Invasions of the oropharynx: microbiome of healthy and infected respiratory tissue

Asha Rani Ram

Submitted in fulfilment of the requirements for the degree of *Doctor of Philosophy*

September 2016

Institute of Biodiversity, Animal Health and Comparative Medicine
College of Medical, Veterinary and Life Sciences
University of Glasgow

Abstract

Research aiming to understand pathogens in infection is shifting rapidly towards considering not only the individual pathogen but the whole microbial community. Therefore, understanding microbial communities through exploring the key questions in community ecology, such as the relationship between diversity and stability, are relevant here also. Research has made considerable progress in characterising microbial communities of different body sites but the human oropharynx microbiome is still among the less well known despite its importance in hosting various commensal bacteria and being an important entry site for pathogenic intrusion. Determining the healthy oropharynx microbiome will allow comparison to various disease scenarios and the attributes that change a community from a healthy to diseased state.

This thesis represents the most comprehensive survey of looking at the longitudinal bacterial community structure in the oropharynx. Here, analysis was done on the bacterial oropharynx microbiome composition, its natural fluctuations and stability, and relating these to the changes that occur to the microbiome before, during and after an infection. This involved initial swabbing of the oropharynx of eighteen baseline-healthy, non-smoking participants weekly for a total period of 9 months and sequencing the V1-V2 region of the 16S rRNA gene using Illumina MiSeq sequencing. This would determine the community make up that is representative of a healthy state. This was then directly compared to oropharyngeal samples taken weekly from 12 smokers within the same age range for a total period of 6 months to observe the community differences between smokers and non-smokers.

Looking at the healthy participants (non-smokers) alone, the key taxa recovered were *Firmicutes* at phylum level and *Streptococcus, Prevotella* and *Veillonella* at genus level; these were the most abundant taxa in healthy samples. There was variation in taxa within and between participants, but this variability in microbial community structure occurred more at genus and OTU level. Variability was influenced by changes in health status, although environmental factors were also likely to play a role even though they were not investigated here. Disturbances to the oropharynx microbiome were shown in participants that had cold-related symptoms (negative for viruses) and antibiotic treatment.

These communities had decreased diversity (as opposed to high diversity healthy communities) and changes in abundances of certain taxa. However, participants recovered quickly from these disturbances (within one week after the disturbance) in that the microbiome returned to a state similar in community composition prior to the disturbance. This showed the oropharynx microbiome of baseline-healthy participants to be relatively resilient and stable as samples from the same participants were similar on a weekly basis.

Looking at smokers, they had distinct changes in the bacterial community of the oropharynx in comparison to non-smoking healthy participants. This included changes in abundance of taxa with increased *Bacteroidetes*, *Proteobacteria and Actinobacteria* at phylum level and *Streptococcus* at genus level and increased abundances in pathogenic microorganisms such as *S. pneumoniae* which overall affected the functions associated with the bacterial community. These communities also appeared stable (regardless of having an altered state) in that samples from smoking participants were also similar on a weekly basis, but interestingly, were only disrupted during antibiotic treatment and not during an infection from samples with cold related symptoms.

Therefore this thesis provides insight into the oropharynx microbiome of healthy participants (non-smokers) and smokers. It examines the stability and resilience of the oropharynx microbiome during specific scenarios and identifies the key and important taxa in a healthy and unhealthy community. By continuing to develop this research it may be possible to identify, treat and restore respiratory diseases by examining the oropharynx microbiome through identification of taxa and functions.

Table of contents

List of tables

List of figures

Abbreviations

CV = Coefficient of variation

DMSO = Dimethly sulfoxide

DNA = Deoxyribonucleic acid

GI = Gastrointestinal

HMP = Human Microbiome Project

IBD = Inflammatory bowel disease

KEGG = Kyoto encyclopedia of genes and genomes

LCBD = Local contributions to beta diversity

LMM = Linear mixed model

NGS = Next generation sequencing

NIH = National Institute of Health

NMDS = Non-metric multidimensional scaling

OTU = Operational taxonomic unit

PCR = Polymerase chain reaction

Publications in progress

Longitudinal study of the pharyngeal microbiota and its resilience to viral and antibiotic perturbations

A Ram[1], U Z Ijaz[2], C Quince[3], T J Evans[4] & J Lindström[1]

[1]Institute of Biodiversity, Animal Health & Comparative Medicine, University of Glasgow, Scotland, UK;

[2]School of Engineering, University of Glasgow, Scotland, UK;

[3]Warwick Medical School, University of Warwick, England, UK;

[4]Institute of Infection, Immunity & Inflammation, University of Glasgow, Scotland, UK

Acknowledgements

Firstly I would like to thank my supervisors for making this research and thesis possible, in particular to Dr Jan Lindström for his continued help, support, honesty and encouragement. I really value and appreciate all the advice, guidance and freedom I was given in conducting this research and hopefully in making me a better researcher. I would also like to say a special thanks to Dr Umer Ijaz for his patience and help throughout my project. I am truly grateful for all those hours spent working together and making R bearable for me, as well as always responding to my emails quickly regardless of how silly the question was. Thanks are also due to Prof Tom Evans for his insight, feedback and useful discussions throughout the project and to Dr Chris Quince for his expertise and helpful advice. I am extremely grateful to have worked with such talented individuals.

I would like to acknowledge various people who helped in aspects of this research, namely the technicians Julie Russell and Anne McGarrity for their immense help in completing my laboratory work as well giving me advice and support when I was stressed out or having a bad day. Thank you for making the engineering laboratory such a fun and friendly place to work. I would also like to thank the Centre of Genomic Research (University of Liverpool) for doing my sequencing namely Dr John Kenny, Dr Pia Koldejivic and Dr Linda D’Amore for their contributions. Thanks are also due to Dr William Weir, Dr Konstantinos Gerasimidis, Prof Sarah Cleveland and Dr Sarah Haig for the various discussions and questions that challenged me and made me think outside the box. I am also grateful for the Lord Kelvin and Adam Smith Scholarship for the financial support received.

This PhD would not have been bearable without all the friends I have made during these 4 years, in no particular order Amy Sinclair, Quyen Melina Bautista De Los Santos, Stephanie Connelly, Melanie Schirmer, Maria Catalina Sevillano Rivera, Lorna Mulvey, James Minto and Erifyli Tsagkari, with special thanks to my office friends Kathryn Allan, Elaine Ferguson and Yi-Hsui Chen for all the fun times.

Finally I want to acknowledge my family (especially my mum) for their love and support; their belief in me and continuous praise gave me the confidence and motivation to complete this project even when I couldn’t see the end or believe in myself. On that note, I would like to dedicate this thesis to my eldest sister Narita Ram for giving me countless advice over the last 4 years and always telling me “eventually everything just falls into place”. I hope this work makes you proud.

Author's declaration

I declare that, except where explicit reference is made to the contribution of others, that this dissertation is the result of my own work and has not been submitted for any other degree at the University of Glasgow or at any other institution.

1 Introduction

Microorganisms are microscopic living organisms including bacteria, viruses, protozoa and archaea. Many are abundant in the human body, but only recently have the composition, structure and function of the bacterial components of these communities at different body sites been investigated (Cho & Blaser, 2012). Techniques used previously to study microbiology such as culturing were limited, with many bacteria (especially anaerobic bacteria) remaining uncultivable due to requiring different growth conditions or long incubation periods (Jones, 2009). However, advances in culturing due to formation of complex and nutrient rich media have now made it possible to culture various anaerobic bacteria from the gastrointestinal (GI) tract (Browne et al., 2016). There has also been an increase and expansion in non-traditional molecular methods such as DNA sequencing (Petrosino et al., 2009) resulting in fast and less laborious detailed investigations and analysis of microbial communities in the human body, where it has now been discovered that each different body site or niche is home to millions of microorganisms living as a community. In order to develop understanding of these complex ecosystems it is essential to explore and investigate the microbial diversity and variation in a healthy and diseased state which will create the basis for subsequent analyses such as identifying key taxa responsible for shaping the structure and function of these communities.

1.1 Investigating the microbiome

The term microbiome is a relatively new one, first coined by Joshua Lederberg to describe any ecological community of commensal, symbiotic and pathogenic microorganisms that share our body space (Hooper, 2001). Microbiota studies refer to the identification of bacteria whereas microbiome projects identify bacteria, genes and genomes as well as environmental conditions of the community. The collection of genes and genomes within a community is also known as the metagenome. However the term microbiome is increasingly being used in studies that also only refer to the microbial community and so for this reason the term microbiome was specifically used in this project to characterise the bacterial community and to identify the predicted functions associated with 16S rRNA gene datasets. Most microbiological studies have historically focussed on investigating the disease causing microorganisms found in the human body, with little recognition of the benefits of the residential bacteria. With the completion of the human genome sequencing project in 2001 (International Human Genome Sequencing Consortium, 2001) discussions regarding a second human genome project arose that would detail the microbial genes and genomes at particular body sites (Relman & Falkow, 2001) giving an insight into the role of endogenous microorganisms in healthy individuals. The benefits of such a project would be to understand the organisation of microbial communities all over the human body and their potential influence on health. These initial studies and findings were pivotal in determining the aims and procedures of the Human Microbiome Project (HMP).

1.2 The Human Microbiome Project

In 2008, the National Institute of Health (NIH) led a 5 year wide scale research project to determine the components of microbial communities at various body sites with the aim to understand the roles of the human microbiome. The NIH described the concept of the microbiome as the entire community of microbes that inhabit the human body, their genetic elements and their environment (Petrosino et al., 2009). The main aim was to characterise the human microbiome and its role in health. To achieve this aim, healthy individuals were recruited for sampling each individual microbiome (Figure 1.1) to determine if individuals shared a core microbiome at the lowest taxonomic level that was dependent on body site (species level), and to understand if changes in the human microbiome within and between individuals could be correlated with changes in human health.

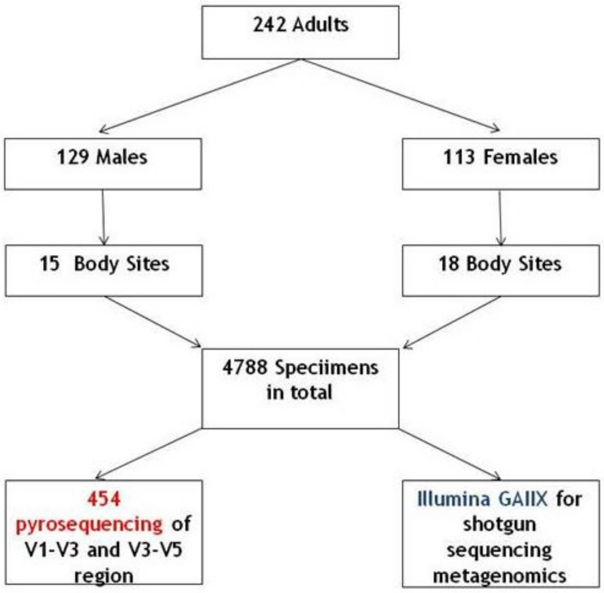

242 Adults
129 Males
113 Females
15 Body Sites
18 Body Sites
4788 Speciimens in total
454 pyrosequencing of V1-V3 and V3-V5 region
Illumina GAIIX for shotgun sequencing metagenomics

Figure 1.1 - The workflow process involved in recruiting and determining the microbiome of various body sites in the Human Microbiome Project (HMP). The areas tested included 9 oral sites, 4 skin specimens, 1 nostril sample and 1 stool sample, with 3 additional samples collected from the vagina in women. Sequencing of the 16S rRNA gene was performed using 454 pyrosequencing whereas Illumina platforms for metagenomics analysis were selected to explore community function.

The HMP also had various goals over the course of the project that would enable them to develop a reference set of microbial genome sequences, explore and develop new tools and technologies for computational analysis and examine the ethical, legal and social implications involved in studying the microbiome. To investigate the microbiome, millions of DNA sequences were analysed through taxonomic assignment and clustering to identify operational taxonomic units (OTU) at the lowest taxonomic assignment, usually genus or species level. An example is oligotyping, a process which allows investigation into the diversity of closely related bacteria through determining variations in the 16S sequences (Eren et al., 2013). It is a supervised computational method that investigates and reveals the microbial diversity concealed within OTUs by focussing on the variable sites in sequences that contain the most discriminating information. This uses Shannon entropy rather than pairwise sequence similarity to discard low-entropy nucleotide positions providing ecological information of microbial communities.

This investigation produced extensive datasets of which the main results are summarised: T*he microbiota consists of 10-100 trillion symbiotic microbial cells* (Cho & Blaser, 2012) (Segata et al., 2012). Each body site contains a vast number of microbes living as a complex microbiota that vastly exceeds the number of human cells. *Body sites showed differences in microbial richness* (Huse et al., 2012). At the lowest taxonomic classification level of identifying bacteria to genus and species level, the oral site had the most number of shared OTUs between healthy people whilst the vagina and skin had the least number of shared OTUs when looking at the numbers and percentages of OTUs; an OTU is classified as a specific type of bacterium based on sequence similarity, usually at a cut off level at 97% identity (Schloss & Westcott, 2011). *The abundance of microbes and diversity of communities on each body site varied widely amongst healthy individuals* (Huse et al., 2012). Even though key taxa were present at each body site (taxa that were always the most abundant) there was still a lot of variation in these abundances within and between healthy individuals. This individual variation of bacteria can be a result of natural variation, as well as external factors such as diet and lifestyle choices (The HMP Consortium, 2012). *A core microbiome representing health was not found* (Huse et al., 2012). A core microbiome was defined as OTUs being present in 100% of samples. However a healthy state could not be described by identifying the OTUs (at species level) as very few shared OTUs were found across all subjects. *The functions of communities at different body sites were not influenced by microbial abundances* (The HMP Consortium, 2012). Even though there was great variation in microbial abundances in body sites across individuals, the metabolic pathways remained the same suggesting that a community of microbes (rather than individual species) are responsible for function. This was shown by Arumugam et al., (2011) where healthy individuals were classified into three different enterotypes based on the taxa found in their GI tract. These enterotypes were each dominated by *Bacteroides*, *Ruminococcus* or *Clostridiales* and *Prevotella* at the genus level and all three enterotypes had the same metabolic pathways such as carbohydrate and amino acid metabolism (Turnbaugh & Gordon, 2009). This research showed that even though the healthy GI tract had variability in taxa in individuals, this did not interfere with the functions provided by that community. This study highlighted the importance of investigating the abundances of taxa present in addition to the function of communities, but to also understand that variation between individuals is still representative of a healthy state.

These initial findings led to a massive increase in microbiome related projects where studies explored taxa compositions in specific disease scenarios. The microbiome may be responsible for influencing various processes such as metabolism (Turnbaugh & Gordon, 2009), immunity (Lozupone et al., 2012) and resistance to pathogens (Gao et al., 2014), all of which are beneficial to the host. But there is now evidence to also link taxa abundances and microbiome changes in various disease states as shown in Table 1.1 which describes associations of bacteria with specific diseases. These studies collectively highlight the importance of studying the microbiome in health and disease as correlations between microbial abundances and disease have been established.

Table 1.1 – Microbial abundances associated with specific diseases.

Disease types	**Microbial abundances in disease**
Childhood onset asthma	*H. pylori* absent in stomach (Blaser et al., 2008)
Colorectal cancer	Increases in *Fusobacterium spp* in colorectal tumour tissue (Warren et al., 2012)
Crohn's disease	Increase in *Enterobacteriaceae* in biopsies taken from the terminal ileum and rectum (Gevers et al., 2014)
Irritable Bowel Syndrome (IBS)	Decrease in *Bacteroidetes* and increase in *Enterobacteriaceae* in faeces (Distrutti et al., 2016)
Periodontitis	Higher diversity in communities and increase in *Spirochaetes* from oral cavities (Abusleme et al., 2013)
Psoriasis	Increased ratio of *Firmicutes* to *Actinobacteria* on skin (Zhan et al., 2008)
Obesity	Increased ratio of *Firmicutes* to *Bacteroidetes* in faeces (Turnbaugh et al., 2009)
Schizophrenia	Increase in lactic acid bacteria in oropharyngeal communities (Castro-Nallar et al., 2015)
Vaginosis	Higher diversity in communities and a reduction in *Lactobacillus* (Ravel et al., 2013)

Therefore, identifying the taxa at different classification levels in the microbiome of a specific body site is only a starting point for understanding its influence on the host. Further challenges exist in considering the biotic interactions, community assemblage and evolution within these communities (Castro-Nallar et al., 2015) (Fierer & Lennon, 2011). In addition, the variation present between healthy individuals and its relevance must be investigated to fully determine the roles of the microbiome in health and then compare this in disease situations.

1.3 Understanding the ecology of microbial communities

Microbial ecology investigates communities of microorganisms living together in a specific environment and the interactions between them and their environment (Konopka, 2009). One goal of ecology is to measure, understand and predict biodiversity and function of an ecosystem whereas microbiome projects aim to identify the microorganisms present and how their functions influence the host. With the increasing microbiome data available there is a growing interest to apply ecological theory to analysis and thereby gain an improved understanding of why these communities have the structure and functions observed. This section will explore the common and increasing ecological terms and theories applied to microbiome studies.

The microbiome can affect the host organism in many ways by influencing attachment of secondary colonised microorganisms and interacting with pathogenic species, consequently affecting risk of disease. Such conclusions require an understanding of the ecological processes in microbial communities. Resilience is an important term that can be described as measuring the fluctuations in a community and its ability to withstand and recover from disturbances through looking at community composition; stable communities tend to have minimal fluctuations as well as quick recovery from drastic community changes. To illustrate the role of ecology in understanding the function of the human microbiome, the GI microbiome, a reasonably well-known part of the human microbiome, provides a good example. The healthy GI tract consists of two main phyla *Bacteroidetes* and *Firmicutes* (Turnbaugh et al., 2009) forming usually a stable community with the former phylum in dominance. *Bacteroidetes* regulates various metabolic activities such as the breakdown of substrates and carbohydrate metabolism. A disturbance in this proportion such as a decrease in *Bacteroidetes* and increase in *Firmicutes* has been associated with a predisposition to obesity in humans (Turnbaugh et al., 2009) resulting in an unstable and disease prone state. Obese subjects were found to have higher levels of *Firmicutes* and a greater expression of obesity prone genes. Therefore the phylum *Bacteroidetes* may be regarded as a key phylum in shaping the community and maintaining the health of an individual whereas an increase in *Firmicutes* may contribute to decreased diversity and instability. Therefore, community properties such as abundance, biodiversity and stability (Table 1.2) may be key features in linking the human microbiome to health and disease. It is crucial to understand how and to what extent these properties shape a community in health, disease and post disease, and many of these theories can now be tested in microbiome studies (Li & Ma, 2016). The details of important processes differ in different communities but the common theme in an ecological viewpoint is that it focuses on a wider context than on a single species; between-species interactions and/or interactions with the environment are considered even if the whole community is not always of interest.

Table 1.2 – Key terms in microbial community ecology.

