiology. 2016;7:1457-.

210. Kwon M, Kim M, Takacs-Vesbach C, Lee J, Hong SG, Kim SJ, et al. Niche specialization of bacteria in permanently ice-covered lakes of the McMurdo Dry Valleys, Antarctica. Environ Microbiol. 2017;19(6):2258-71.

211. Malard LA, Šabacká M, Magiopoulos I, Mowlem M, Hodson A, Tranter M, et al. Spatial Variability of Antarctic Surface Snow Bacterial Communities. Frontiers in Microbiology. 2019;10(461).

212. Dennis PG, Newsham KK, Rushton SP, O'Donnell AG, Hopkins DW. Soil bacterial diversity is positively associated with air temperature in the maritime Antarctic. Scientific Reports. 2019;9(1):2686.

213. Vincent WF. Evolutionary origins of Antarctic microbiota: invasion, selection and endemism. Antarctic Science. 2004;12(3):374-85.

214. Archer SDJ, Lee KC, Caruso T, Maki T, Lee CK, Cary SC, et al. Airborne microbial transport limitation to isolated Antarctic soil habitats. Nature Microbiology. 2019;4(6):925-32.

215. Salazar G, Sunagawa S. Marine microbial diversity. Current Biology. 2017;27(11):R489-R94.

216. Poulain S, Bourouiba L. Biosurfactants Change the Thinning of Contaminated Bubbles at Bacteria-Laden Water Interfaces. Physical Review Letters. 2018;121(20):204502.

217. Salazar G, Cornejo-Castillo FM, Benítez-Barrios V, Fraile-Nuez E, Álvarez-Salgado XA, Duarte CM, et al. Global diversity and biogeography of deep-sea pelagic prokaryotes. The Isme Journal. 2015;10:596.

218. Milici M, Vital M, Tomasch J, Badewien TH, Giebel H-A, Plumeier I, et al. Diversity and community composition of particle-associated and free-living bacteria in mesopelagic and bathypelagic Southern Ocean water masses: Evidence of dispersal limitation in the Bransfield Strait. Limnology and Oceanography. 2017;62(3):1080-95.

219. Bottos EM, Scarrow JW, Archer SDJ, McDonald IR, Cary SC. Bacterial Community Structures of Antarctic Soils. In: Cowan DA, editor. Antarctic Terrestrial Microbiology: Physical and Biological Properties of Antarctic Soils. Berlin, Heidelberg: Springer Berlin Heidelberg; 2014. p. 9-33.

220. Cuthbertson L, Amores-Arrocha H, Malard AL, Els N, Sattler B, Pearce AD. Characterisation of Arctic Bacterial Communities in the Air above Svalbard. Biology. 2017;6(2).

221. Kobayashi F, Maki T, Kakikawa M, Noda T, Mitamura H, Takahashi A, et al. Atmospheric bioaerosols originating from Adélie penguins (Pygoscelis adeliae): Ecological observations of airborne bacteria at Hukuro Cove, Langhovde, Antarctica. Polar Science. 2016;10(1):71-8.

222. Weisleitner K, Perras A, Moissl-Eichinger C, Andersen DT, Sattler B. Source Environments of the Microbiome in Perennially Ice-Covered Lake Untersee, Antarctica. Frontiers in Microbiology. 2019;10:1019.

223. Bowman JP, Cavanagh J, Austin JJ, Sanderson K. Novel Psychrobacter species from Antarctic ornithogenic soils. International journal of systematic bacteriology. 1996;46(4):841-8.

224. Herlemann DPR, Lundin D, Labrenz M, Jürgens K, Zheng Z, Aspeborg H, et al. Metagenomic De Novo Assembly of an Aquatic Representative of the Verrucomicrobial Class <em>Spartobacteria</em&gt. mBio. 2013;4(3):e00569-12.

225. Galperin MY, Mekhedov SL, Puigbo P, Smirnov S, Wolf YI, Rigden DJ. Genomic determinants of sporulation in *Bacilli* and *Clostridia*: towards the minimal set of sporulation-specific genes. Environmental microbiology. 2012;14(11):2870-90.

226. McHugh AJ, Feehily C, Hill C, Cotter PD. Detection and Enumeration of Spore-Forming Bacteria in Powdered Dairy Products. Frontiers in Microbiology. 2017;8(109).

227. Barka EA, Vatsa P, Sanchez L, Gaveau-Vaillant N, Jacquard C, Meier-Kolthoff JP, et al. Taxonomy, Physiology, and Natural Products of *Actinobacteria*. Microbiol Mol Biol Rev. 2015;80(1):1-43.

228. Wilkins D, Lauro FM, Williams TJ, Demaere MZ, Brown MV, Hoffman JM, et al. Biogeographic partitioning of Southern Ocean microorganisms revealed by metagenomics. Environ Microbiol. 2013;15(5):1318-33.

229. Peter H, Hörtnagl P, Reche I, Sommaruga R. Bacterial diversity and composition during rain events with and without Saharan dust influence reaching a high mountain lake in the Alps. Environmental microbiology reports. 2014;6(6):618-24.

230. Amato P, Joly M, Schaupp C, Attard E, Möhler O, Morris CE, et al. Survival and ice nucleation activity of bacteria as aerosols in a cloud simulation chamber. Atmos Chem Phys. 2015;15(11):6455-65.

231. Lee J, Cho J, Cho Y-J, Cho A, Woo J, Lee J, et al. The latitudinal gradient in rock-inhabiting bacterial community compositions in Victoria Land, Antarctica. Science of The Total Environment. 2019;657:731-8.

232. Aagaard K, Ma J, Antony KM, Ganu R, Petrosino J, Versalovic J. The Placenta Harbors a Unique Microbiome. Science Translational Medicine. 2014;6(237):237ra65.

233. Dickson RP, Erb-Downward JR, Freeman CM, McCloskey L, Falkowski NR, Huffnagle GB, et al. Bacterial Topography of the Healthy Human Lower Respiratory Tract. mBio. 2017;8(1):e02287-16.