Ecological term	**Definition and importance to microbiome studies**
Colonisation resistance	The process of when a bacterial community forms and maintains a barrier for protection from invading pathogens (Robinson et al., 2010)
Community	A group of different microorganisms living together in a particular environment (Konopka, 2009)
Community assembly	Processes that build and shape the community. These processes can include dispersal, diversification, environmental selection and ecological drift (Costello et al., 2012)
Dispersal	Movement of microorganisms across space (Costello et al., 2012)
Diversification	Evolution of divergent ecological traits (Costello et al., 2012)
Diversity	Species richness and evenness present in a community (Fierer et al., 2012). This can be broken down into : **Alpha diversity -** diversity measured in a single habitat or community **Beta diversity** – diversity between habitats **Gamma diversity** – diversity of an area that is composed of many habitats
Ecological drift	The processes of birth, death, colonisation and extinction to determine the diversity and species of local communities that are independent of traits and niches (Rosindell et al,. 2012)
Environmental selection	Role of the environment in shaping the community (Costello et al., 2012)
Functional redundancy	The concept of where the function of a community remains the same after a decline of one type of species resulting in other species to compensate to provide the same function (Lozupone et al., 2012)
Relative abundance	Proportion of a microorganism relative to the total number of microorganisms in a community
Resilience	Rate of recovery after a disturbance to a community, it can be a measure of stability (Robinson et al., 2010)
Resistance	Degree to which a community is unchanged when the environment changes, it can be a measure of stability (Robinson et al., 2010)
Stability	The ability of a community to withstand or recover from disturbances (Robinson et al., 2010)

Diversity is an important measure in microbiome studies that can be calculated in many ways. Alpha diversity can be measured by various indicators that take into account species richness (the number of different species present) and species evenness (the spread of the species present). Taking into account the species richness and evenness is known as the relative abundance. This includes the Shannon-Wiener Index (H) that ranges from values of 0 to 5, with the higher value representing higher diversity. Simpson's Index (D) is a measure of dominance and gives the probability that individuals drawn at random from a large community belong to different species. Values range from 0 to 1, with the higher value representing dominance in the community. On the other hand, beta diversity measures the change in species diversity between communities by calculating the number of species that are not the same in 2 communities. Indices used to calculate beta diversity can measure the similarity and dissimilarity of communities by investigating abundances or presence and absence data. An example is the Bray-Curtis dissimilarity index which uses clustering to determine the dissimilarities between samples using abundance data. This gives values ranging from 0 to 1, with a value of 1 indicating 2 samples not sharing any species and so being extremely dissimilar to each other.

Determining the diversity of communities is important because high diversity communities may prevent extinction of bacteria as well as allowing bacteria to adapt to changes in the community (reduction or elimination of some species) therefore regulating the behaviour and function of communities. Stability on the other hand, refers to the ability to withstand and return from disturbances to a state similar to before the disturbance and stable communities tend to be highly resilient (Fierer et al., 2012). In community ecology, stability has been linked to diversity. Since the seminal work by ecologists MacArthur and Elton, low diversity has been linked to many diseases, whereas high diversity has been linked to a more stable and resilient environment that may be more immune to changes (Richardson & Pysek, 2007). Elton argued that the simplest communities are more vulnerable to invasion (McCann, 2000) suggesting that high diversity communities are more prepared for perturbations due to having more species that will respond differently to the perturbation. If this is the case then diversity may be responsible for providing functional redundancy as a means to protect key processes for community survival (Konopka, 2009). If individual species can contribute to a range of functions, the community as a whole may survive on less diversity when challenged to still maintain stability as long as they have those key species present; also known as the insurance hypothesis (McCann, 2000).

High levels of diversity in communities have been seen as beneficial and favourable to protect communities from disturbances by broadening the sustainable conditions in which the community can endure.

The diversity/stability debate aims to understand whether diversity is associated with community stability. This can be tested by identifying perturbations that disturb the community to test the resilience and ability to resist invasion, whilst observing if any changes to diversity occur throughout the process. Various studies suggest a high diversity community is indicative of a healthy and stable state (McCann, 2000), whereas low diversity communities represent disease and unstable environments. This has been shown in obesity which is linked to a low diversity state (Turnbaugh & Gordon 2009). By having a high diversity community, the community will still be able to maintain the stability and functions of the community. The role of the gut microbiome has also been shown to be important in Inflammatory Bowel Disease (IBD) where a low diversity community has been observed from the intestinal microbiome of IBD patients in comparison to healthy controls. In IBD there are decreases in the abundance of anti-inflammatory species such as *Faecalibacterium prausnitzii* which are known to reduce inflammation by releasing anti-inflammatory cytokines and produce short chain fatty acids such as butyrate (Khan et al., 2012) which the host cannot produce itself. Therefore certain bacteria in the GI tract are needed for metabolism and to maintain a healthy state. By having a high diverse community, it provides insurance to the community that another species can help contribute to the overall functions of the community. However, there are exceptions to this theory as the vaginal tract consists of low diversity communities dominated by *Lactobacillus* which is representative of a healthy but stable state (Ravel et al., 2013). The onset of vaginosis results in a reduction in *Lactobacillus* and an increase in diversity resulting in decreased stability which allows colonisation by other microorganisms and the possibility of other infections. More studies investigating this in the microbiome are required to see if there is a general pattern of diversity and stability linked to health and disease.

1.4 Exploring the human oropharynx microbiome

This thesis investigated the human oropharynx microbiome in two groups of populations, non-smokers and smokers. The oropharynx (the part of the throat immediately below the nasal cavity) was specifically chosen due to the presence of commensal bacteria in addition to being an important entry point for pathogenic bacteria.

1.4.1 The healthy oropharynx microbiome

The human pharynx consists of three main parts as shown in Figure 1.2. The oropharynx is constantly exposed to inhaled and ingested microbes, those cleared by mucociliary mechanisms from the respiratory tract and those contained in saliva, food and water. It is a niche for various microorganisms made up of bacteria, viruses and yeast, with the majority of the community dominated by bacteria. The oropharynx is home to various commensal species, many belonging to the *Streptococcus* and *Prevotella* species but is also a site for many pathogenic bacteria such as *Streptococcus pneumoniae* (Pelton, 2012) *Haemophilus influenzae* and *Neisseria meningitides* (Gazi et al., 2004). The oropharynx is dominated by the specific phyla *Firmicutes* and *Bacteroidetes* with other phyla (*Proteobacteria, Actinobacteria* and *Fusobacteria*) residing at less prevalent numbers (Lemon et al., 2010).

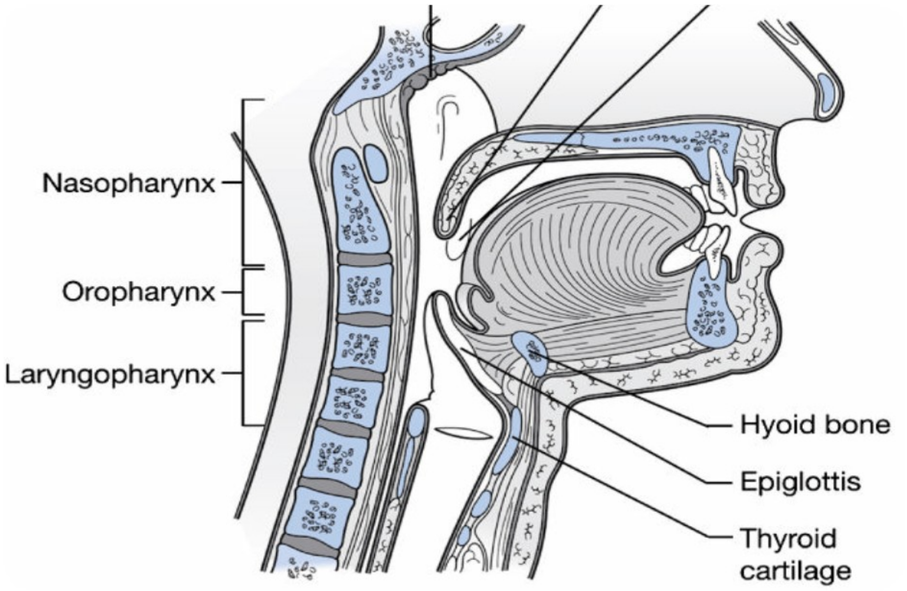


Figure 1.2 – A representation of the different parts of the human pharynx as adapted from Matsuo et al., (2009).

1.4.1.1 Diversity and variation in the healthy oropharynx microbiome

The oropharynx microbiome is a diverse habitat that has previously been shown to include a range of 400-800 different taxa from individuals (Lemon et al., 2010). These individuals varied in the spread of abundance (evenness) in samples but samples from the same individual were shown to be more similar in comparison to samples from other individuals; there was also greater variation in the diversity between individuals. However, individuals did have large intra-personal variation which was increased at further taxonomic levels showing that variation in microbial abundances and taxa within samples is common and representative of a healthy state. This highlights the extent of variation in microbes within and between individuals but also the importance of identifying variation to determine what is considered healthy for individuals.

1.4.1.2 Microbial community functions in the oropharynx

Microbiome studies are now focusing on identifying functions of microbial communities. This gives an indication of the processes the community contributes to in the host in both healthy and diseased states which in turn determines how exactly the microbiome changes during a specific disease situation and what consequences this has for the host. Literature investigating the functions of the oropharyngeal microbial community is limited. Evidence regarding the functional diversity of the respiratory tract has shown microbial communities from the healthy oropharynx to be associated with pathways involved in ATP synthesis and lipid and carbohydrate metabolism (Castro-Nallar et al., 2015). Detailed characterisation of the functions associated with oropharyngeal microbial communities in health is needed to fully understand how variation of microbial communities affects function. This can then be compared against disease scenarios to show how the oropharynx microbiome affects the host during a specific disease and if this impairs function.

1.4.2 The diseased oropharynx microbiome

In order to fully understand the potential role of the oropharynx microbiome in health, it is also required to investigate the oropharynx microbiome in a diseased state. The effects of infections such as the common cold, pharyngitis and tonsillitis on the oropharynx microbiome are currently being investigated through exploring and examining the changes that occur in the microbiome during these infections. These infections tend to be viral; however, pharyngitis and tonsillitis could also be caused by bacterial pathogens. The relationship between viral infections and the oropharynx microbiome is gaining interest and there has been evidence showing Rhinovirus (one of the viruses responsible for the common cold) resulting in increases in *Neisseria* and *Haemophilus* in the nasopharynx (Hofstra et al., 2015) which possibly leads to secondary bacterial infections. In another study, Leung et al., (2012) showed that infection by a certain strain of influenza (pH1N1) resulted in a decrease of commensals *Prevotella* and *Veillonella* and an increase in pathogens such as *Pseudomonas* showing that viral infection does interfere with the host microbiome. However, additional studies are needed to explore this relationship in order to fully understand the role of viral infections on bacterial populations.

1.4.2.1 Can the oropharynx microbiome be a marker for other illnesses?

Understanding the oropharynx microbiome in healthy individuals and its response to a disturbance could potentially be used as a model for disease scenarios. Changes in the oropharynx microbiome (in comparison to a control group) have been discovered in patients with laryngeal cancer. Significant differences were found in abundances of the following organisms with increases in *Fusobacterium nucleatum*, *Fusobacterium* sp. oral taxon and *Prevotella intermedia* with reductions in *Streptococcus* sp. oral taxon and *Streptococcus parasanguinis* (Gong et al., 2014). At the phylum level there was also a significant increase in *Fusobacteria* which is not considered as a dominant phylum of a healthy community.

Changes in the oropharynx microbiome at the phylum and genus level were also observed in individuals who suffered from schizophrenia (Castro-Nallar et al., 2015). Schizophrenic subjects had higher abundances of the fungi Ascomycota and lactic acid bacteria such as *Lactobacilli* (especially *Lactobacillus gasseri*) and *Bifidobacterium*, as well as a reduction in *Neisseria* and *Capnocytophaga*. Functionally, schizophrenic patients had an increased number in pathways related to metabolite transport systems, whereas controls had more pathways involved in energy metabolism (Figure 1.3). This showed that the different microbial communities in controls and patients resulted in different functional abilities. However, whether these functions play a role in exacerbating disease or symptoms still remains to be discovered especially as schizophrenic patients may have completely different behaviours and diets in comparison to control groups. This shows that studies investigating the microbiome in a disease and control group need to be carefully planned and executed to take into account external factors that may influence the results.

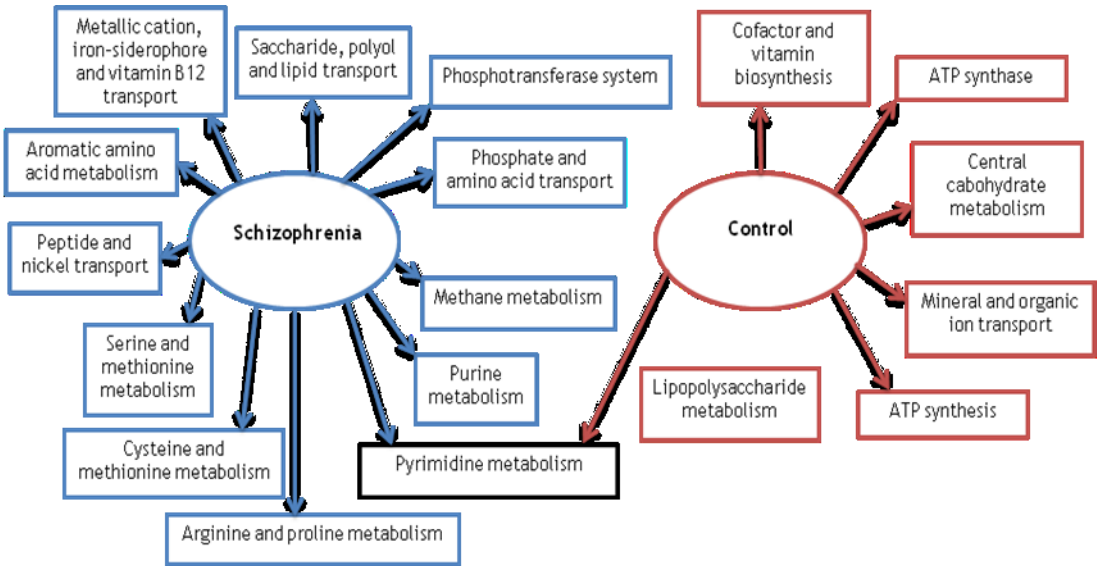

Metallic cation, iron-siderophore and vitamin B12 transport
Saccharide, polyol and lipid transport
Phosphotransferase system
Aromatic amino acid metabolism
Phosphate and amino acid transport
Peptide and nickel transport
Schizophrenia
Methane metabolism
Serine and methionine metabolism
Purine metabolism
Cysteine and methionine metabolism
Pyrimidine metabolism
Arginine and proline metabolism
Cofactor and vitamin biosynthesis
ATP synthase
Central cabohydrate metabolism
Control
Mineral and organic ion transport
Lipopolysaccharide metabolism
ATP synthesis

Figure 1.3 – The most abundant functional pathways present in schizophrenic patients (blue) and the control group (red). The pathway for pyrimidine metabolism was an exception in that it was abundant in both groups.

1.4.3 The effects of smoking on the oropharynx microbiome

The role of the microbiome in disease and the effects of environmental stimuli such as smoking are now being investigated. With smoking increasing the risk of infectious diseases (Bagaitkar et al., 2008) it is only expected that smoking will also alter the microbiome; smoking may promote pathogenic microbial colonisation by disrupting mucocilliary processes and impairing host immune responses against pathogens (Tamashiro et al., 2009). Charlson et al., (2010) reported evidence for the presence of distinct communities in smokers and non-smokers but this has not been done on a longitudinal basis. The effects of cigarette smoke have been explored in the oral cavity with increases in *Parvimonas, Fusobacterium* and *Campylobacter* species, whereas the oropharynx has shown increased abundances in *Megasphaera* and *Veillonella* with decreased proportions of *Fusobacterium and Peptostreptococcus* (Charlson et al., 2010).

Smokers have also shown increased diversity in their samples – there was an increase in pathogens associated with disease, but also microorganisms not previously recognised with disease (Charlson et al., 2010). However more investigation is required when studying the effects of smoking in all microbiomes – there needs to be more focus in how exactly smoking changes the microbiome and whether it is a cause or effect process; does the effects of cigarette smoke actually kill off some species whilst enabling others to survive, or does smoking result in defective host immune responses and increased inflammatory responses which in turn changes the microbial community structure enabling infection to occur. It is also unknown if the effects of smoking on the microbiome are reversible and whether these changes affect the overall function of the community.

1.4.4 Significance of the oropharynx microbiome

It is estimated that 6 million deaths occur globally as a result of smoking and smoking related causes, whereas acute respiratory infections result in an additional 4 million deaths (Ferkol & Schraufnagel, 2014). Smokers also have a higher death rate from respiratory diseases compared to non-smokers (Carter et al., 2015). Understanding how smoking affects the microbial communities over a period of time and the functions associated with the communities is important to investigate on any microbiome. However, in order to fully understand the effects of smoking, there must be an investigation in non-smoking and smoking individuals in both healthy and unhealthy scenarios. Investigating these two groups will be the foundation to determine what changes occur to the microbiome of a non-smoker during a respiratory infection and what changes occur to the microbiome of a smoker during a respiratory infection, as well as determining how long these changes last. Therefore, it is of interest to understand the structure and stability determinants of the oropharyngeal microbial community in non-smokers (both healthy and unhealthy states) and smokers. Therefore characterising the oropharynx microbiome of non-smokers will improve our understanding of why some individuals become colonised with pathogens while others do not. It is still not known how these microbes interact together, the stability of these communities in regards to health and smoker status and what changes occur in these communities on a longitudinal basis before, during and after a disturbance (cold symptoms) or infection. This research could be a starting point in making a clinical impact, where in the future it may be possible to diagnose oropharyngeal disease, restore diseased communities from the effects of smoking or even just give an indication of overall oropharyngeal health status from looking at the microbial populations alone.

1.5 Molecular techniques used to explore microbial communities

Modern microbial ecology requires using molecular techniques. The goal of using such molecular techniques is to study an entire microbial community sampled directly from its natural habitat. DNA-based microbiome studies usually fall into two approaches: either a marker gene (targeted amplicon studies) or the entire metagenome of the community is sequenced. Research presented in this thesis is based on amplicon sequencing of the 16S rRNA gene.

1.5.1 DNA extractions

Before investigating specific genes, genomic DNA must be isolated from bacteria. DNA can be isolated through various commercial kits which either involve bead beating to break down cells and expose the genetic material or a lysozyme step to hydrolyse the peptidoglycan present in the cell walls. The goal of DNA extraction is to have a final volume of DNA which can then be used for molecular processes. However there are various drawbacks and challenges in using DNA extraction kits. This includes obtaining low concentrations of DNA from difficult to lyse samples or low density communities. This can be overcome through optimising the DNA extraction protocol or using the bead beating protocol rather than lyzosyme based kits to ensure higher DNA concentrations. DNA contamination can also be introduced into the sample from any dead external bacteria remaining in the kit, which could be especially problematic for low density communities. This can be overcome through the use of DNA extraction controls where the protocol is run (not using any samples) to help determine the contaminating bacteria present in the kits. The controls are sequenced along with the samples and any reads present just in the controls are manually removed from the samples.

1.5.2 The 16S rRNA gene

The gold standard of bacterial identification and community diversity characterisation is to sequence the highly conserved 16S rRNA gene which is ubiquitous in all prokaryotes (Pace, 1997). The 16S rRNA gene is conserved enough to enable the design of PCR (polymerase chain reaction) primers to target different taxonomic groups, but also has enough variability to provide phylogenetic comparisons of microbial communities (Woese, 1987). All prokaryotes contain the 1500 base pair long 16S rRNA gene for protein production, making it a useful tool in evolutionary studies. The gene consists of conserved regions and 9 hyper variable regions known as V1-V9 (Figure 1.4) varying from 50-100 bases; these sequences are used as targets for microbial identification and can be amplified through PCR. However there are drawbacks in using the 16S rRNA gene as some bacteria have multiple copies of this gene resulting in over amplification during PCR resulting in some bacteria being over or under represented. However, this can be overcome by statistical analyses to take into account over representation of taxa. PCR can also not determine the functions of the gene of interest, as PCR only indicates detection of the gene amplified.

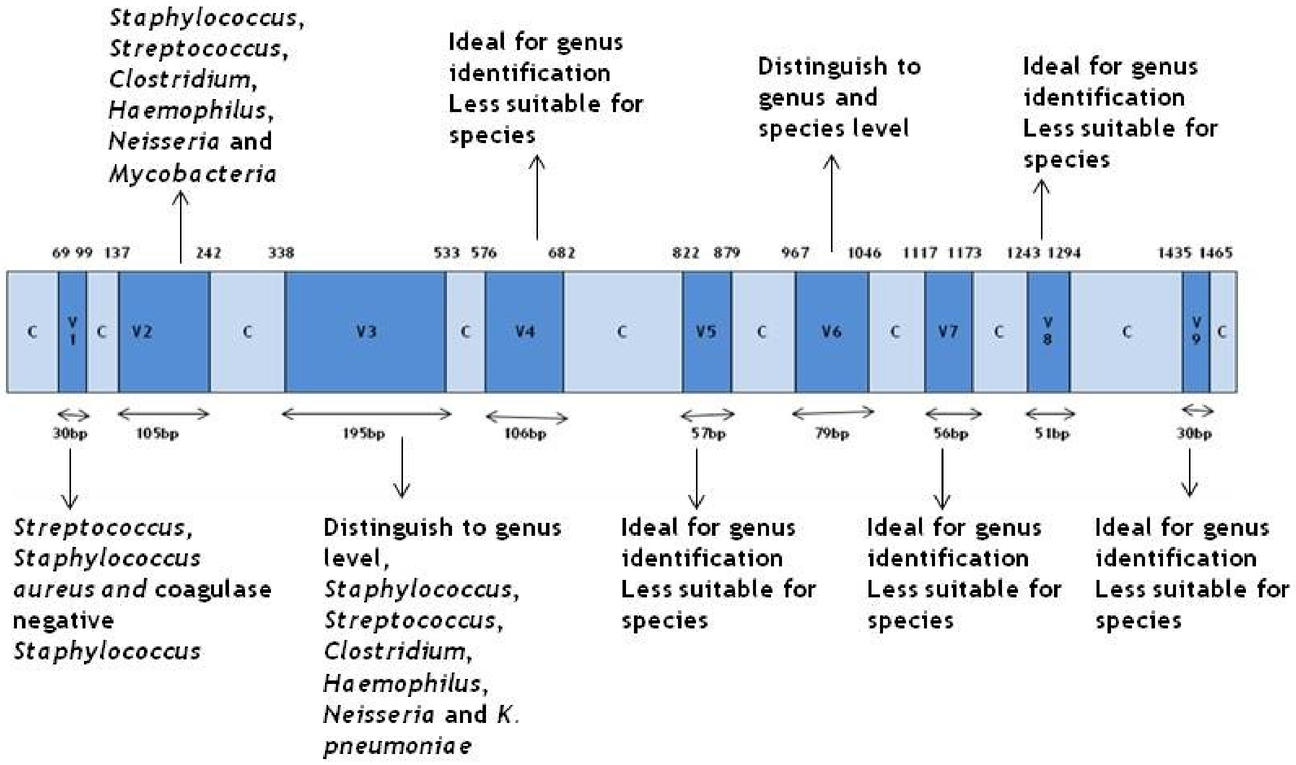

Staphylococcus, *Streptococcus*, *Clostridium*, *Haemophilus*, *Neisseria* and *Mycobacteria*
Ideal for genus identification Less suitable for species
Distinguish to genus and species level
Ideal for genus identification Less suitable for species
69 99 137
242
338
533 576
682
822 879
967
1046
1117 1173
1243 1294
1435 1465
C
V1
C
V2
C
V3
C
V4
C
V5
C
V6
C
V7
C
V8
C
V9
C
30bp
105bp
195bp
106bp
57bp
79bp
56bp
51bp
30bp
Streptococcus, *Staphylococcus aureus and* coagulase negative *Staphylococcus*
Distinguish to genus level, *Staphylococcus*, *Streptococcus*, *Clostridium*, *Haemophilus*, *Neisseria* and *K. pneumoniae*
Ideal for genus identification Less suitable for species
Ideal for genus identification Less suitable for species
Ideal for genus identification Less suitable for species

Figure 1.4 - Conserved and hyper variable regions of the 16S rRNA gene showing what regions are best for bacterial identification (as adapted from Chakravorty et al., 2008).

1.5.3 PCR methods

The PCR protocol was developed in 1983 by Kary Mullis (Bartlett & Stirling, 2003) in order to multiply genes to a significant concentration that is necessary for molecular and genomic analyses. The technique works by producing millions of copies of a desired gene within a few hours as shown in Figure 1.5.

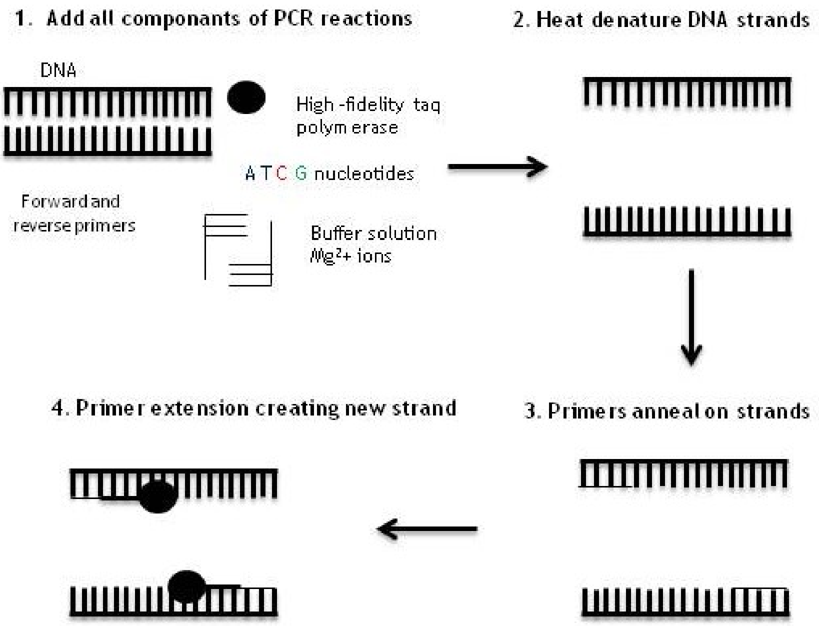

1. Add all componants of PCR reactions
2. Heat denature DNA strands
DNA
High -fidelity taq polym erase
A T C G nucleotides
Forward and reverse primers
Buffer solution
Mg2+ ions
3. Primers anneal on strands
4. Primer extension creating new strand

Figure 1.5 – The stages involved in a simple PCR reaction showing the crucial steps of denaturation, annealing and extension to produce copies of a target gene.

Certain regions of the 16S rRNA gene can be amplified to investigate microbial community structure due its occurrence in all prokaryotes. However, as with all molecular methods, there are biases associated to it and its efficiency is reliant on the quality of the starting DNA, presence of inhibitory substances which can be extracted alongside the DNA, as well as other biases that need to be considered such as template concentration, number of cycles and chimera formation (Janda & Abbott, 2007). However, in spite of this, 16S rRNA gene PCR is a fast, effective and reliable technique used in microbiome studies (Petrosino et al., 2009).

1.5.4 Preparation of a DNA clone library for Sanger sequencing

Clone libraries are a collection of DNA fragments that are stored in vectors each containing a different insert of DNA. Clone libraries are very useful for a preliminary observation into the community make up - the creation of a clone library for a particular gene such as the 16S rRNA gene is one of the most useful and widely used methods for initial community exploration (Leigh, 2010) which can then be used for the production of mock communities as the community make up is already known. Universal primers targeting the 16S rRNA gene are used to amplify the gene of interest. The amplified products are then cloned and inserted into *E. coli* vectors through transformation, with fragments digested with restriction enzymes and separated by gel electrophoresis (Figure 1.6).

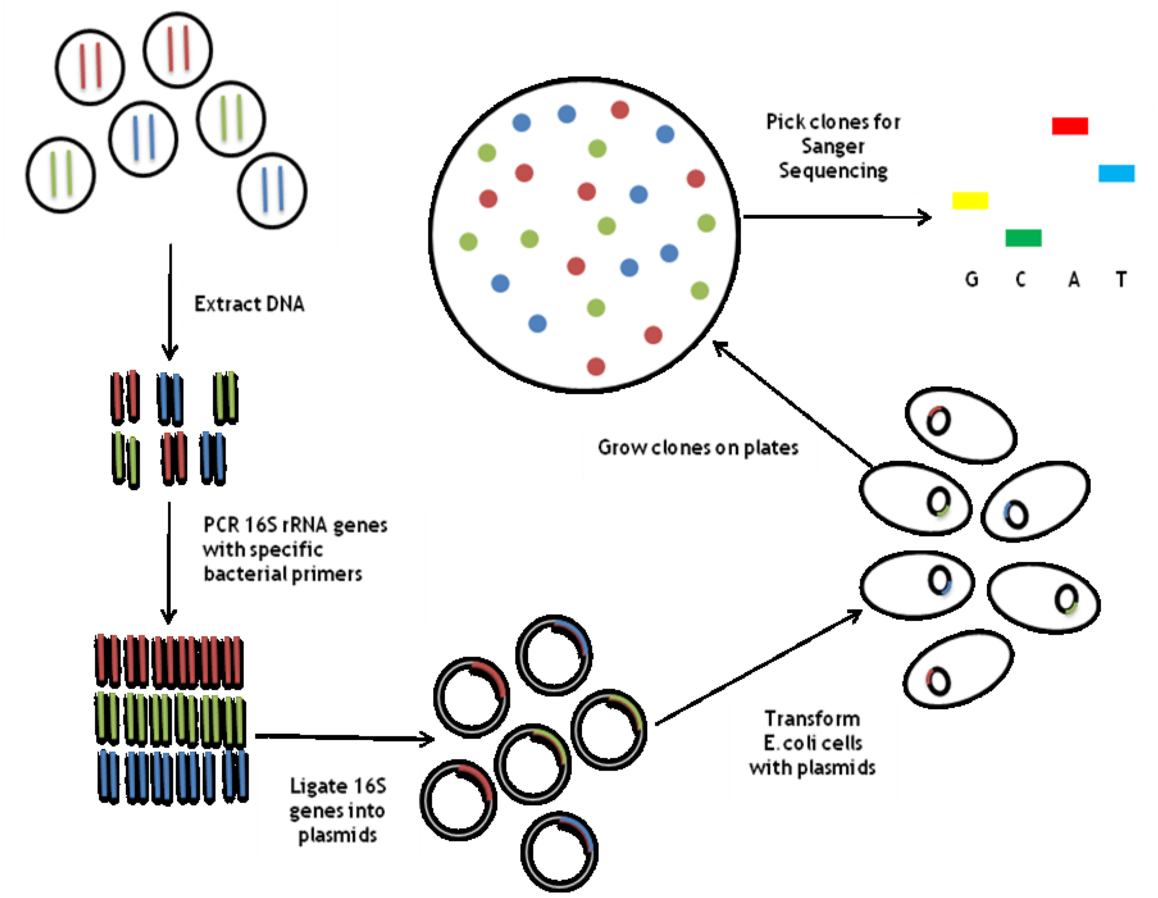

Extract DNA
PCR 16S rRNA genes with specific bacterial primers
Ligate 16S genes into plasmids
Transform E.coli cells with plasmids
Grow clones on plates
Pick clones for Sanger Sequencing
G
C
A
T

Figure 1.6 – The stages involved in preparing a clone library for initial community exploration. The 16S rRNA gene is PCR amplified and cloned into an *E. coli* vector in preparation for Sanger sequencing.

Due to the differences in DNA content, unique banding patterns are generated for each microorganism. Representatives of each different band are identified as a single operational taxonomic unit (OTU) - an OTU is classified as a specific type of bacterium based on sequence similarity, usually at a cut off level at 97% identity (Schloss & Westcott, 2011). Different OTUs are then selected and sent away for Sanger sequencing (Sanger & Nicklen, 1977), a technique that involves replicating single stranded DNA through the use of DNA polymerases to add nucleotides to a growing chain. This is stopped when it randomly incorporates a fluorescently labelled dideoxynucleotide (ddNTP) which results in DNA strands of different lengths. This process is repeated various times which in turn generates a large number of fragments that end in fluorescently labelled bases. The fragments run through a thin glass capillary where an electrical charge separates the fragments by size; the shorter fragments move faster than the longer fragments. The final fluorescent base of each fragment is recorded as it passes through the glass capillary allowing the original DNA sequence to be read.

Sanger sequencing remains a useful method; however there are faster, cheaper and more efficient sequencing methods in use today. Thousands of clones may be required to document the actual richness of the community, hence the preferred choice of next generation sequencing (NGS) techniques. Nevertheless, clone libraries and Sanger sequencing are still beneficial for preparation of mock communities for use as positive controls and quality control for a sequencing run.

1.5.5 Next generation sequencing

DNA sequencing is now routinely used in various fields of study due to the introduction of NGS platforms allowing scientists to sequence a large number of DNA fragments in a single run. NGS platforms have made it possible to recover and characterise genomic material from a wide range of samples, allowing microbial community structure to be explored at a cost effective rate. An example is shotgun sequencing which allows the sequencing of the whole genome by separating the DNA into smaller fragments which can be individually sequenced and then reassembled. The advantages of shotgun sequencing are that it is a fast process that can produce large amounts of data. The disadvantages are that it requires a lot of computing power and errors can occur during reassembly. The reassembly of genomes may also be difficult if there is not a reference genome already available. Different sequencing platforms are in use for amplicon studies, but they all require extraction of nucleic acids, library preparation for sequencing and bioinformatics processing (Vincent et al., 2016). However each platform comes with advantages and disadvantages in terms of their read length, quantity of data, run time and cost (Table 1.3) all of which must be considered when choosing a platform.

Table 1.3 - Comparison of next generation sequencers commonly used today as adapted from Vincent et al., (2016).

Platform	**PacBio**	**Ion Torrent**	**Illumina MiSeq**	**Illumina HiSeq 2000**	**Roche 454 FLX +**	**Minion**
Technology	Phospholinked fluorescent nucleotides	Proton detection	Reversible dye terminator	Reversible dye terminator	Pyrosequencing	Nanopore technology
Read Length (bp)	50% of reads over 10kb	400	300	50-100	600-800	Long read lengths from 5-200kb
Cost per Mb ($)	2	0.75	0.74	0.10	1	Low start up cost
Advantages	No PCR involved Long read lengths	Short run times	Most widely used platform High throughput	Most widely used platform High throughput	Can be used for longer fragments	Small portable instrument Large read lengths
Disadvantages	Fewer reads High error rate	Problems with homopolymer reads Coverage bias	Substitution errors	Substitution errors	Has high homopolymer error rates Expensive	Lower base call accuracy

1.5.5.1 Illumina technology

Illumina platforms are the most popular and economical NGS platform, accounting for 60% of all platforms used (Eisenstein, 2012). Today, five versions of Illumina sequencer are commercially available: HiSeq 2500, HiSeq 1000, Genome Analyser, Genome Analyser IIx and MiSeq. The Illumina MiSeq technology uses a unique paired end strategy where the DNA strand is sequenced from both ends with the forward and reverse reads aligned as a read pair to increase the length of the sequence reads.

The process of MiSeq amplicon sequencing (Hodkinson & Grice, 2015) involves attaching library adapters to DNA fragments (Figure 1.7). The library is loaded onto a flow cell where the adapters from the DNA fragments attach to oligonucleotide adapters on the flow cell. These fragments are amplified locally within clusters to create high densities ready for sequencing, a process known as bridge amplification where the single stranded molecule flips over and forms a bridge by hybridising to an adjacent complementary primer. This in turn forms a double stranded bridge through extension by polymerases, but the double stranded DNA is then denatured resulting in single stranded templates. This process is repeated many times until a high density cluster is formed ready for sequencing. The actual sequencing process uses a reversible terminator based method incorporating one base at a time. Flows of 4 different fluorescently dyed deoxynucleotides (dNTP) are run over the plate and block incorporation once a dNTP attaches to the growing chain which in turn releases a fluorescent signal. After signal detection the dyes are cleared and another cycle of reagents is added as before allowing identification of the DNA sequence. Illumina sequencers are considered the best choice for use in microbiome studies due to the fast improving technology, longer read lengths and reasonable cost.

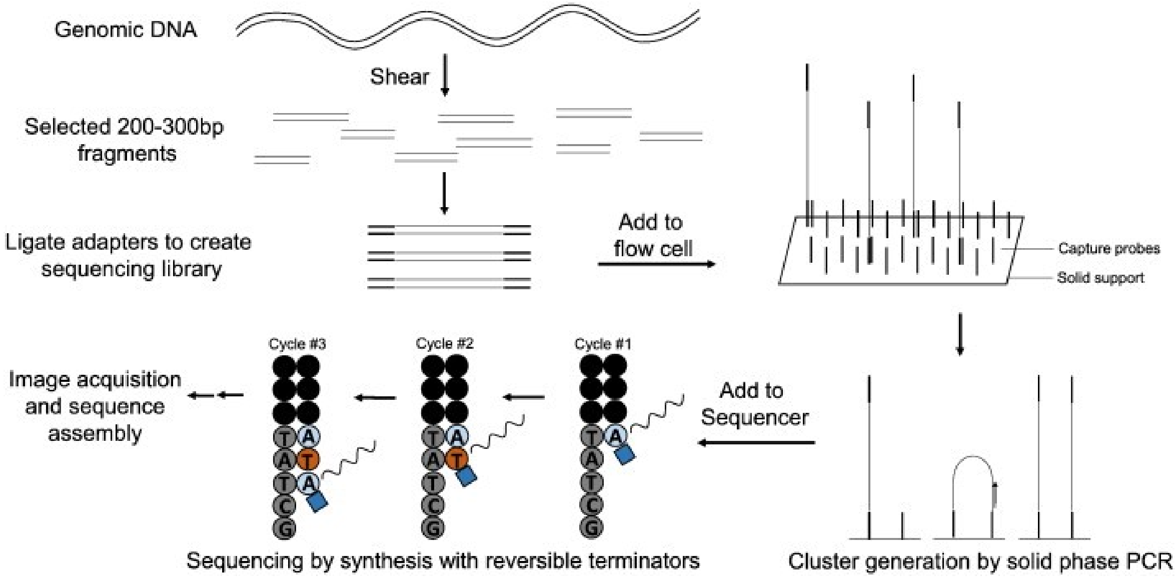

Genomic DNA
Shear
Selected 200-300bp fragments
Ligate adapters to create sequencing library
Add to flow cell
Capture probes
Solid support
Cycle #3
Cycle #2
Cycle #1
Image acquisition and sequence assembly
Add to Sequencer
Sequencing by synthesis with reversible terminators
Cluster generation by solid phase PCR

Figure 1.7 – The process of MiSeq sequencing as taken from Biggar & Storey, 2014.