234. Zhu L, Luo F, Hu W, Han Y, Wang Y, Zheng H, et al. Bacterial communities in the womb during healthy pregnancy. Frontiers in microbiology. 2018;9:2163.

235. Karcher SJ. 6 - POLYMERASE CHAIN REACTION. In: Karcher SJ, editor. Molecular Biology. San Diego: Academic Press; 1995. p. 215-27.

236. Kennedy NA, Walker AW, Berry SH, Duncan SH, Farquarson FM, Louis P, et al. The Impact of Different DNA Extraction Kits and Laboratories upon the Assessment of Human Gut Microbiota Composition by 16S rRNA Gene Sequencing. PLOS ONE. 2014;9(2):e88982.

237. Kulakov LA, McAlister MB, Ogden KL, Larkin MJ, O'Hanlon JF. Analysis of bacteria contaminating ultrapure water in industrial systems. Applied and environmental microbiology. 2002;68(4):1548-55.

238. Rand KH, Houck H. Taq polymerase contains bacterial DNA of unknown origin. Molecular and cellular probes. 1990;4(6):445-50.

239. Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. Appl Environ Microbiol. 2005;71(12):7724-36.

240. Huber T, Faulkner G, Hugenholtz P. Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. Bioinformatics (Oxford, England). 2004;20(14):2317-9.

241. Sze MA, Schloss PD. The Impact of DNA Polymerase and Number of Rounds of Amplification in PCR on 16S rRNA Gene Sequence Data. mSphere. 2019;4(3):e00163-19.

242. Ghyselinck J, Pfeiffer S, Heylen K, Sessitsch A, De Vos P. The Effect of Primer Choice and Short Read Sequences on the Outcome of 16S rRNA Gene Based Diversity Studies. PLOS ONE. 2013;8(8):e71360.

243. Lauder AP, Roche AM, Sherrill-Mix S, Bailey A, Laughlin AL, Bittinger K, et al. Comparison of placenta samples with contamination controls does not provide evidence for a distinct placenta microbiota. Microbiome. 2016;4(1):29.

244. Eisenhofer R, Minich JJ, Marotz C, Cooper A, Knight R, Weyrich LS. Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. Trends in Microbiology. 2019;27(2):105-17.

245. Stinson LF, Keelan JA, Payne MS. Identification and removal of contaminating microbial DNA from PCR reagents: impact on low-biomass microbiome analyses. Letters in Applied Microbiology. 2019;68(1):2-8.

246. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome. 2018;6(1):226.

247. Knights D, Kuczynski J, Charlson ES, Zaneveld J, Mozer MC, Collman RG, et al. Bayesian community-wide culture-independent microbial source tracking. Nature methods. 2011;8:761.

248. Karstens L, Asquith M, Davin S, Fair D, Gregory WT, Wolfe AJ, et al. Controlling for Contaminants in Low-Biomass 16S rRNA Gene Sequencing Experiments. mSystems. 2019;4(4):e00290-19.

249. Jari Oksanen FGB, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs and Helene Wagner. vegan: Community Ecology Package. R package version 2.5-6. 2019.

250. Wickham H. ggplot2: Elegant Graphics for Data Analysis: Springer Publishing Company, Incorporated; 2009. 216 p.

251. Arfken AM, Song B, Sung JS. Comparison of airborne bacterial communities from a hog farm and spray field. Journal of microbiology and biotechnology. 2015;25(5):709-17.

252. Emerson JB, Keady PB, Clements N, Morgan EE, Awerbuch J, Miller SL, et al. High temporal variability in airborne bacterial diversity and abundance inside single-family residences. Indoor Air. 2017;27(3):576-86.

253. Jeon EM, Kim HJ, Jung K, Kim JH, Kim MY, Kim YP, et al. Impact of Asian dust events on airborne bacterial community assessed by molecular analyses. Atmospheric Environment. 2011;45(25):4313-21.

254. Kembel SW, Meadow JF, O'Connor TK, Mhuireach G, Northcutt D, Kline J, et al. Architectural design drives the biogeography of indoor bacterial communities. PLoS One. 2014;9(1):e87093.

255. Kumari P, Choi HL. Seasonal variability in airborne biotic contaminants in swine confinement buildings. PloS one. 2014;9(11):e112897-e.

256. Leung MHY, Wilkins D, Li EKT, Kong FKF, Lee PKH. Indoor-Air Microbiome in an Urban Subway Network: Diversity and Dynamics. Applied and Environmental Microbiology. 2014;80(21):6760.

257. Brandt J, Albertsen M. Investigation of Detection Limits and the Influence of DNA Extraction and Primer Choice on the Observed Microbial Communities in Drinking Water Samples Using 16S rRNA Gene Amplicon Sequencing. Frontiers in Microbiology. 2018;9(2140).

258. Chakravorty S, Helb D, Burday M, Connell N, Alland D. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. Journal of microbiological methods. 2007;69(2):330-9.

259. Vierna J, Dona J, Vizcaino A, Serrano D, Jovani R. PCR cycles above routine numbers do not compromise high-throughput DNA barcoding results. Genome. 2017;60(10):868-73.

260. Archer SDJ, Lee KC, Caruso T, Maki T, Lee CK, Cowan DA, et al. Microbial dispersal limitation to isolated soil habitats in the McMurdo Dry Valleys of Antarctica. bioRxiv. 2018:493411.

261. Goodrich JK, Di Rienzi SC, Poole AC, Koren O, Walters WA, Caporaso JG, et al. Conducting a microbiome study. Cell. 2014;158(2):250-62.

262. Lindgreen S, Adair KL, Gardner PP. An evaluation of the accuracy and speed of metagenome analysis tools. Scientific Reports. 2016;6:19233.

263. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Applied and Environmental Microbiology. 2009;75(23):7537.