1.5.5.2 Bioinformatics

Next generation sequencing platforms are now capable of generating a high number of reads in a single run with an increased need for improved software to be able to handle large datasets. A read refers to a data string of nucleotides A, T, C and G that correspond to a DNA sequence. In order to remove low quality reads from true reads, a series of bioinformatics processes are necessary to produce high quality DNA sequences that can be used for statistical analysis. The workflow involves quality filtering of raw sequences which discards reads that do not meet the required quality or length thresholds. Reads are also checked against chimeric sequences. These are artificial sequences that are produced during PCR that do not represent amplicon products. The remaining sequences are then clustered into OTUs against a reference database which assigns DNA sequences to microbial species. There are various programs available that consist of multiple steps to prepare sequencing reads with the most common programs described in Table 1.4. AmpliconNoise was chosen due to decreased error rates, a greater number of reads and longer read lengths in comparison to Mothur (Gaspar & Thomas, 2013).

Table 1.4 – Various bioinformatics programs used in amplicon sequencing projects.

Program	**Description of use**
AmpliconNoise	Used mainly for 454 and Illumina generated sequences (Quince et al., 2011) and consists of 3 main processes: Pyronoise – detects any misreads in sequences SeqNoise – removes PCR mutations Perseus – removes PCR chimeras
Mothur	Open source software package for bioinformatics data processing (Schloss & Westcott, 2011) of raw sequences to OTU construction and phylogenetic construction that can be adapted for processing data from various sequencing platforms including Sanger, PacBio, IonTorrent, 454 and Illumina
QIIME (Quantitative insights into microbial ecology)	Open source pipeline for analysis from raw DNA sequencing data (Caporaso et al., 2011). It can be used on sequences from 454 and Illumina platforms. The processes involve demultiplexing and quality filtering, OTU picking, taxonomic assignment and phylogenetic reconstruction

Quality control steps can also be added through using mock communities to check error rates and ensure the correct sequences are being sequenced. This can also be a way to check for any external contamination introduced into samples through preparing DNA extraction kit controls. Once reads have been filtered and quality checked, rarefaction is then used to assess the species richness from sampling depth and reads coverage to determine if all species within the community have been sampled. The reads can then be used for data analysis. This can include analysis investigating the genetic distance between the DNA sequences and testing variables in a statisitical model to investigate whether there are any correlations in taxa abundance and specific metadata. However there are many issues to take into account when analysing sequence data. This includes ensuring there are enough reads present in each sample for sufficient sampling depth which could especially be a concern in samples with low reads. It is also important to determine what statistical tests to use and this will depend on whether data follows a normal distribution or not; data that has unequal samples sizes, a very small size or does not follow a normal distribution would require non-parametric tests. However there are disadvantages of using non-parametric tests some of which include losing some data and these tests not being as powerful in comparison to using parametric tests. Microbiome data can be visually analysed through various clustering and ordination methods with the statistical testing performed from using univariate and multivariate analysis. However, testing for more than one variable in microbiome studies is now extremely common especially in hypotheses that concern the effects of treatments as well as various factors on bacterial communites. Therefore multivariate analysis is useful for studies assessing the association of many variables with the microbiome in human health.

1.5.5.3 Metagenomic techniques: assigning function to communities

The majority of microbiome studies focus solely on the 16S rRNA gene, even though it is now possible to sequence all genes from a community. The benefit of this is to gain an insight into the overall function of a community as well as the potential functional properties of individual members. This technique is known as metagenomics and provides greater and richer data in the description and quantification of genes in a microbial community. The increase in metagenomic studies has also made it possible to assign predicted functions to communities from using only sequences from the 16S rRNA gene that have been previously assigned to a functional pathway from earlier metagenomic studies. As the 16S rRNA gene is a powerful marker gene as well as more cost effective than performing whole metagenome sequencing, this gene can be used to predict the functional capabilities of microbial communities based entirely on 16S rRNA gene datasets. This is the basis of the Tax4Fun package available in R (Aßhauer et al., 2015) that is used to estimate the metabolic profile of a metagenome based on taxonomic abundance estimates and references using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (a collection of databases dealing with identification of genomes and biological pathways). This package is useful for initial exploration of the functions which could then be further investigated by metagenomic sequencing. This package has also been shown to have greater accuracy in predicting functions of 16S genes in comparison to PICRUSt (Aßhauer et al., 2015). However the disadvantages of using these packages is that it only predicts function from DNA sequences and not from RNA or proteins which are actually used to measure gene expression.

1.5.6 Implications of using next generation sequencing in microbial ecology

The use of NGS in microbial ecology is constantly advancing and expanding at a fast rate. However, various issues have occurred. The main issues with NGS technology include DNA extraction quality, PCR amplification, computational power, storage space, cost and analysis. In order to taxonomically assign DNA sequences, it is assumed that the reference database must contain DNA sequences that are correctly identified and annotated. However, current sequence databases may also be limited and not up to date. Although there are various databases such as BLAST, NAST and GenBank to identify and group sequences, the quality of these sequences are questionable. A study conducted by Clayton et al., (1995) showed that 26% of identical 16S rRNA gene sequences in GenBank had random sequencing errors questioning the true accuracy of that sequence representing the labelled species. In order to overcome this problem it is now reasonable to search a query sequence in at least 2 databases or alignment tools to show that the sequence does represent the query sequence. In spite of all this, NGS has and will continue to revolutionise and accelerate biological and biomedical research, allowing scientists to explore complex and new areas of microbiome study.

1.6 Aims and hypothesis

The main aim of this PhD project is to characterise the microbiome of the oropharynx in terms of microbial community structure and temporal stability in non-smoking individuals and smokers, and to identify the key factors associated with infection and recovery from these.

The project will first characterise the oropharynx microbiome in non-smokers. This will determine the baseline oropharynx microbiome from healthy samples in non-smokers in terms of occurrence and relative abundance of different phyla, genera and OTUs. This will then be compared to unhealthy samples from non-smokers to determine what happens to the healthy community during a disturbance, infection and antibiotic treatment. The stability of the oropharynx microbiome in non-smokers will also be addressed. This will determine how much temporal variation is there in the oropharynx microbiome, how stable the oropharynx microbiome is, how long recovery takes from a disturbance and whether diversity is linked to stability.

The next section will then compare the non-smoking microbiome to a smoker's microbiome in terms of phyla, genera and OTU abundance. This will also compare the stability of non-smokers and smokers microbiomes to determine if smokers have a longer recovery time from a disturbance compared to non-smokers. Finally the predicted functions of bacterial communities will be characterised in non-smokers and smokers to determine whether smokers have changed functions in comparison to non-smoking participants.

The significance of investigating this work is that the oropharynx microbiome is less defined compared to other body sites but still a very important one to consider. Longitudinal studies are needed to determine the fluctuations that occur naturally and during disturbances in many participants, but to also investigate how the community is constructed first and how it changes. The healthy microbiome can then be established and compared to various disease scenarios and disrupted communities. For this reason, the main hypothesis for this project is as follows – healthy non-smoking participants will have a distinct and stable oropharynx microbiome over time, in comparison to a smoker's microbiome which is expected to be unstable with a different microbial structure.

2 Materials & methods

2.1 Recruitment and consent of participants

The study was approved by the University of Glasgow Ethics Committee (Ethics Application 2012107 & 200140023) and recruitment of participants occurred through mass email targeting mostly students.

2.1.1 Non-smoking participants

Eighteen participants between the ages of 18 - 37 (39% male, 61% female) were recruited on the basis that they were healthy, non-smokers, had no respiratory disease or infection and were not on any long-term medication.

2.1.2 Smoking participants

Twelve smoking participants were recruited (using the same requirements for the non-smokers) between the ages of 19 – 40 (17% male, 83% female) on the basis that they had no underlying health issues and were not on any long term medication.

2.2 Sampling periods and collection

Swabs were collected for non-smoking participants and smokers over two separate time periods. The swabbing period for non-smoking participants started on 20 January 2013 until end of May 2013, recommencing in September 2013 until December 2013 to represent semester times, although some participants continued to hand in swabs over the summer period. The swabbing period for smokers took place from 17 November 2014 to 14 June 2015 giving a sampling period of 30 weeks (excluding Christmas holidays). The smokers sampling period was shorter than the swabbing period for the non-smoking participants due to time and funding restrictions. As non-smokers and smokers were sampled in different years, this could also potentially introduce bias into the project, especially as non-smokers were also sampled in autumn whereas smokers were not. Other factors that could have been different during the two years include weather, pollen levels and levels of circulating microorganisms. Prior to sampling, all participants were given a briefing of the project, a swabbing demonstration and a consent form to sign.

2.2.1 Non-smoking participants

Participants were provided with 2 swabs – Sigma Transwabs in liquid amies (Medical Wire Ltd, UK) for bacterial detection and flocked dry swabs (Copan Diagnostics Ltd, UK) for viral detection. Participants were shown how to take a swab of the oropharynx and were provided with swabs for practice before the official swabbing start date. The swabbing procedure involved washing hands before taking the swab, opening the mouth as wide as possible and touching the swab over the following areas – tonsil, posterior wall to tonsil as shown in Figure 2.1.

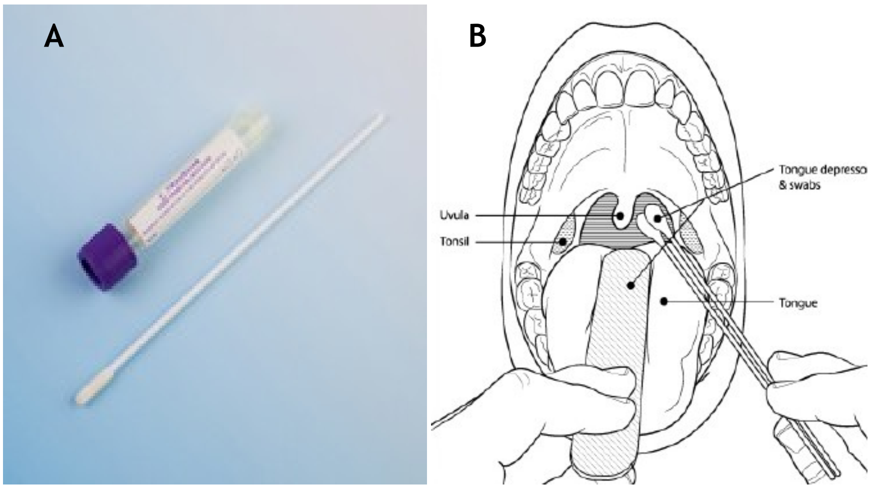


Figure 2.1 – The bacterial swab used for weekly swabbing (Fig. 2.1A) and a diagram showing how to take an oropharynx swab (Fig. 2.1B).

This motion was repeated at least 5 times to ensure all the areas of the swab were used, with the swab then inserted into transport medium ready for storage. Participants were asked to take a swab every Monday morning as soon as they got up and before they had breakfast or brushed their teeth, in order not to disrupt their microbial community. Participants kept a diary stating the time when they took the swab and to record their overall health status. This involved noting if they were healthy or had any illnesses or symptoms, whether there had been any change in their normal routine (on holiday, change in diet or commencing any medication) and if they had touched any other surface in the mouth such as teeth, tongue or cheek. It was likely that participants would become ill during the swabbing period, therefore participants were asked to record this and continue taking swabs throughout. During routine weekly swabbing, participants used the Sigma Transwabs for bacterial detection only. By contrast, if they had a fever, cold symptoms including runny nose, sore throat or a cough, illness, flu or were prescribed any antibiotic treatment participants were also asked to take a dry swab for viral detection (tested for a respiratory screen at Gartnavel hospital). This respiratory screen tested for the following viruses – Influenza A, Influenza B, Respiratory Syncytial Virus A, Respiratory Syncytial Virus B, Parainfluenza, Adenovirus, Rhinovirus, Human metapneumovirus and Coronaviruses NL63, 229E and CC43.

Samples were collected on a weekly basis through a collection box so participants could drop off used swabs and collect new ones. Participants typically took a swab as soon as they got out of bed and all swabs were received before midday. During this period swabs were stored at room temperature in Amies transport medium but were frozen at -20°C if they were not received on the same day of swabbing. Bacterial swabs were processed as soon as possible (typically within 2 hours after collection) for DNA extractions whereas viral swabs were taken immediately to West of Scotland Virology Centre (Gartnavel hospital) for a respiratory screen. Participant metadata including the number of samples received from each participant throughout the sampling period is shown in Table 2.1.

Table 2.1 – Summary metadata for non-smoking participants.

Participant	**Age at time of sampling (years)**	**Sex**	**Total number of samples received**	**Number of unhealthy and antibiotic treated samples**		
				Cold	**Antibiotics**	**Viral**
HA	26	F	34	2	0	0
HB	21	M	2	0	0	0
HC	18	F	16	2	0	0
HD	23	F	24	0	0	1
HE	23	F	7	1	0	0
HF	21	M	22	0	4	1
HG	20	F	8	1	0	0
HI	37	F	31	4	0	2
HJ	21	F	12	1	3	0
HL	18	F	6	0	1	0
HM	18	M	14	3	0	0
HN	19	F	12	1	0	0
HO	30	F	26	1	0	0
HQ	21	M	9	1	0	0
HR	19	M	14	0	0	0
HS	31	M	30	1	0	0
HT	23	M	30	0	0	0
HV	18	F	16	1	0	3

At the end of the sampling period, 313 samples were received in total and samples were broken down into the following groups – healthy (n=279 from 18 participants), cold (n=19 from 12 participants - this includes samples that had the symptoms of a cold but were detected as negative for the standard viruses tested from the respiratory screen at Gartnavel hospital), antibiotics (n=8 from 3 participants – this includes samples positive for antibiotic treatment) and viral (n=7 from 4 participants – this includes samples with confirmed viruses from the respiratory screen). The cold, antibiotics and viral samples were grouped as unhealthy samples. Even though the participants on antibiotics were on treatment for acne and not because of an infection, they were still categorised with the unhealthy group due to disturbing the community structure. The metadata of atypical and unhealthy samples are shown in Table 2.2.

Table 2.2 - Metadata for atypical healthy samples and all unhealthy samples from non-smoking participants (all samples were included in analysis). Oropharyngeal samples are colour coded: green = healthy but with swabbing deviations, black = antibiotics, red = cold related symptoms and blue = viral.

Samples	Health status	Samples	Health status
HA14	Cold symptoms, dry throat	HI39	Cold, sore throat
HA26	Runny nose	HJ2	250mg erythromycin antibiotics twice daily (acne)
HC9	Hit tooth after swabbing	HJ5	250mg erythromycin antibiotics twice daily (acne)
HC11	Touched back of tongue slightly before swabbing	HJ7	250mg erythromycin antibiotics twice daily (acne)
HC18	Cold, sore throat	HJ37	Cold, sore throat
HC44	Cold, sore throat	HL3	Antibiotics
HD9	Rhinovirus - sore throat	HM5	Cold, sore throat
HD27	Swab taken after holiday	HM28	Cold, runny nose
HE12	Cough, sore throat	HM31	Cold, runny nose
HF5	Respiratory synctial virus - fever, sore throat	HN10	Hit tooth during swabbing
HF10	Swab taken on holiday	HN36	Cold
HF13	Swab taken after holiday	HO8	Runny nose
HF20	Tetracyclines for acne	HQ38	Cold
HF23	Tetracyclines for acne	HR14	Dropped swab on desk before swabbing[1]
HF24	Tetracyclines for acne	HS3	Swab taken after breakfast
HF25	Tetracyclines for acne	HS43	Cold, sore throat
HG6	Cough, sore throat	HT12	Swab taken after holiday
HI2	Cold	HV36	Rhinovirus - runny nose
HI14	Sore throat	HV37	Rhinovirus - fever, dry throat
HI18	Cold, sore throat	HV38	Rhinovirus - dry cough
HI27	Rhinovirus - sore throat	HV43	Sore throat
HI28	Rhinovirus - sore throat		

[1] Sample included in analysis

2.2.2 Smoking participants

Participants were given the same sampling instructions as described in section *2.2.1*, but were also asked to record the rough number of cigarettes smoked per week, with Monday as the start of the week. Participant metadata from each smoking participant throughout the sampling period is shown in Table 2.3.

Table 2.3 - Summary metadata of smoking participants.

Smoking participant	Age at time of sampling (years)	Sex	Number of years smoking	Total number of samples received	Number of unhealthy and antibiotic treated samples	
					Cold	Antibiotics
SA	40	M	20	23	0	0
SB	19	F	5	3	0	3
SC	19	F	2	23	2	0
SD	19	F	1	8	1	0
SE	19	F	5	1	1	0
SF	19	F	4	19	0	15
SG	33	M	15	25	0	0
SH	30	F	13	25	5	0
SI	30	F	10	24	3	0
SK	19	F	5	10	2	0
SL	19	F	3	10	0	0
SM	19	F	3	6	0	0

At the end of the smokers sampling period, 177 samples were received in total; samples were broken down into the following groups – healthy smokers (n=147 from 12 participants - classified as healthy samples in that there were no symptoms of disease or infection), cold (n=14 from 6 participants - this includes samples that had the symptoms of a cold but were detected as negative for the viruses tested from the respiratory screen at Gartnavel hospital) and antibiotics (n=18 from 2 participants – this includes samples positive for antibiotic treatment). No viruses were detected from the viral swabs. Cold and antibiotics samples were grouped as unhealthy samples as described before with the metadata of atypical and unhealthy samples from smoking participants shown in Table 2.4.

Table 2.4 - Metadata for atypical and unhealthy smoker samples (green = healthy with swabbing deviations, black = antibiotics and red = cold related symptoms).

Samples	Health status	Samples	Health status
SB2	Lymecycline (408mg) one per day (acne)	SF19	Tetralysal (300mg) once a day (acne)
SB3	Lymecycline (408mg) one per day (acne)	SF22	Tetralysal (300mg) once a day (acne)
SB4	Lymecycline (408mg) one per day (acne)	SF23	Tetralysal (300mg) once a day (acne)
SC3	Cough, sore throat	SF24	Tetralysal (300mg) once a day (acne)
SC4	Cough	SF25	Tetralysal (300mg) once a day (acne)
SD2	Cold symptoms	SH2	Touched tongue
SE3	Cold symptoms, sore throat	SH4	Sore throat
SF9	Tetralysal (300mg) twice daily (acne)	SH5	Sore throat, cold symptoms
SF10	Tetralysal (300mg) twice daily (acne)	SH8	Cough
SF11	Tetralysal (300mg) twice daily (acne)	SH9	Cough
SF12	Tetralysal (300mg) twice daily (acne)	SH18	Cold, sore throat symptoms
SF13	Tetralysal (300mg) once a day (acne)	SI4	Flu-jab received
SF14	Tetralysal (300mg) once a day (acne)	SI13	Cold, sore throat
SF15	Tetralysal (300mg) once a day (acne)	SI14	Cold, sore throat
SF16	Tetralysal (300mg) once a day (acne)	SI16	Cough
SF17	Tetralysal (300mg) once a day (acne)	SK3	Cold symptoms, sore throat
SF18	Tetralysal (300mg) once a day (acne)	SK14	Chesty cough

2.3 DNA extractions

DNA was extracted using the QIAamp DNA Mini kit (Qiagen Ltd, UK) following the bacteria, swab and tissue protocol (Biesbroek et al., 2012) (Salter et al., 2014) – Appendix 1. A negative extraction (containing no sample) was performed each time the kit was used. These negative extractions were then sequenced to ensure minimal contamination from the reagents in the kit into samples. Extracted DNA was quantified using the Qubit and picogreen HS DNA assay (Invitogen Ltd, UK). A volume of DNA (5µl) was mixed with 2µl of loading dye on a 1% agarose gel (1g agarose to 100ml TBE) along with a 1Kb Invitrogen DNA ladder and ran at 100v for 60 minutes to check presence. The DNA was then stored at -20°C until required.

2.4 16S rRNA gene PCR

Due to the variability in DNA concentration extracted swabs (a range from 0.2-50ng/µl), 16S rRNA gene PCR reactions were set up to ensure bacterial DNA was present. A 25µl reaction was set up with the following reagents – 12.5µl Bioline PCR Mix (Bioline Ltd, UK), 1µl of forward primer (12.5pmol), 1µl of reverse primer (12.5pmol), 2µl of DNA template (ensuring a DNA concentration of 10-15ng depending on initial DNA template concentration) and 8.5µl of DNAse free water. The universal prokaryotic 16S primers used were 27F (5'-GAGTTTGATCCTGGCTCAG-3') and 1392R (5'-ACGGGCGGTGTGTRC-3').

The PCR amplification cycle was carried out at the following conditions:
initial denaturation – at 95°C for 5 minutes, 30 cycles of denaturation at 94°C for 1 minute, annealing at 62°C for 1 minute and extension at 72°C for 1 minute. A final extension was carried out at 72°C for 10 minutes with a holding stage at 4°C. A 1% agarose gel was prepared to run 10µl of amplicon product along with a 1Kb Invitrogen DNA ladder at 100v for 60 minutes to ensure correct length of the amplicon product which represents the targeted 16S region (Figure 2.2). A positive control was set up using DNA from the mock community. A negative control was also set up to show no DNA was present in the PCR reagents and resulting in no amplification of the 16S rRNA gene.

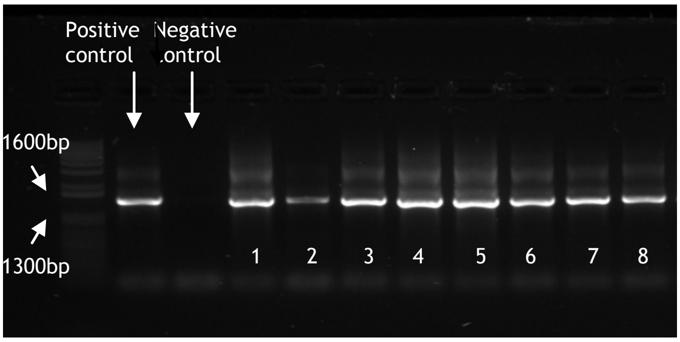


Figure 2.2 – Diagram of a 16S rRNA gene gel showing the 1365bp amplicon product (labelled as samples 1-8) against a 1Kb Invitrogen DNA ladder.

2.5 Preparation of a clone library for Sanger sequencing

A clone library was prepared using the Invitrogen Topo-Seq Kit (Invitrogen Ltd, UK) as a quality control step to produce a mock community for future Illumina MiSeq runs. The purpose of the mock community was to act as a positive control for each MiSeq run to ensure that the correct sequences were being sequenced. DNA from 10 participants was mixed together in equal concentrations (3µl of 5ng) for preparation of a clone library. The DNA was amplified for the 16S rRNA gene using specific bacterial primers and the amplicon product was loaded onto a 1% agarose gel along with a 1Kb Invitrogen DNA ladder and run for 60 minutes at 100V. The full protocol of performing a clone library is shown in Appendix 2.

Each type of OTU (operational taxonomic unit) that had a unique banding pattern after restriction enzyme digest (as observed on the gel) was assumed to be a different species and was sent for Sanger sequencing to Source Biosciences Ltd (Cambridge, UK). In total 96 clones were analysed of which 38 different banding patterns were identified and were labelled as different OTUs; these OTUs were sent for sequencing. The forward and reverse strands were trimmed at each end using a program called DNA Dynamo Sequence Analysis Software (BlueTractorSoftware Ltd, UK) to produce a contig sequence which were then checked against BLAST (Altschul et al., 1990) to identify each OTU to species or genus level; sequences were identified to species level at a >97% identity cut off. The mock DNA community was prepared by diluting each OTU to 13ng/µl and using 3µl of DNA from 26 chosen OTUs; these were high quality sequences that had an overlap of at least 20 base pairs when joining the forward and reverse sequences to produce a contig sequence.

2.6 Optimisation of 16S rRNA PCR

The V1-V2 region of the 16S rRNA gene was chosen for amplification and sequencing. The significance of this region was to try to identify bacterial sequences to genus and OTU level and to differentiate the different types of *Streptococcus* species in the oropharynx as research showed the genus *Streptococcus* as a dominant member of the healthy oropharyngeal community (Charlson et al., 2010) (Charlson et al., 2011).

Primers were provided by the Sanger Centre (Cambridge, UK) and were specific golay barcoded primers covering regions 27F (AGMGTTYGATYMTGGCTCAG) and 338R (GCTGCCTCCCGTAGGAGT). The total amplicon product expected after PCR amplification was 398bp including lengths of adapter and linker sequences. Primers were tested in a 25µl PCR assay using the HotStart High Fidelity Taq KAPA kit (Anachem Ltd, UK) with the following components: 12.5µl of readymix, 0.75µl of dNTPs, 1µl of forward primer, 1µl of reverse primer, 1.25µl of dimethyl sulfoxide (DMSO), 6.25µl of nuclease free water and 2µl of DNA (DNA from the mock community at a concentration of 3ng/µl). Optimisation of the protocol included setting up a temperature gradient to determine annealing temperatures of primers, increasing primer concentration from 0.3mM to 0.4mM, decreasing extension time and increasing the number of cycles from 23 to 26 cycles to accommodate for low DNA templates as shown in Figure 2.3.

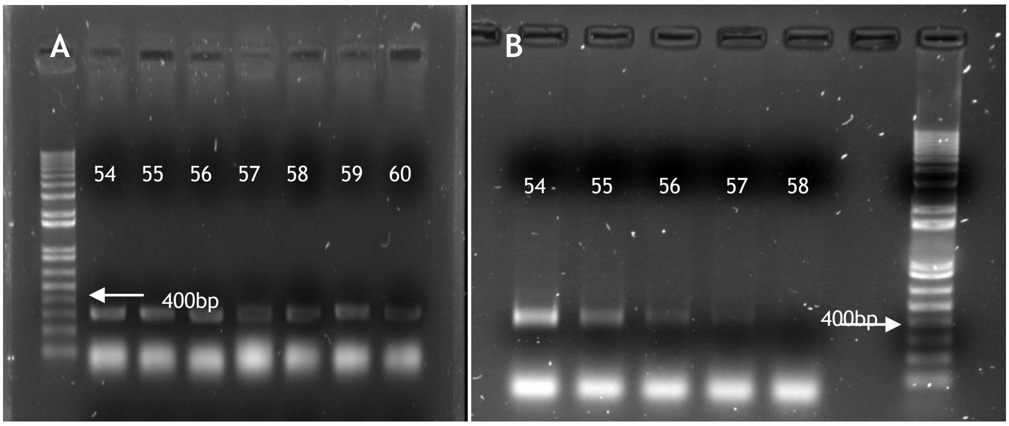


Figure 2.3 – Optimisation gels of the V1-V2 region showing amplicon products between 300bp – 400bp at different temperature gradients (from 54°C - 60°C) at 23 cycles (Fig. 2.3A) and 26 cycles (Fig. 2.3B).

2.7 Amplification of V1-V2 region of 16S rRNA gene

All samples were amplified in triplicate and each sample was assigned a different barcode to ensure identification during the sequencing process. The final optimised PCR reaction (25µl) was carried out at the following conditions: initial denaturation – at 98°C for 5 minutes, 26 cycles of denaturation at 98°C for 20 seconds, annealing at 54°C for 15 seconds and extension at 72°C for 30 seconds. A final extension was carried out at 72°C for 1 minute with a holding stage at 4°C. A 1% agarose gel was prepared to run amplicon product at 100V for 60 minutes with a 1Kb Invitrogen DNA ladder (Figure 2.4) to check length of amplicon product.

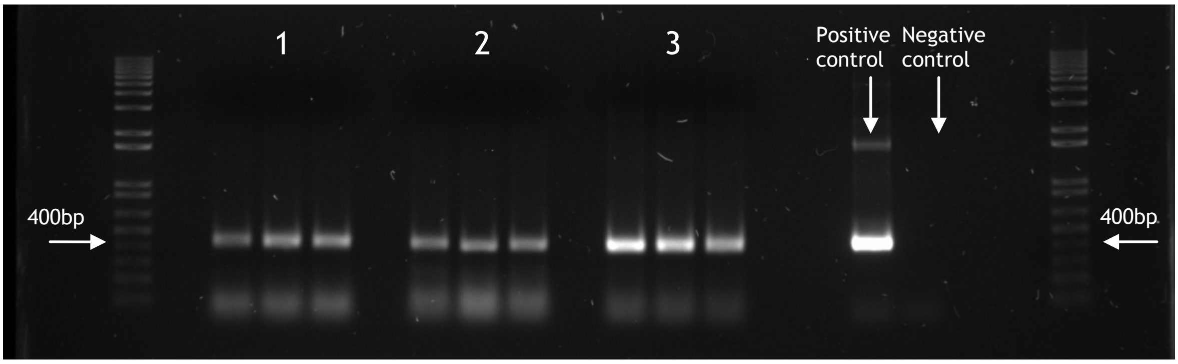


Figure 2.4 - Example of a V1-V2 region amplified gel (samples 1-3 done in triplicate) at an annealing temperature of 54°C at 26 cycles with necessary controls.

2.8 Quality assurance

To reduce PCR biases, all template DNA was diluted to the same concentration and 26 cycles were used in each PCR run to reduce non-specific binding with appropriate controls in place. A positive (DNA from mock community) and a negative control (for each different reverse barcode using nuclease free water) was set up for each PCR run. To ensure that the source of bacterial sequences was not the swab itself or the DNA isolation reagents, PCR was performed on DNA isolated from an unused swab. To confirm that the PCR reagents were not the source of bacterial sequences, PCR of the no-template extraction control was also performed. Neither of these control PCRs yielded products visible on a gel, indicating that there was no or minimal contamination from the swab or reagents. Sequencing of the mock community resulted in 93% matched reads for the forward sequences (error rate of 0.9) and 90% matched reads for the reverse sequences (error rate of 0.9) using the Bioinformatics tool AMPLImock (https://bitbucket.org/umerijaz/amplimock/src) (D'Amore et al., 2016). This is a pipeline used to quantify error and matched rates of a mock community against known reference sequences. Quality assurance was done through presence and absence by identifying species that were present in the mock community but this did not account for species abundances.

2.9 Normalisation and pooling of PCR amplicons for sequencing

After amplification, each sample had a total volume of 75µl that was run on a 1% gel for extraction to clean up the DNA. Gel extractions were performed using the Qiagen Gel Extraction kit (Qiagen Ltd, UK) – protocol shown in Appendix 3. After the DNA clean up, the concentration of DNA was measured using the Qubit (HS DNA assay). Samples with low DNA concentration (less than 2ng/µl) were precipitated using ethanol to increase the final DNA concentration.

Each sample was normalised to a concentration of 2.5ng/µl using nuclease free water. A mixed pool consisting of 130-150 samples was produced using 2µl of each sample. Each pool also consisted of a positive mock community control and a negative extraction kit control (Salter et al., 2014). The pool was checked for the correct length on a 2% agarose gel using 5µl of amplicon product at 90V for 40 minutes with a 1Kb Invitrogen DNA ladder as shown in Figure 2.5. Each pool was then sent for 2 x 250bp MiSeq sequencing at the Centre for Genomic Research at the University of Liverpool to conduct the sequencing.

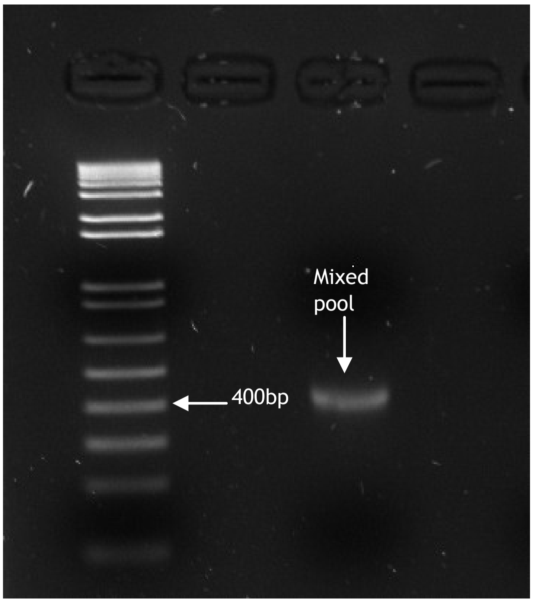


Figure 2.5 – Example of a final gel showing a mixed pool of 130-150 samples (amplicon product roughly at 400bp) against a 1Kb Invitrogen DNA ladder.

2.10 Bioinformatics

Trimming and filtering of paired-end sequencing reads was done using Sickle (version 1.2) by applying a sliding window approach and trimming regions where the average base quality drops below 20 (Joshi et al., 2011). This applied a 10bp length threshold to discard reads that fall below this length. BayesHammer (Nikolenko et al., 2013) was used from the SPAdes assembler (version 2.5) to error correct the paired-end reads followed by PANDAseq (version 2.4) with a minimum overlap of 50bp to assemble the forward and reverse reads into a single sequence spanning the entire V1-V2 region (Masella et al., 2012). The above choice of software showed a reduction in substitution errors by 77-98% with an average of 93.2% for MiSeq datasets (Schirmer et al., 2015). After having obtained the consensus sequences from each sample, UPARSE (version 7.0.1001) was used (https://bitbucket.org/umerijaz/amplimock/src) for OTU construction as described in Edgar, 2013. The approach pools together the reads from different samples and adds barcodes to keep an account of the samples these reads originate from. The reads are then dereplicated and sorted by decreasing abundance and discarding singletons. In the next step, the reads are clustered based on 97% similarity. Even though the cluster_otu() command in usearch removes reads that have chimeric models built from more abundant reads, a few chimeras may be missed, especially if they are present in very low abundance.

Therefore, in the next step, a reference-based chimera filtering step using a gold database was used (http://drive5.com/uchime/uchime_download.html) that is derived from the ChimeraSlayer reference database in the Broad Microbiome Utilities (http://microbiomeutil.sourceforge.net/).

The original barcoded reads were then matched against OTUs with 97% similarity (a proxy for species level separation) to generate OTU tables for different samples. The representative OTUs were then taxonomically classified against the RDP database using the standalone RDP classifier (version 2.6) (Wang et al., 2007). Phylogenetic distances between OTUs were produced by first using MAFFT (version 7.040) (Katoh & Standley, 2013) to align the OTUs against each other and then by using FastTree (version 2.1.7) on these alignments to generate an approximately-maximum-likelihood phylogenetic tree (Price et al., 2010). The OTU table, phylogenetic tree, taxonomic information and metadata were then used in multivariate statistical analysis. A summary diagram showing the stages involved in the bioinformatics process is shown in Figure 2.6.

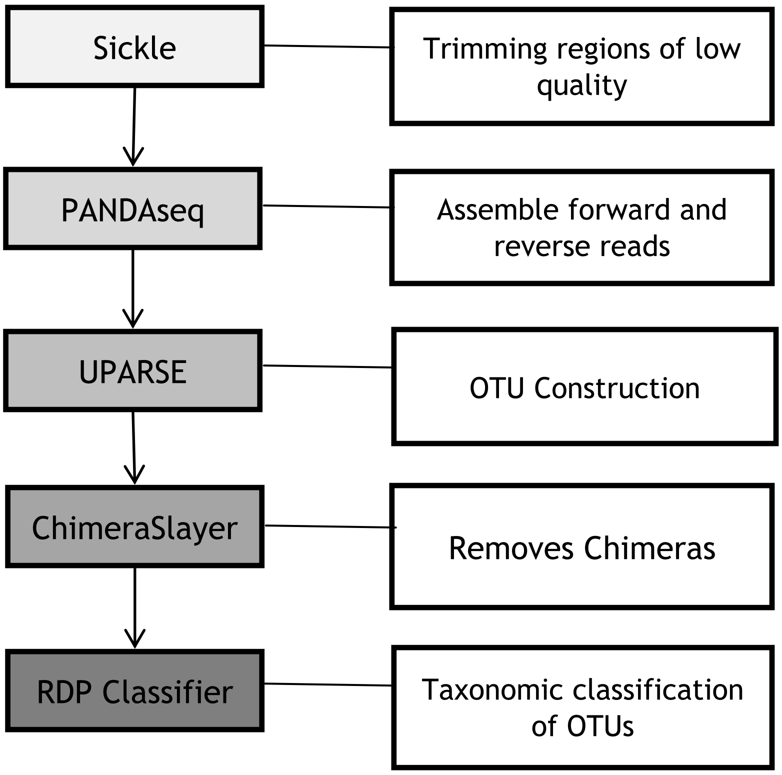

Sickle
Trimming regions of low quality
PANDAseq
Assemble forward and reverse reads
UPARSE
OTU Construction
ChimeraSlayer
Removes Chimeras
RDP Classifier
Taxonomic classification of OTUs

Figure 2.6 – Stages involved in the bioinformatics workflow from raw data to taxonomic classified operational taxonomic units (OTUs).