264. Xue Z, Kable ME, Marco ML. Impact of DNA Sequencing and Analysis Methods on 16S rRNA Gene Bacterial Community Analysis of Dairy Products. mSphere. 2018;3(5).

265. López-García A, Pineda-Quiroga C, Atxaerandio R, Pérez A, Hernández I, García-Rodríguez A, et al. Comparison of Mothur and QIIME for the Analysis of Rumen Microbiota Composition Based on 16S rRNA Amplicon Sequences. Frontiers in Microbiology. 2018;9(3010).

266. Edgar RC. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. Bioinformatics (Oxford, England). 2018;34(14):2371-5.

267. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics (Oxford, England). 2018;35(3):526-8.

268. Plummer E, Twin J, Bulach DM, Garland SM, Tabrizi SN. A comparison of three bioinformatics pipelines for the analysis of preterm gut microbiota using 16S rRNA gene sequencing data. Journal of Proteomics & Bioinformatics. 2015;8(12):283.

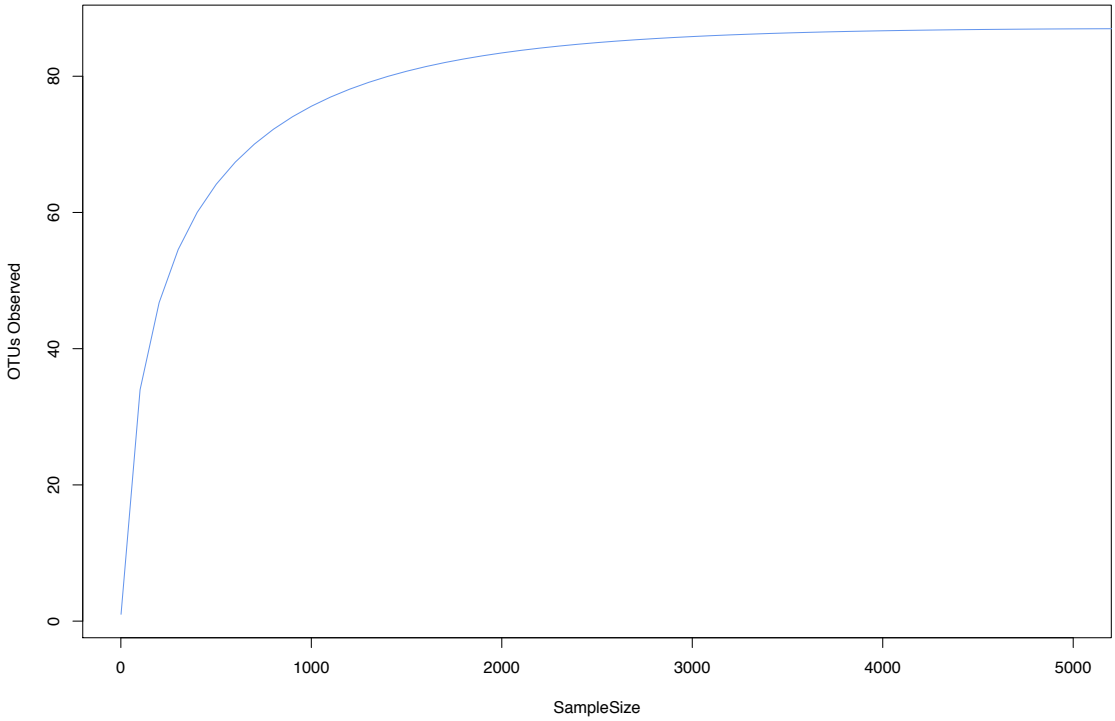
269. Šantl-Temkiv T, Gosewinkel U, Starnawski P, Lever M, Finster K. Aeolian dispersal of bacteria in southwest Greenland: their sources, abundance, diversity and physiological states. FEMS Microbiology Ecology. 2018;94(4).

270. Lee S, Choi B, Yi SM, Ko G. Characterization of microbial community during Asian dust events in Korea. The Science of the total environment. 2009;407(20):5308-14.