2.11 Statistical analysis

2.11.1 Initial analysis

Results from samples that contained less than 5000 reads were discarded in the analysis (for non-smokers and smokers) to allow comparison of all samples with enough statistical power. This resulted in 490 samples altogether (313 from non-smoking participants and 177 from smoking participants). The relative abundance of taxa for each sample was calculated by dividing the read counts of that taxon by sample size whereas prevalence was calculated as the percentage of samples containing a given taxa. Abundance was shown as the count of reads (or percentage of reads) belonging to a particular taxon. Statistical analysis was performed in R software (version 3.1.2). Where appropriate before specific analyses, the abundance data was normalised (McMurdie & Holmes, 2014).

2.11.2 Community analysis

Microbial compositional structure was assessed using a non-metric multidimensional scaling plot (NMDS) at genus and OTU level (at 3% divergence) to determine the differences in communities of all samples of non-smokers and smokers. This determined the effects of various variables such as smoker or health status on community composition. Here, Bray-Curtis dissimilarity index was used which considers bacterial taxon abundance. Additionally, the unweighted UniFrac distance analysis from the Phyloseq package (version 1.17.2) (McMurdie & Holmes, 2013) was used which takes into account the phylogenetic distances (relatedness) of the bacterial taxa through presence or absence, without accounting for their proportional representation. A covariance ellipse using ordiellipse() and veganCovEllipse() in Vegan (version 2.4.0) (Oksanen, 2013) was added (95% confidence interval calculated from the standard error of the mean of each group) with the centroid of the ellipse representing the group mean. Covariance for each group was calculated using cov.wt() and the shape of the ellipse was defined by the covariance within each group; the bigger the ellipse, the more variability in community structure in samples within the group.

To find OTUs that were significantly different in abundance between the conditions, the DESeq2 package (version 1.12.3) was used (Love et al., 2014). This uses a negative binomial GLM to model the abundance data (OTU frequencies) and empirical Bayes to shrink OTU-wise dispersions to identify OTUs that have log-fold changes between different conditions (at a cut off value of *P* < 0.01). This is then tested by performing a Wald test on shrunken log-fold changes and adjusting for multiple comparisons showing *P* adjusted values.

Samples were rarefied to the minimum number of reads (5118) to test for alpha diversity. Rarefaction curves were done for each participant showing a sampling depth of 5118 reads could be used for adequate coverage. Even though some samples saturated at a higher cut off (10,000 reads), a sampling depth of 5118 reads was chosen as choosing a higher cut off would result in the removal of too many samples. Rarefaction curves were obtained for each participant to approximate OTUs detected as a function of sequencing depth. The rarefaction curves of selected participants (Figure 2.7) suggest that the total number of observed OTUs in samples from participants vary between 100 and 300 OTUs.

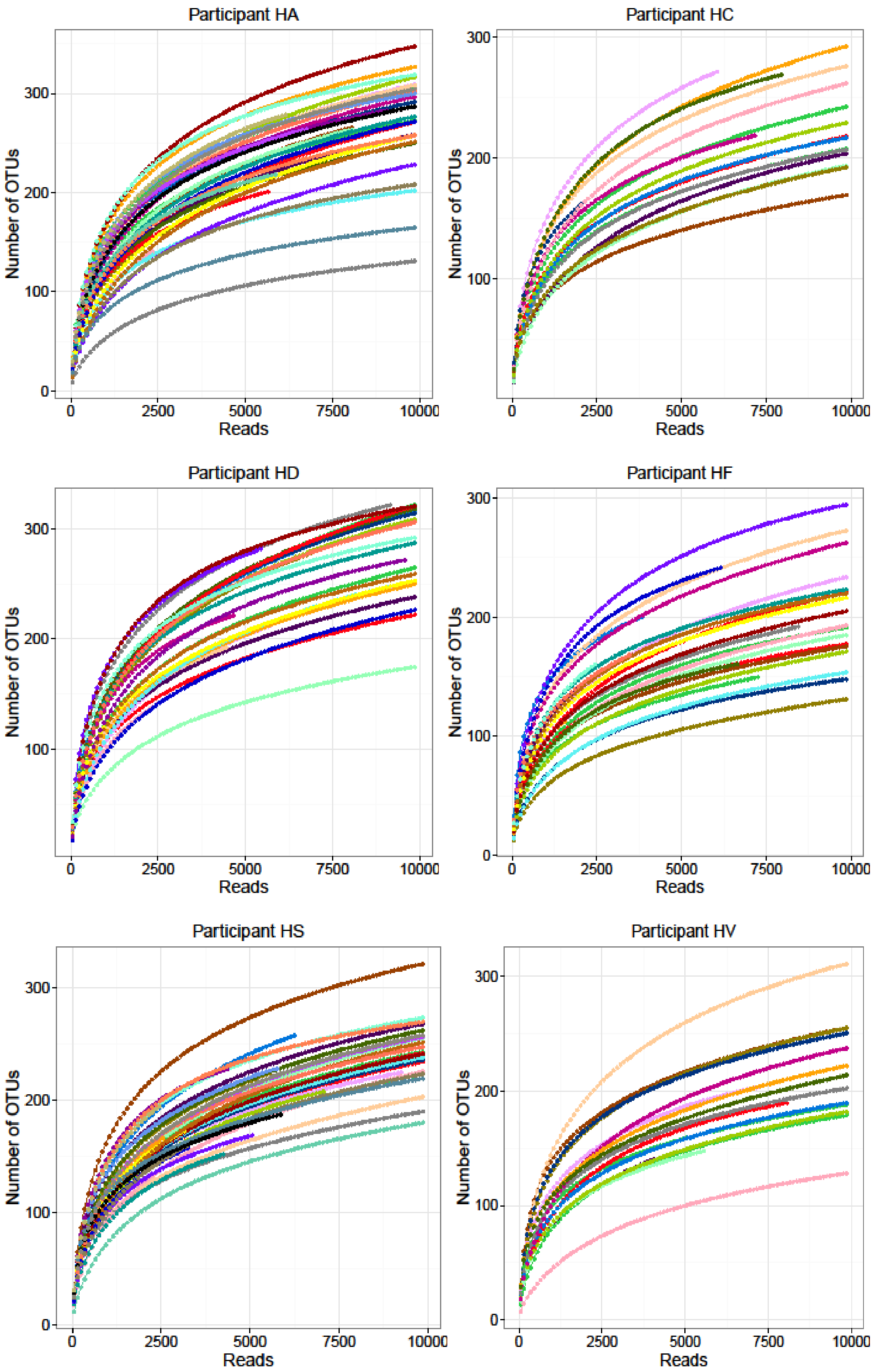


Figure 2.7 - Rarefaction curves displayed for selected participants (non-smokers) choosing a minimum cut off value at 5118 reads.

Alpha diversity at OTU level was investigated to determine the possible associations in oropharyngeal community structure from non-smokers and smokers. The diversity indices calculated for healthy and unhealthy samples (from non-smokers and smokers) were species richness, Shannon H index and Simpson index. Statistical testing used aov() from Vegan (version 2.4.0) to calculate pair-wise ANOVA (analysis of variation) *P* values (taking into account repeated sampling from participants) which were displayed on top of alpha diversity figures.

Co-occurrence networks and sub-community analysis (Williams et al., 2014) were produced in healthy samples from non-smokers at genus level to explore the interactions between specific genera and identify keystone species.

2.11.3 Stability analysis

The Vegan package (version 2.4.0) was used, in particular the two functions adonis() for PERMANOVA and betadisper() for the analysis of multivariate homogeneity of group dispersions, using Benjamini-Hochberg correction for multiple testing to report *P* values. The variability of microbial community structure between participants and health status (for both non-smokers and smokers) was also investigated at OTU level using betadisper() to measure the distance of each individual sample to that group's centroid (mean) allowing for comparison between participants and the different health groups. To understand multivariate homogeneity of groups dispersions (variances) between multiple conditions, betadisper() was used. Non-euclidean distances between objects and group centroids are handled by reducing the original distances (Bray-Curtis or unweighted UniFrac) to principal coordinates and then performing ANOVA on them. Adonis() was also used for analysis of variance using distance matrices (Bray-Curtis/unweighted UniFrac). This function, referred to as PERMANOVA, fits linear models to distance matrices and uses a permutation test with pseudo-F ratios, while using the strata command to take into account repeated sampling from participants.

The community stability of the microbiome for each participant (non-smokers and smokers) was quantified by producing stability plots at OTU level using the distances produced from *betadisper* to display the timeline of sampling and deviations. To address the question about a possible connection between community structure (diversity) and stability, community stability was quantified by calculating the coefficient of variation (ratio of standard deviation to the mean) (Donohue et al., 2016) for each individual participant using the distances from *betadisper* and testing them against diversity variables (number of phyla, genera and OTUs) and number of cold disturbances and viral infections using Spearman rank correlation analysis. The distances from *betadisper* were also tested in a linear mixed model (LMM) to determine if changes in distances could indicate a change in microbial community structure before, during and after a disturbance such as a cold. Sampling week (to accommodate for potential temporal trends) was fitted as a fixed effect and participant ID as a random effect using lme4 (version 1.1-9) and MASS (version 7.3-44) packages. AIC values and likelihood ratio testing were used to compare models.

2.11.4 Assigning functions to communities

Predicted functional profiles of bacterial taxa using 16S rRNA gene sequences were identified using the Tax4Fun package (Aßhauer et al., 2015) which links sequences with the functional annotation of sequenced prokaryotic genomes using a nearest neighbour identification based on a minimum of 16S RNA sequence similarity. It works by blasting the OTUs against silva database (all prokaryotic KEGG organisms are available in Tax4Fun for SILVA SSU Ref NR database release 115 and KEGG database release 64.0) and then utilizing ultrafast protein classification (UProC) tool (Meinicke, 2015) to generate metabolic functional profiles after normalising the data for 16S rRNA gene copy numbers. This shows the pathways as KEGG K numbers which are significantly up/down regulated between multiple conditions, as determined through using Kruskal-Wallis test showing *P* values and multiple testing correction (Benjamini-Hochberg) *P* adjusted values.

3 Characterising the healthy oropharynx microbiome of non-smokers

3.1 Introduction

High throughput sequencing has revealed each body site harbors a vast number of microbes living as a complex microbiome which may contribute to, or even be solely responsible for specific roles in the host (Jones, 2009) (Cho & Blaser, 2012) functioning like any macrobiotic ecological community. Research is now moving on from simple characterisation of the composition in microbiome communities, to improving knowledge about functioning of these communities (Robinson et al., 2010) and the possible links between the microbiome and health (Cho & Blaser, 2012). However, a necessary starting point of understanding any microbiome is the identification of the microbes present; investigations into the species interactions and putative health implications become possible only then (Costello et al., 2012) (Lemon et al., 2010).

One body site that remains relatively unexplored is the oropharynx. The oropharynx is the middle part of the throat and is a component of the upper respiratory tract. The oropharynx is a niche for commensal bacteria that is constantly exposed to various environmental sources and consequently an important entry point for pathogenic bacteria. For instance, upper respiratory infections (which affect the nose and throat) are very common in all ages of people (Ferkol & Schraufnagel, 2014). It is therefore of interest to know the normal oropharynx microbiome which interacts with invading pathogens and either prevents or facilitates the growth of them, as well as that of opportunistic pathogens normally present. Previous studies (Segata et al., 2012) (Botero et al., 2014) have shown the oropharynx microbiome to consist of an array of microorganisms lining the epithelium and existing as a complex community. Five major bacterial phyla have been identified in the oropharynx: *Firmicutes, Bacteroidetes, Proteobacteria, Fusobacteria* and *Actinobacteria* (Lemon et al., 2010). Whilst the majority of microbes are commensal in the oropharynx, many opportunistic bacteria may be present such as *Streptococcus mitis* (Mitchell, 2011) as well as pathogenic bacteria such as *Streptococcus pneumoniae* and *Neisseria meningitides* (Gazi et al., 2004). Investigating the bacterial composition in healthy communities will aid in determining the co-occurrence patterns and whether there are correlations between types of bacteria (Williams et al., 2014). This is something that has previously had little recognition in the oropharynx although determining these relationships is important for characterisation of communities, as well as identifying keystone taxa which could be beneficial for therapeutic purposes.

The impact of environmental factors on any microbiome is still poorly understood and there is now growing recognition of the importance of these factors to try and understand the relationship between the environment and the microbiome (Conlon & Bird, 2015). Although not investigated in this study, diet has shown to impact the GI tract microbiome (Turnbaugh et al., 2009). Host characteristics such as sex and age also affect the microbiome; there is now increasing evidence that the microbiome changes as age progresses (Whelan et al., 2014) and could also differ in regards to sex (Bolnick et al., 2014). All these factors may contribute to differences in microbial communities in healthy people and aid in understanding of community structure.

Characterisation studies should also involve longitudinal sampling as this will determine what bacteria are present overall in healthy communities in the oropharynx and how participants naturally vary in community composition (measured by alpha diversity) over a defined time period. There are still minimal studies involving longitudinal sampling, but taking samples over various weeks also gives an indication of how many samples are needed to sample a naturally fluctuating oropharyngeal community in order to recover as many of the taxa present, but also determine the natural variation present and investigate the cause and effect relationship of certain diseases and disorders. Therefore this chapter will explore the community composition of the bacterial oropharynx microbiome in healthy samples from non-smoking participants through longitudinal sampling. The objectives are listed as follows: to characterise the oropharynx microbiome in healthy samples from non-smoking participants, to determine how sex and age affects the healthy microbiome and to produce co-occurrence networks present in healthy samples.

3.2 Methods

3.2.1 Initial exploration of the oropharynx microbiome

All analyses were carried out using R (version 3.1.2). After sequencing (*Chapters 2.2.1, 2.7, 2.8 & 2.9*) all samples below 5000 reads were removed from subsequent analyses because of lack of statistical support. This resulted in 313 samples from healthy participants (n=18).

Taxonomic classification at phylum, genus and OTU level was done through the RDP database classifier using the standalone RDP classifier version 2.6 *(Chapter 2.10)*. For alpha diversity analysis (such as species richness) samples were rarefied using rarefaction to the minimum number of reads (5118). To address the question of how many samples from a naturally-fluctuating, healthy oropharynx would be needed to capture 100% of taxa recovered per participant (at the phylum, genus and OTU level) cumulative box plots were produced. Participants that had a minimum of 5 healthy samples were included – this involved all participants apart from participant HB. The number of samples per participant instead of weeks was used as participants did not hand in a swab every week.

3.2.2 Community composition

The effects of age and sex were explored in healthy communities through NMDS plots using Bray-Curtis distance and unweighted UniFrac distance at OTU level. Variance ellipses were added as an indication of the variability of each group; the covariance was calculated using cov.wt() in Vegan (version 2.4.0) (Oksanen 2013) and the shape of the ellipse was defined by the covariance within each group (*Chapter 2.11.2*). Significant difference testing between the different groups in the sex and age categories at OTU level was done using PERMANOVA (permutational ANOVA) through adonis() in Vegan using the command strata to take into account repeated sampling from participants. Due to the small sampling size, a cut of *P* value of < 0.1 was used to determine significance between the sex and age categories. The most significant OTUs present in terms of differing abundance between the different groups in sex and age were determined from a negative binomial GLM and was displayed showing the log relative transformation of samples using DESeq() from the DESeq2 package (version 1.12.3) (Love et al., 2014). This uses a negative binomial GLM to model the abundance data (OTU frequencies) and empirical Bayes to shrink OTU-wise dispersions to identify OTUs that have log-fold changes between different conditions (at a cut off value of $P < 0.01$) (*Chapter 2.11.2*).

3.2.3 Co-occurrence networks

Co-occurrence network analyses (Williams et al., 2014) were performed on data from healthy samples (non-smokers) to explore the interactions between bacteria and to identify important members of the community. Analysis was done at the genus level to determine if there were any interactions between specific genera. Samples were rarefied to 5118 reads representing the minimum number of counts per sample. Co-occurrence patterns were investigated through generating a dissimilarity matrix consisting of Spearman correlation coefficients to represent co-occurrence between all pairs of genera from samples. Networks were produced using the igraph package (version 1.0.1) (Csardi & Nepusz, 2006) where microbial taxa were represented as nodes and the presence of a co-occurrence relationship based on a 0.5 correlation level was represented by edges. Keystone taxa were identified as having the largest node size and the greatest number of connections to other taxa as used in Williams et al., (2014).

3.3 Results

3.3.1 Initial exploration of the healthy oropharynx microbiome

At the end of the sampling period, 313 samples were received in total, with 279 designated as healthy due to not having any symptoms of illness (Table 3.1). The taxonomic profiling of samples from the oropharynx of individual participants identified with RDP classifier revealed 5 to 10 phyla, 20 to 70 genera and 140 to 340 assignments at OTU level.

Table 3.1 – The total number of healthy samples from non-smokers and the range in taxa numbers received from each participant.

Participant	**Healthy samples**	**Phylum**		**Genus**		**OTU**	
		Min - Max	**Median**	**Min - Max**	**Median**	**Min - Max**	**Median**
HA	34	8-10	8	24-61	51	156-299	247
HB	2	8-8	N/A	36-42	N/A	174-184	N/A
HC	14	7-9	8	33-58	45	145-294	207
HD	23	6-9	9	36-66	53	158-299	266
HE	6	8-9	9	39-55	45	191-279	218
HF	17	6-9	8	28-51	40	148-261	180
HG	7	7-8	8	30-48	40	170-245	202
HI	26	7-9	8	36-58	45	175-335	221
HJ	8	5-9	8	34-51	44	162-209	192
HL	5	7-9	8	41-67	46	180-268	210
HM	11	5-9	7	33-53	36	150-289	177
HN	11	8-9	8	33-59	42	174-297	211
HO	25	7-10	8	35-57	44	165-291	204
HQ	8	7-9	8	38-55	49	201-268	240
HR	14	7-9	9	38-63	51	189-318	222
HS	29	8-9	9	35-54	46	175-306	216
HT	30	7-9	9	36-65	47	167-298	241
HV	12	6-9	9	33-64	43	147-291	202

Cumulative plots showed the total number of taxa detected from each participant as a percentage of the participant total thereby indicating the minimum number of samples needed from each participant to recover the total numbers of different taxa present. There was relatively little variation in this between the participants, and a minimum of 2, 3 and 4 samples was needed at phylum, genus and OTU level to recover all the taxa present (Figure 3.1).