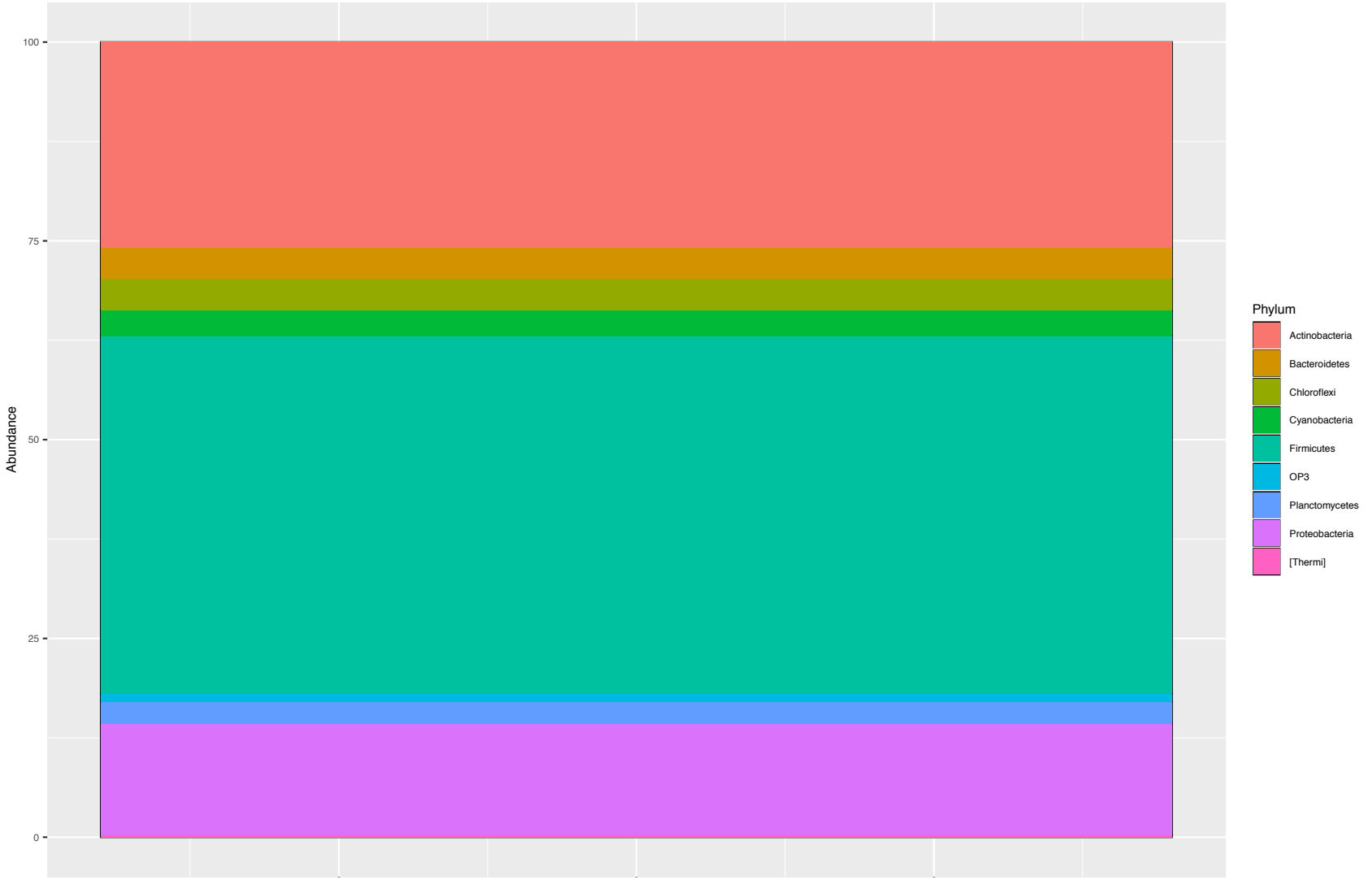
Appendices

Appendix I - Svalbard air sample collected in July 2017 over a 3-day period

An air sample was collected on Svalbard as described in chapter 3. DNA extraction and Illumina MiSeq sequencing was carried out as described in chapter 4. Following decontamination, and the removal of none-bacterial taxa, the sample had 5763 reads. The Shannon index for the sample was 3.09, whilst the Simpson index was 0.85. The amplified sample was sampled in sufficient depth as shown by the rarefaction curve reaching asymptote. There were 9 phyla present, similar to the number of phyla present in the 3-day sample collected in 2015, where there were 10. The top 10 taxa were notably different to those from seen in the 3-day 2015 sample. qPCR revealed the sample DNA extract contained an estimated 101667 CFU per mL^{-1} of bacteria, whilst the relevant kit negative contained 231 CFU per mL^{-1}.



Rarefaction curve for 3-day July 2017 sample collect on Svalbard



Graph of the relative abundance of all phyla for 3-day July 2017 sample collect on Svalbard

Graph of the relative abundance of the top 10 most abundant taxa for 3-day July 2017 sample collect on Svalbar

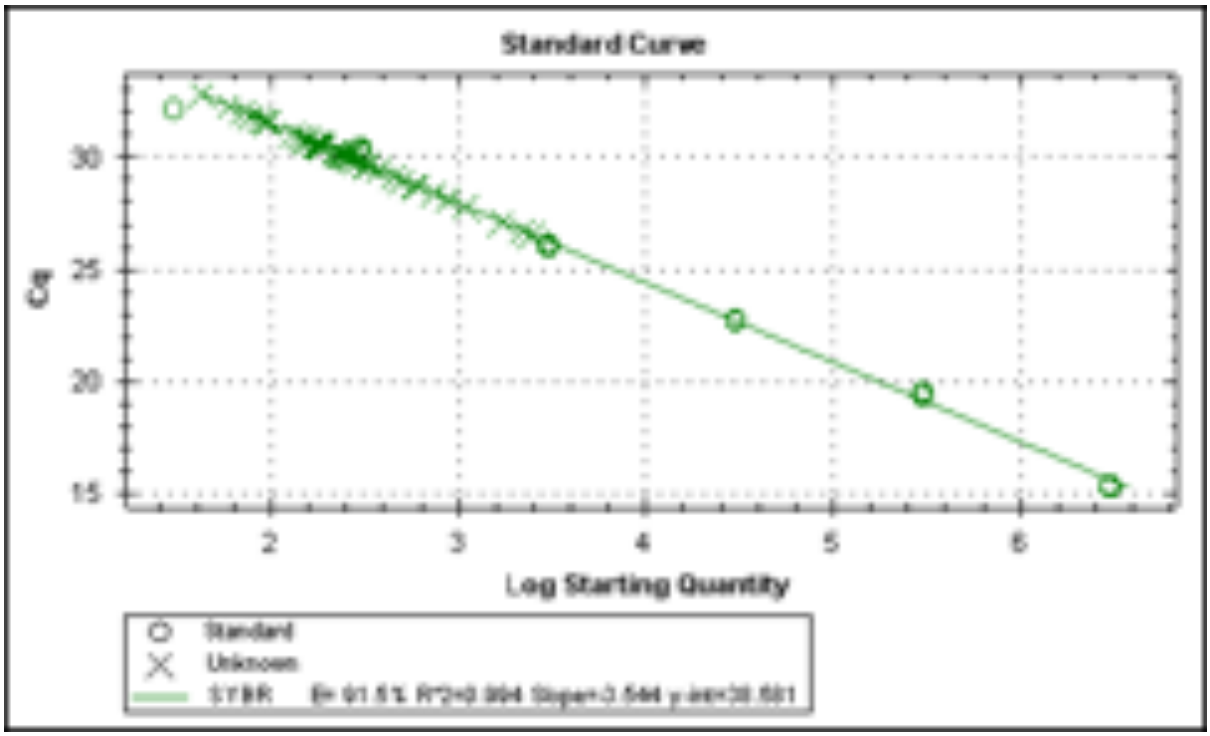
Appendix II – Published work

Cuthbertson L, Amores-Arrocha H, Malard AL, Els N, Sattler B, Pearce AD. Characterisation of Arctic Bacterial Communities in the Air above Svalbard. Biology. 2017;6(2)