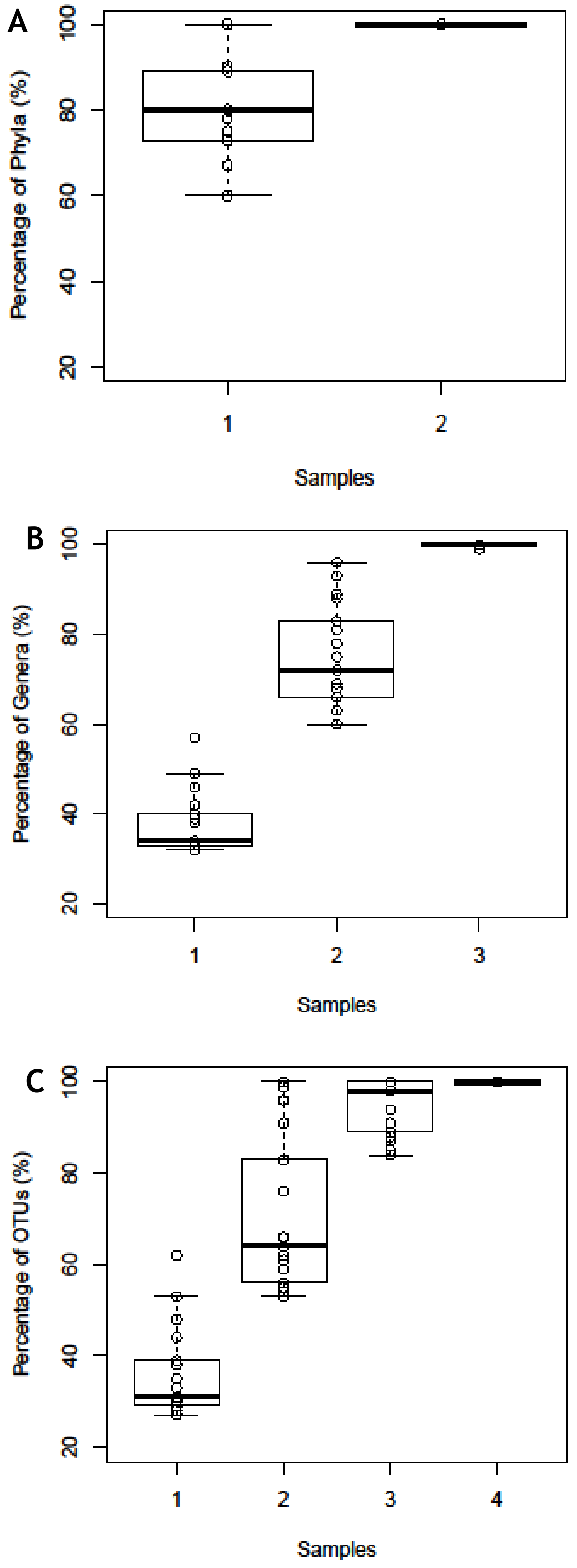


Figure 3.1 - Box plots showing cumulative percentage of phyla (Fig. 3.1A), genera (Fig. 3.1B) and operational taxonomic units (OTUs) (Fig. 3.1C) recovered from each participant over subsequent samples (100% = number of taxa recovered per participant). Box plots show the minimum, 25th percentile, median, 75th percentile and maximum values for the cumulative percentage of taxa recovered from consecutive samples within a single participant.

3.3.2 Community composition of the healthy oropharynx microbiome

At phylum level 99.7% of reads were taxonomically classified with the remaining 0.3% belonging to unknown or unclassified bacteria. There were 5 main phyla that were always present in healthy samples: *Actinobacteria, Bacteroidetes, Firmicutes, Fusobacteria* and *Proteobacteria*. The occurrence of *Spirochaetes, TM7, SR1* and *Synergistetes* was more variable. Considering the abundance of a given taxonomic level in a microbiome sample, the most abundant phylum overall was *Firmicutes* (mean ± SEM = 61% ± 1%) followed by *Bacteroidetes* (16% ± 1%), *Proteobacteria* (11% ± 1%), *Actinobacteria* (7% ± 0.2%) and *Fusobacteria* (5% ± 0.2% (Figure 3.2 & Appendix 4*)*.

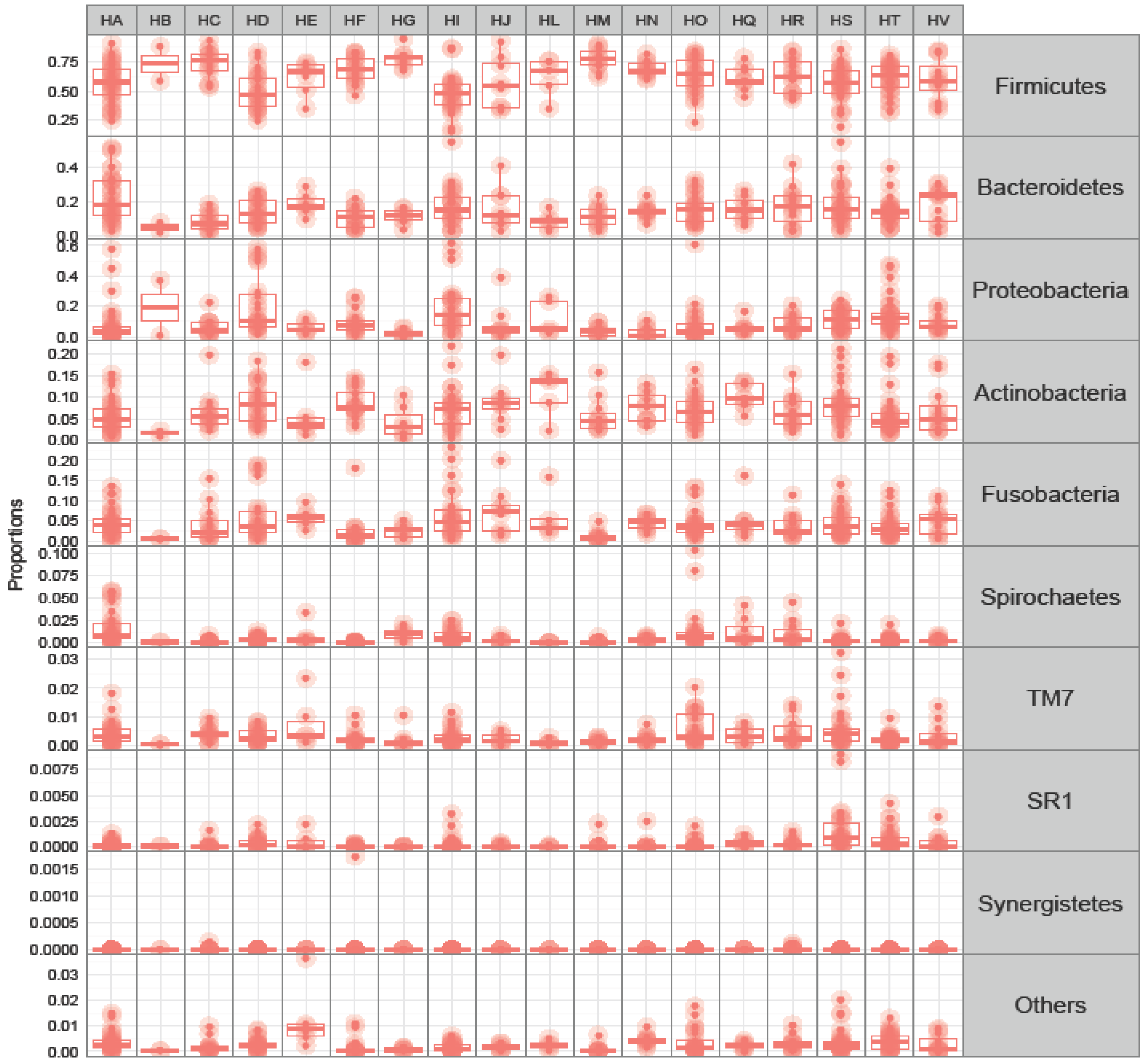


Figure 3.2 – Box plot showing the most abundant phyla (n=9) (the rest pooled in the category ‘Others’) and the median abundance in each participant in healthy samples.

At genus level 95% of reads were taxonomically classified with the remaining 5% belonging to unknown or unclassified bacteria. At the genus level, the patterns of presence and occurrence were similar to the phylum level with the most common genera present in all participants. However, some genera have conspicuously more variable occurrence (*Porphyromonas*). *Streptococcus* (mean ± SEM = 47% ± 1%) was the most dominant genus in the majority of healthy samples followed by *Prevotella* (9% ± 0.4%) and *Veillonella* (5% ± 0.2%) (Figure 3.3 & Appendix 5).

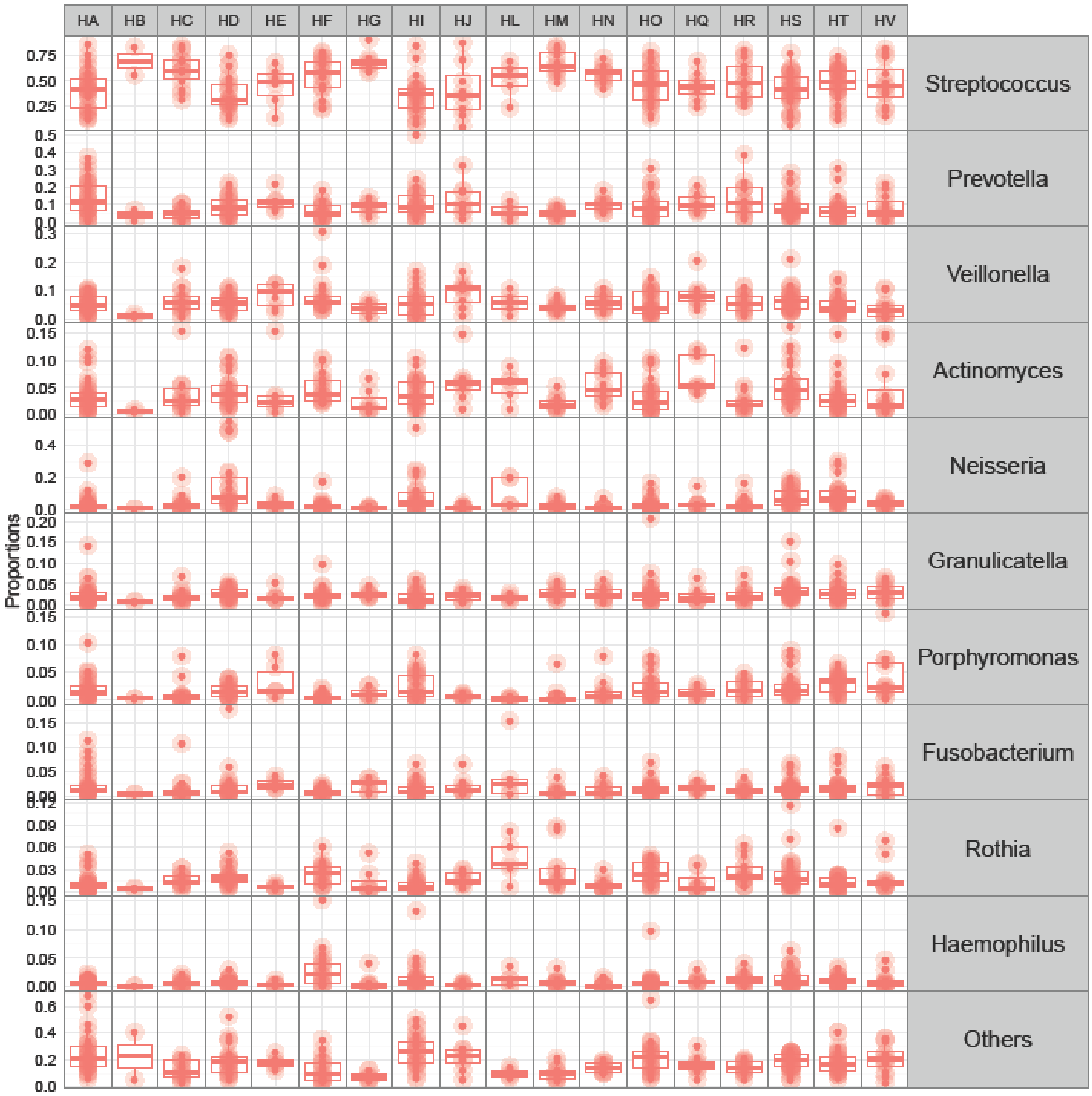


Figure 3.3 – Box plot showing the most abundant genera (n=10) (the rest pooled in the category 'Others') and the median abundance in each participant in healthy samples.

The most abundant OTUs belonged to *Streptococcus* species reflecting the general abundance of their phylum, *Firmicutes* (Figure 3.4). Some *Streptococcus*

OTUs could be identified to species level, however not all *Streptococcus* OTUs could be identified due to the V1-V2 region not being variable enough for adequate species identification. *Streptococcus* OTUs identified included *Streptococcus mitis*, *Streptococcus salivarius* and *Streptococcus parasanguinis* which are all commensal but can be opportunistic (Mitchell, 2011). Even though some *Streptococcus* OTUs could be named, there were *Streptococcus* OTUs from the oropharynx that remain unidentified and so could not be determined whether they are commensal or pathogenic.

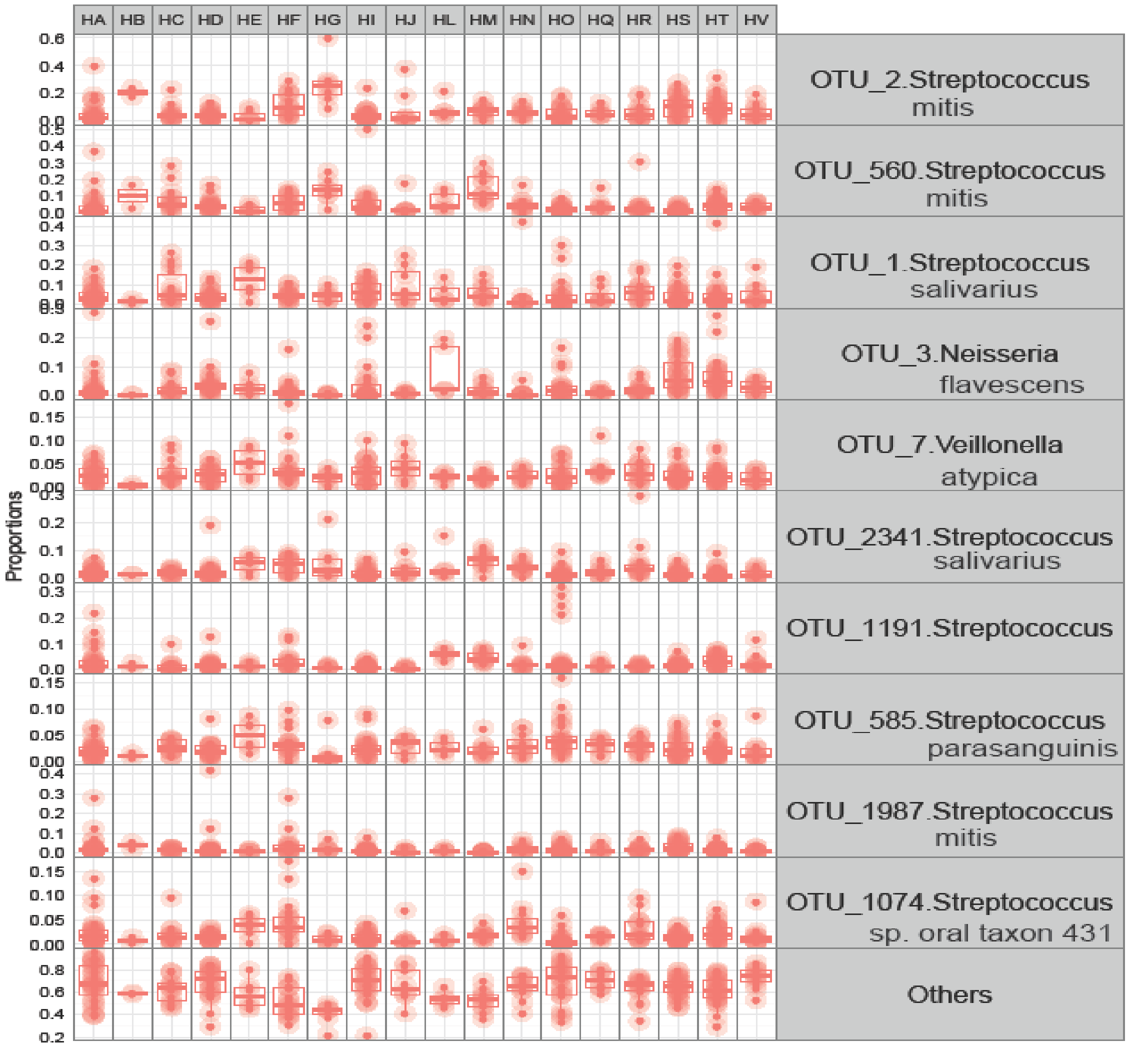


Figure 3.4 – Box plot showing the most abundant operational taxonomic units (OTUs) (n=10) with the rest pooled in the category 'Others' and the median abundance in each participant in healthy samples.

3.3.3 Host characteristics affecting the microbiome

NMDS plots at OTU level gave an indication of the similarity of community composition in males and females (Figure 3.5A) and between the different age groups (Figure 3.5B). There was a visual divide in the communities between males and females; however these communities were not significantly different using Bray-Curtis distance (*P* = 0.8) but were significantly different using unweighted UniFrac distance (*P* = 0.07) at a cut off value at *P* < 0.1. This states that overall there was not a difference between the abundance and but there was a difference in the presence and absence of OTUs - the oropharynx microbiome between males and females differs in that males have different OTUs than females. However, further investigation did show specific OTUs being significantly different in abundance; females had higher abundances of *Sneathia* and *Catonella*, whereas males had increased *Porphyromonas* (Appendix 6).

The NMDS plot for age showed the teens and the thirties group to be the most distinct of the groups. Again there was no significant difference in communities between the different age groups using Bray-Curtis distance (*P* = 0.6) but significant differences were present using unweighted UniFrac distance (*P* = 0.05) which showed communities to be distinct by having different OTUs rather than differences in abundance. However, there were specific OTUs that were significantly different in terms of abundance between the different age categories; the most significant differences in increasing age (from the teenage years to twenties and thirties) included a significant increase in the abundance of *Neisseria, Sneathia* and *Prevotella* with a decrease in the abundance of *Streptococcus* (Appendix 7). This showed that overall the different groups within the sex and age categories had similar community composition; *Streptococcus, Prevotella* and *Veillonella* were the most abundant genera in all healthy communities regardless of age or sex. However, communities were significantly different in that males and females had different OTUs present, as did the separate age groups.