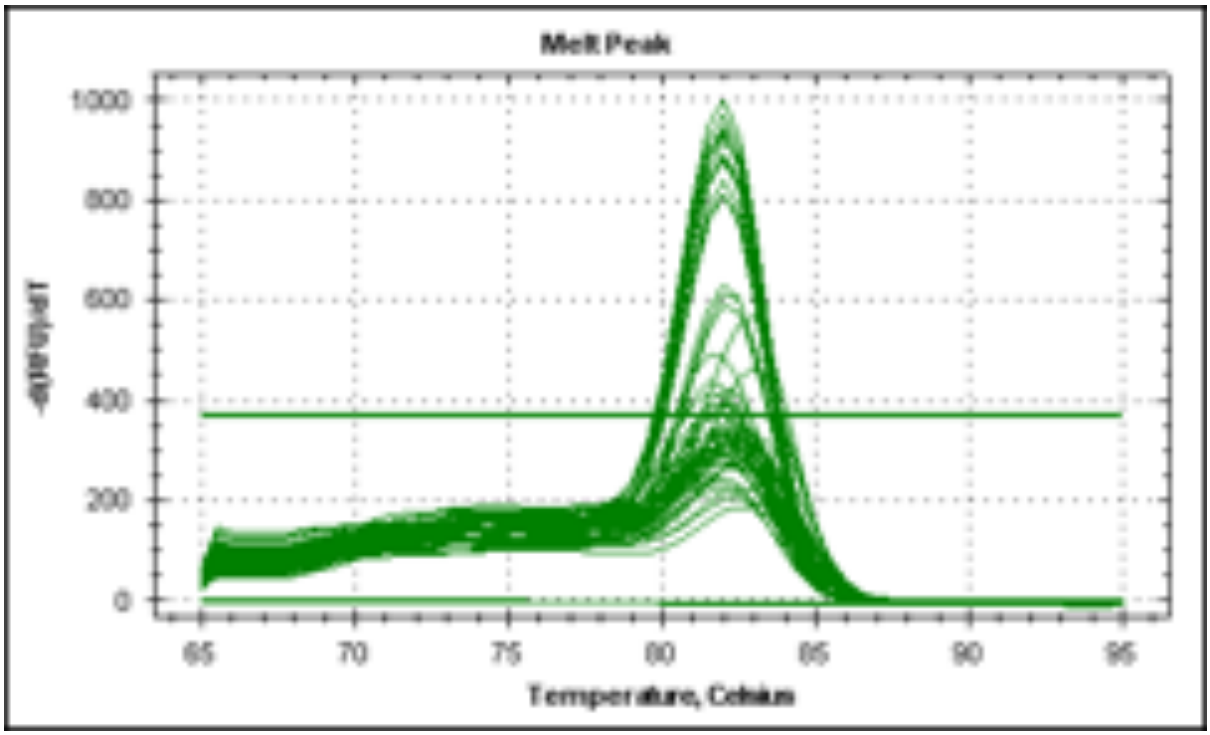
Cuthbertson L, Pearce DA. Aeromicrobiology. Psychrophiles: From Biodiversity to Biotechnology: Springer; 2017. p. 41-55

Pearce DA, Alekhina IA, Terauds A, Wilmotte A, Quesada A, Edwards A, et al. Aerobiology Over Antarctica – A New Initiative for Atmospheric Ecology. Frontiers in Microbiology. 2016;7:16

Appendix III – Qiagen Powersoil vs Qiagen Powersoil Powerlyzer comparison qPCR data



Standard curve for qPCR assay comparing the two kits.



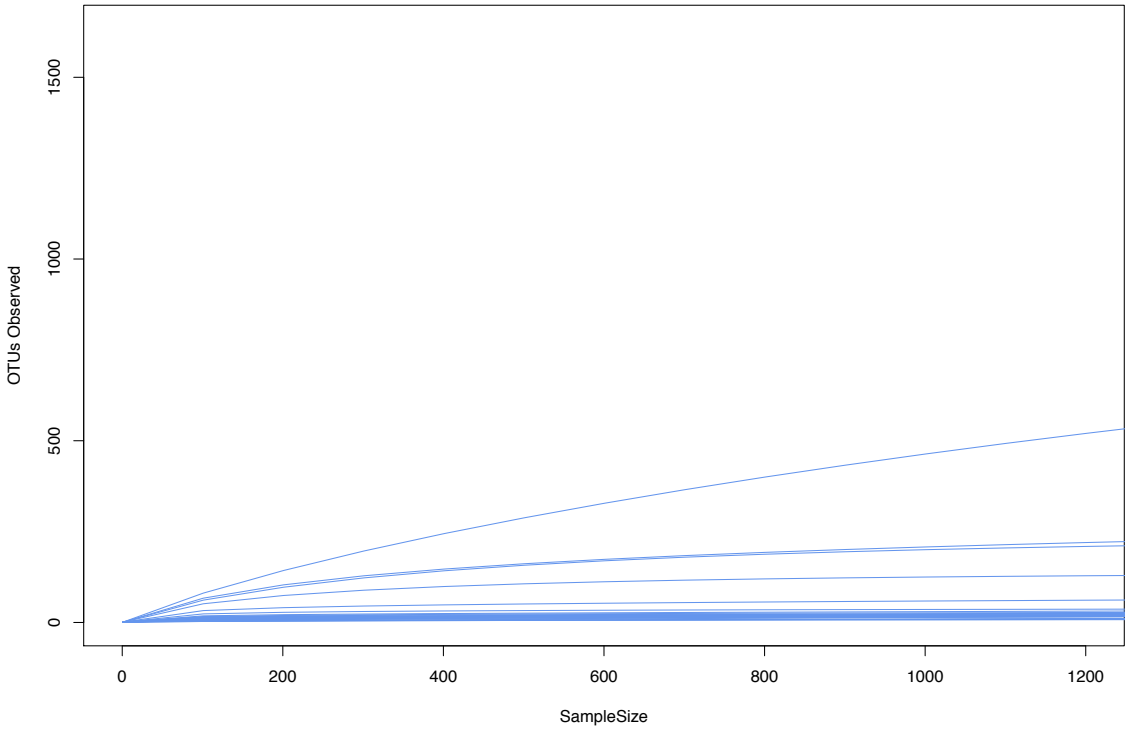
Melt curves for the qPCR assay comparing the two kits.

	16S rRNA copies per mL	Approximate CFU per mL^{-1}
Powersoil kit negative	43591	10379
Powersoil sample	69764	16611
Powerlyzer kit negative	33001	7857
Powersoil sample	70273	16732

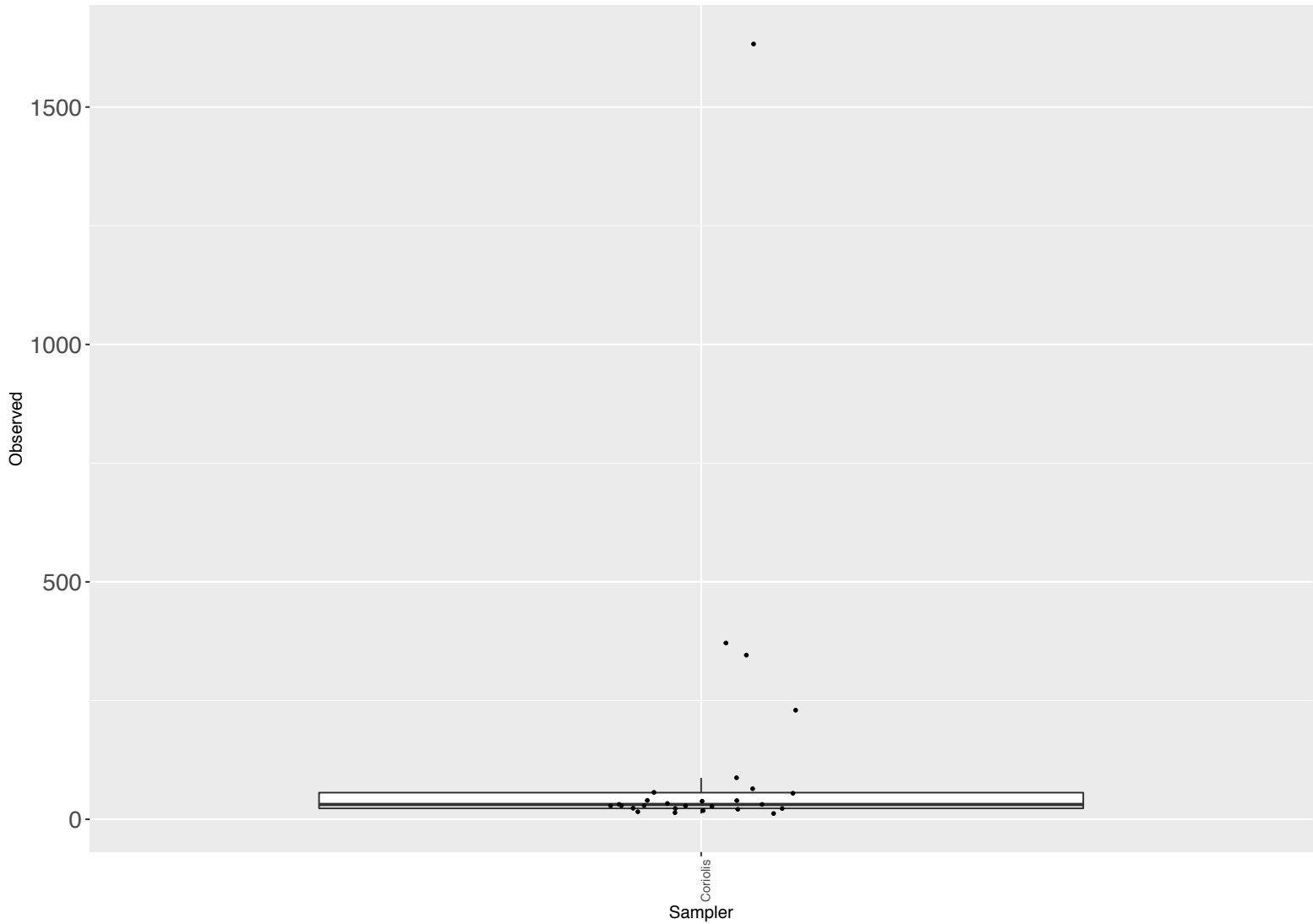
Average biomass of extracts using Powersoil vs Powerlyzer for a 3H air sample collected at Northumbria University.

Appendix IV - Coriolis U data for Antarctic air samples

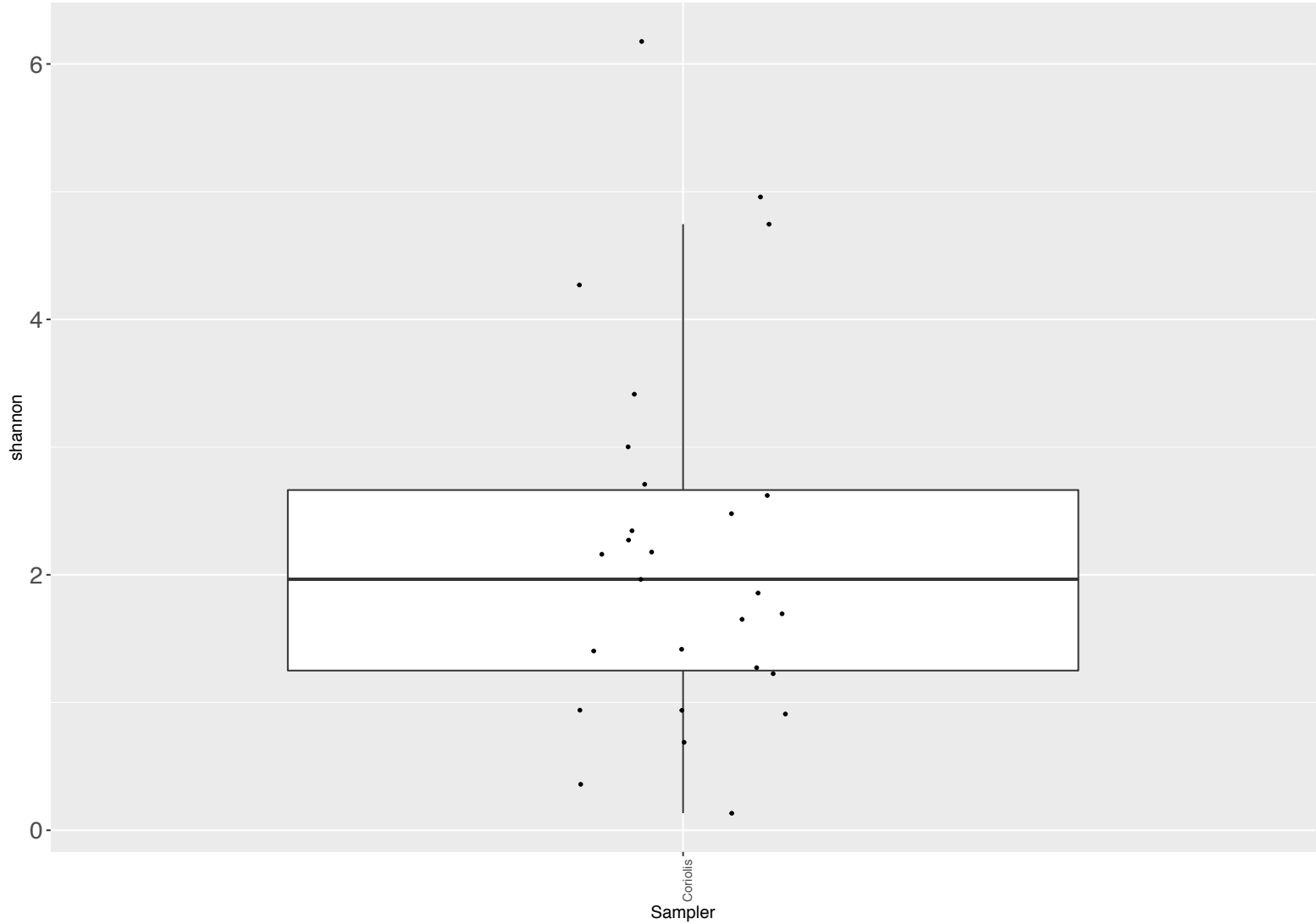
27 of 34 Coriolis U samples were retained following denoising, contaminant screening, removal of non-target taxa and removal of low read samples. The 27 samples contained a sum of 426678 reads, with 15803 on average. The smallest sample had 1245 reads whilst the largest sample contained 83967. The median observed OTUs for all samples was 31, whilst the median Shannon index was 1.97. No core community was present at sequence level, with no taxa present in more than 50% of all samples. The mean estimated cells per Coriolis sample was 3.3E+04 CFU per mL^{-1} based on qPCR data for 11 Coriolis samples.



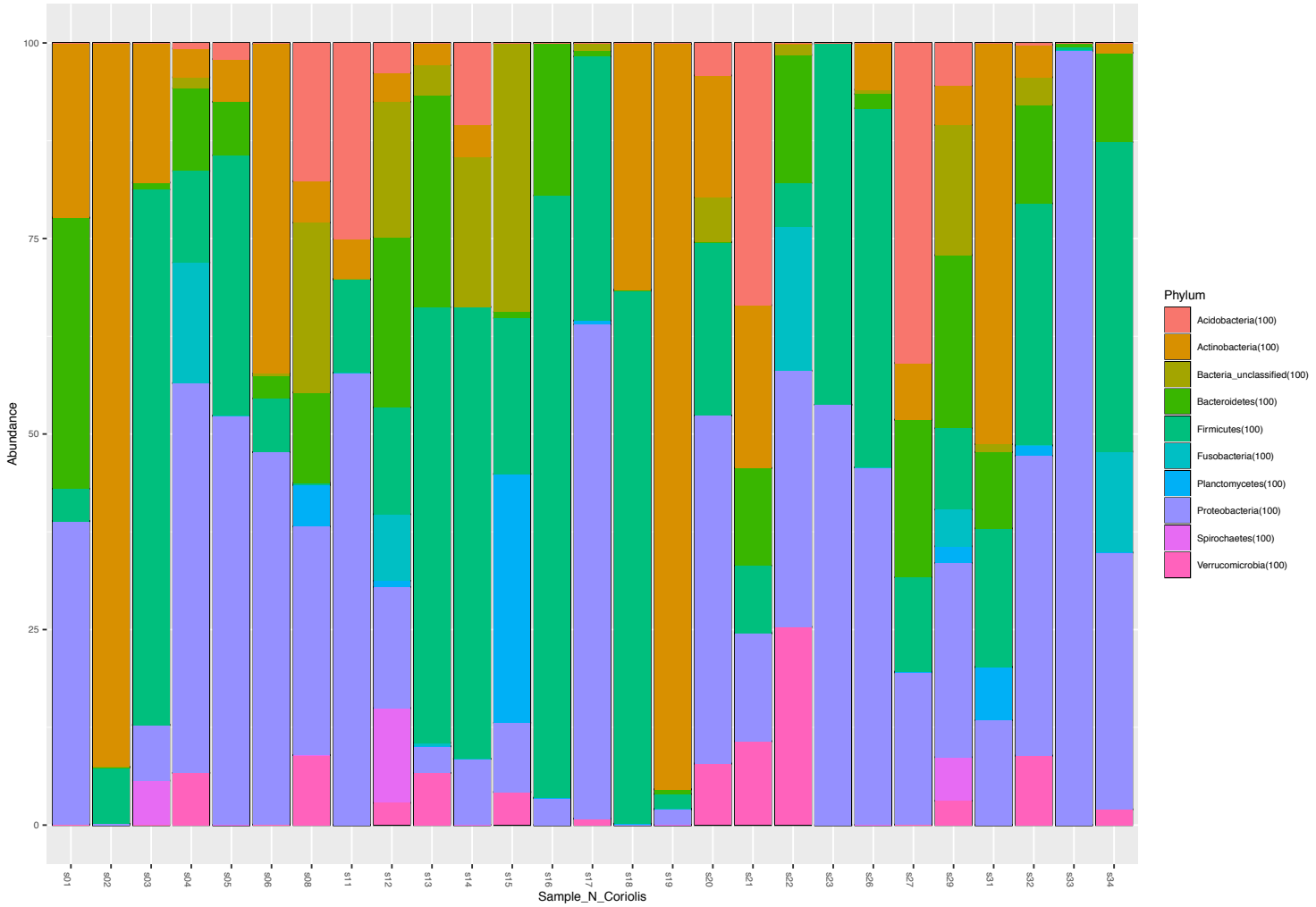
Rarefaction curves for Coriolis data



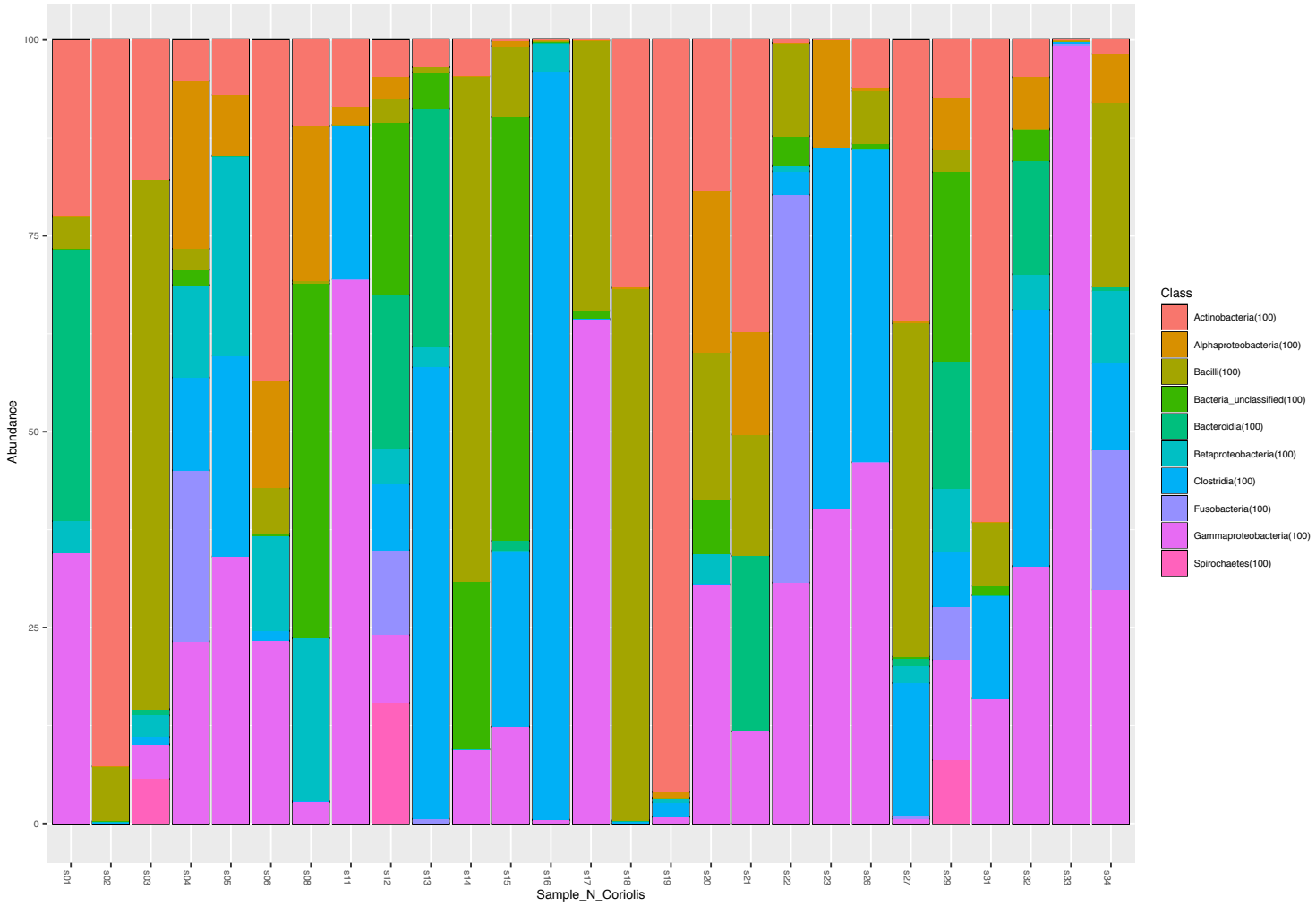
Observed OTUs for Coriolis data



Shannon index for Coriolis data



Relative abundance of the top 10 most abundant phyla in Coriolis samples. *Sample_N_Coriolis does not equate to Sample_N for membrane filtration samples.



Relative abundance of the top 10 most abundant classes in Coriolis samples. *Sample_N_Coriolis does not equate to Sample_N for membrane filtration sample