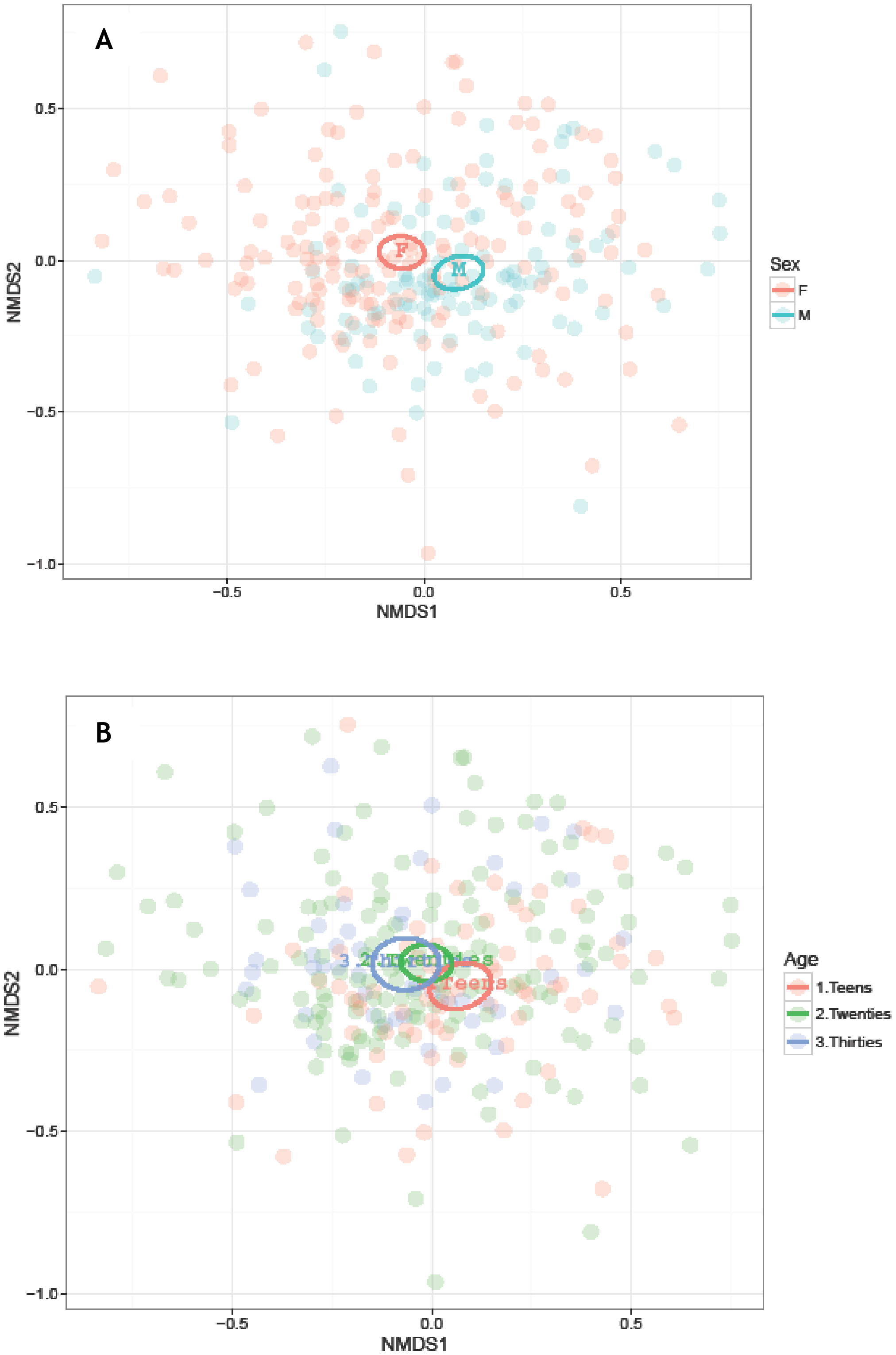


Figure 3.5 – Non-metric multidimensional scaling (NMDS) plots using Bray-Curtis distance at operational taxonomic unit (OTU) level showing the effects of sex (Fig. 3.5A) and age (Fig. 3.5B) on healthy samples (n=279).

3.3.4 Co-occurrence of bacteria in the healthy oropharynx

Co-occurrence relationships showed co-existence patterns of bacteria in healthy samples (Figure 3.6), with only the most significant co-occurrence patterns of taxa shown at a correlation value of 0.5. Most correlations between the taxa were positive in that as one taxon increased in abundance, the other taxon would also increase in abundance. However a negative co-occurrence pattern was observed between unclassified *Veillonellaceae* and *Streptococcus*. There was a dominant sub-community where *Megasphaera*, *Prevotella*, *Veillonella* and unclassified *Veillonellaceae* had the biggest nodes due to having the most connections. These genera were also seen as having the greatest betweenness values (as displayed by the betweenness and eigenvalue plots) as they were shown to be located on the edge of the betweenness plot showing various connections to other nodes. Therefore these genera were seen as important members of the healthy oropharynx microbiome. By having the greatest number of connections in the co-occurrence network (being connected to the most number of nodes) they were shown to be the genera that interacted with other taxa the most.

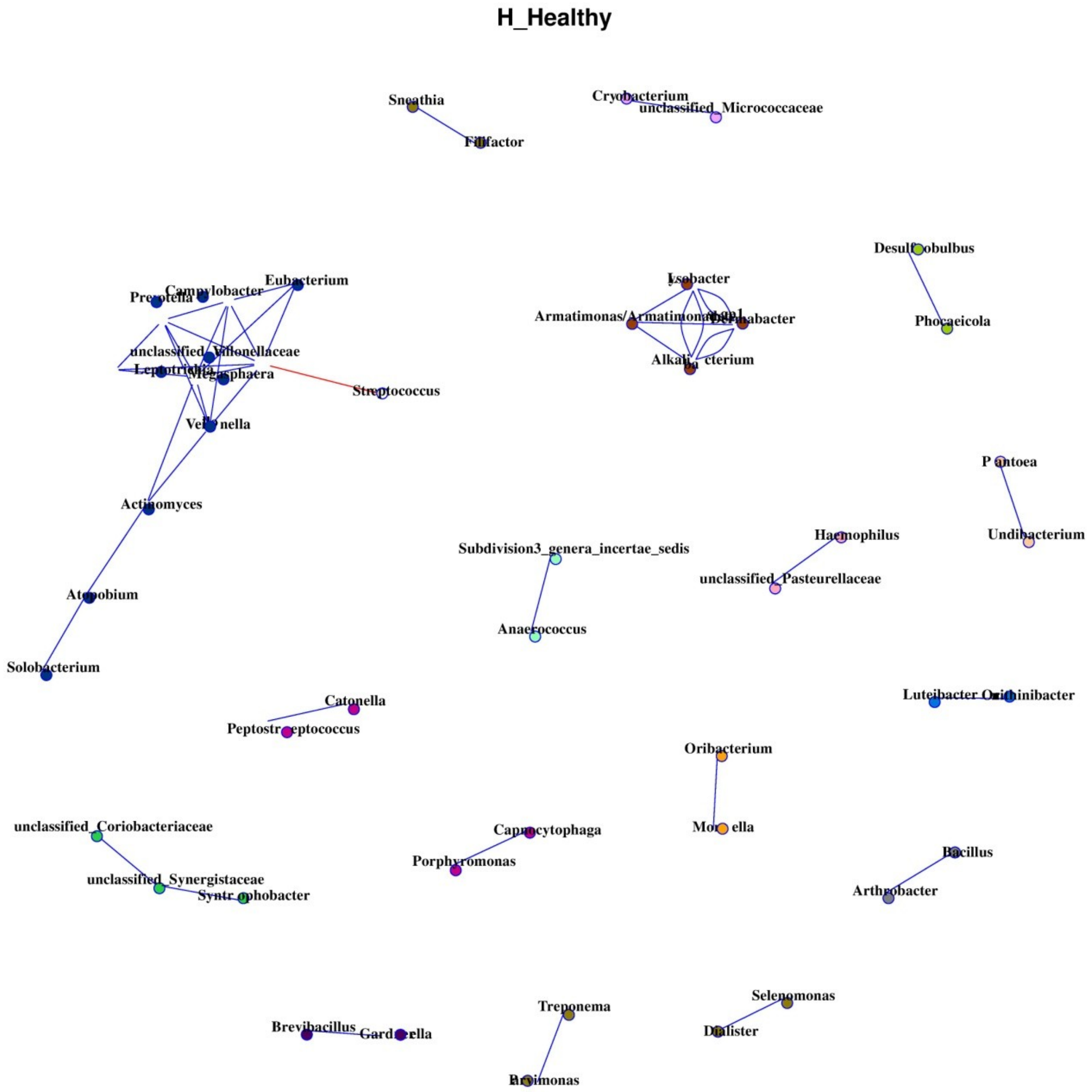


Figure 3.6 – A co-occurrence network generated from all healthy samples (n=279) at genus level using a correlation value of 0.5. Sub-communities consist of various nodes and different genera are defined by colour coded nodes. Larger nodes show genera that have the greatest number of connections to other genera (through having a lower *P* value).

3.4 Discussion

The healthy oropharynx microbiome was investigated to determine community composition, the changes that occur to communities in regards to host characteristics such as age and sex, and the interactions of specific taxa within these communities. The results showed that *Firmicutes* and *Streptococcus* are the most dominant phylum and genus in healthy samples which was also seen in previous studies (Segata et al., 2012) (Lemon et al., 2010) (Lazarevic et al., 2009) and so, these could be considered as important taxa as an indicator of health status. In order to fully identify the key *Streptococcus* species, all *Streptococcus* OTUs must be identified and categorised to commensal and pathogenic OTUs (Mitchell, 2011), something that was not possible for these samples. However, it was possible to identify the most abundant *Streptococcus* OTUs present in healthy samples which included *Streptococcus mitis* and *Streptococcus salavarius*.

In this study, sex and age did not dramatically alter the microbiome, even though there were significant differences in the presence/absence and abundance of specific OTUs in regards to age. The effects of aging on the microbiome have been investigated previously showing that the microbiome changes as we age and elderly people have distinct communities compared to younger adults (Saraswati & Sitaraman, 2014). However, as this study only involved adults within specific age ranges (18-37) and did not investigate between the different extremes of age (young adults to elderly), only a subset of the wider age range was observed and smaller differences between age groups could therefore be expected. The communities of males and females also had similar microbiomes. Other factors may be responsible for changes in the microbiome such as diet. Diet has shown to affect the GI microbiome (David et al., 2014) which could also potentially affect the oropharynx microbiome. Males and females may have different diets also – males may have more meat consumption in their diet compared to females or females may be more aware of their diet compared to males. However a study has shown that the same diet in males and females has different effects on the GI microbiome showing that the host must also influence the microbiome in the different sexes (Bolnick et al., 2014). Diet could be partly responsible for changes in the microbiome (and between the different sex and age groups) but diet was not investigated in this study. Therefore the role of sex and age (coupled with diet) requires further investigation in how it affects the oropharynx microbiome in healthy participants.

Investigating the OTUs in healthy communities improved understanding of the bacterial community structure in the oropharynx and the interactions between specific bacteria. This was partly investigated through co-occurrence network plots which showed the co-existence patterns of bacteria in healthy samples. In healthy communities *Prevotella* and *Veillonella* were important gatekeepers in that they had the biggest nodes and many edges showing they interacted with various bacteria in the oropharynx - these bacteria were also considered dominant bacteria as they were usually the most abundant following *Streptococcus*. This may influence the structure of the oropharyngeal community as a whole by controlling abundances of other taxa and overall functioning of the community. Most of these interactions were also positive, showing that a healthy state is created by the presence of these bacteria which in turn increases the abundance of other bacteria. Interestingly, in the healthy samples *Streptococcus* was not seen to have any positive significant co-occurrence networks with any other taxa at this correlation level even though it has been recognised as the most dominant genus in healthy oropharyngeal communities. This could be due to only creating co-occurrence networks at genus level indicating that *Streptococcus* species may prefer to interact with other species of *Streptococcus*. There is evidence that *Streptococcus* species interact in the oral tract (Kreth et al., 2009); the commensal *S. sanguinis* is able to produce hydrogen peroxide to inhibit the growth of the pathogenic *S. mutans* suggesting that the oropharyngeal community has a massive sub-community of *Streptococcus* species that are in constant existence and interaction with each other. This can be explored further by naming all the *Streptococcus* OTUs present, identifying which OTUs are commensal, opportunistic and pathogenic and then creating these co-occurrence networks at both the genus and OTU level in healthy samples making it possible to understand the interactions in key *Streptococcus* OTUs present in healthy and eventually unhealthy states.

3.5 Conclusions

Overall the healthy oropharynx microbiome in non-smoking participants was found to be similar at the phylum level with increasing differences at genus and OTU level. The most dominant taxa in healthy communities were identified as *Firmicutes* at phylum level and *Streptococcus* at genus level, but the oropharynx microbiome was not majorly impacted by sex and age in this study. Co-occurrence networks did show interactions of specific bacteria in healthy samples, but this needs to be further explored to show how *Streptococcus* OTUs co-exist in the different health states.

4 Characterisation of the unhealthy oropharynx microbiome in non-smokers

4.1 Introduction

A necessary starting point in understanding any microbiome is the identification of the microbes present in a healthy setting. This was explored in *Chapter 3* where the oropharyngeal community composition in healthy samples from non-smokers was characterised. This provided the foundation of identifying universal features of the healthy microbiome such as community composition and the most abundant taxa present which can then be directly compared to microbiomes in specific disease scenarios. Links between disease and microbiome compositions have been reported in a variety of conditions such as IBS (Willing et al., 2010) and periodontitis (Abusleme et al., 2013) where deviations from the healthy state, dysbiosis, are associated with disease in the host.

There are few studies comparing the healthy oropharynx microbiome to specific disease scenarios. The majority of these studies compare the healthy oropharynx microbiome to lower respiratory tract diseases that affect the lungs such as asthma (Park et al., 2014), chronic obstructive pulmonary disorder (Cabrera-Rubio et al., 2012) or tuberculosis (Botero et al., 2014). Few studies have investigated upper respiratory tract infections such as the impact of viral infections like the common cold (Yi et al., 2014) or bacterial infections such as tonsillitis (Stenfors et al., 2003) and their impact on the oropharynx microbiome. However these studies do not take into account longitudinal sampling and only compare a healthy control group to diseased samples. It is important to consider the possible effects of upper respiratory tract infections, how they affect the whole population and the differing outcomes of disease depending on health status or age of the population affected. To improve our understanding of the possible role of the oropharynx microbiome in upper respiratory tract infections, the issue of causality needs to be addressed, i.e. whether infections follow from disturbances or if these are merely associated with them. This requires characterising the microbiome through longitudinal sampling and determining its overall stability *(Chapter 5)*.

This chapter, in contrast, explores the differences in community composition of the bacterial oropharynx microbiome in unhealthy samples compared to those from the same non-smoking participants when they are healthy. Specifically, these differences were characterised during colds, viral infections and antibiotic treatment, and compare the alpha diversity (species richness and Shannon/Simpson Index) as well as specific beta diversity measures of healthy and unhealthy samples. As previous literature has demonstrated diseased states resulting in altered, low diversity communities, it is a possibility that this pattern will also be seen in the oropharynx. Therefore the objective of this chapter is to determine the changes in community composition between healthy and unhealthy samples from non-smokers. The hypothesis is that the healthy oropharynx microbiome from a non-smoker is a high diversity community consisting of keystone species which will become unbalanced during a respiratory disturbance (due to loss of keystone species and diversity).

4.2 Methods

4.2.1 Initial analysis

All samples below 5000 reads were removed from analyses resulting in 313 samples from non-smokers (n=18), of which 34 samples were identified as unhealthy samples. The definition of an unhealthy sample is a sample that was collected at the time a participant reported symptoms of a disease or illness or they were on antibiotic treatment. Even though participants on antibiotics were on treatment for acne and not because of an infection, they were still categorised with the unhealthy group due to disturbing the community structure. These samples were categorised into the following groups: cold (n=19 from 12 participants), antibiotics (n=8 from 3 participants) and viral (n=7 from 4 participants). The cold group includes samples from participants self reporting symptoms of a cold but were detected as negative for the standard viruses tested during the respiratory screen at Gartnavel hospital as described in *Chapter 2.2.1*. The antibiotics group included any participants undergoing antibiotic treatment and the viral group included only symptomatic samples with confirmed viruses from the respiratory screen.

4.2.2 Comparison of community composition between healthy and unhealthy samples

Taxonomic classification at phylum, genus and OTU level was done through the RDP database classifier using the standalone RDP classifier version 2.6 *(Chapter 2.10)*.

The abundances of specific taxa at phylum and genus level were log transformed and tested in a linear mixed model (LMM) to determine if changes in abundance could indicate a change in health status. Sampling week (to accommodate for potential temporal trends) was fitted as a fixed effect and participant ID as a random effect using lme4 package (version 1.1.9) from R (version 3.1.2). Species richness at OTU level was also tested in a LMM against sample status (healthy and unhealthy) to determine if unhealthy samples had reduced species richness overall.

Local contributions to beta diversity (Vegan, version 2.4.0) in R (version 3.1.2) was performed to show dysbiosis of unhealthy communities as described in Legendre & De Cáceres, 2013. This method involves measuring beta diversity to show the variation in species composition by generating a single number estimate of beta diversity in the different health groups. This was done by using Hellinger transformation to compute the total sum of squares of the species composition from which the local contributions to beta diversity could be derived generating the total beta diversity. This generates values known as local contributions to beta diversity (LCBD) where a large LCBD value indicates samples that have different species composition. This was performed at OTU level producing a timeline of samples for each participant displaying the *P* values for the unhealthy samples.

The abundance changes in community composition between the different health groups were tested by identifying the most significant OTUs present in regards to health status using the DESeq2 package (version 1.24.0) (Love et al., 2014) in R (version 3.1.2). This was determined from a negative binomial GLM to model the abundance data (OTU frequencies) and empirical Bayes to shrink OTU-wise dispersions to identify OTUs that have log-fold changes between different conditions. The cut off value was $P < 0.01$ with P adjusted values being used for multiple comparisons (*Chapter 2.11.2*).

Alpha diversity at OTU level was investigated to determine the possible associations between health status and oropharyngeal community structure. For alpha diversity analysis, samples were rarefied using rarefaction to the minimum number of reads (5118). The diversity indices calculated for healthy and unhealthy samples were species richness, Shannon H index and Simpson index (*Chapter 2.11.2*). Significant differences between the different health groups were measured using aov() from Vegan (version 2.4.0) taking into account repeated sampling from participants to calculate pair-wise ANOVA generating P values which were displayed on top of alpha diversity figures.

4.3 Results

4.3.1 Comparison of the healthy and unhealthy oropharynx microbiome from non-smokers

4.3.1.1 Initial exploration of the unhealthy oropharynx microbiome

The taxonomic profiling of unhealthy samples from the oropharynx of non-smoking participants identified with RDP classifier revealed 5 to 9 phyla (median = 8), 20 to 70 genera (median = 38) and 100 to 300 assignments at OTU level (median = 182) (Table 4.1). This was broadly similar to the taxonomic profiling of healthy samples (5 to 10 phyla (median = 9), 20 to 70 genera (median = 45) and 140 to 340 OTUs (median = 218), but showing a reduction in OTUs in unhealthy samples as determined in a linear mixed model (t value = -5.693, $P <$ 0.001).

Table 4.1 – The total number of unhealthy samples from non-smokers and the range in taxa numbers received from each participant. Participants HB, HR and HT are excluded due to having no unhealthy samples.

Participant	Total samples	Phylum		Genus		OTU	
		Min - Max	Median	Min - Max	Median	Min - Max	Median
HA	2	8-8	N/A	27-47	N/A	109-245	N/A
HC	2	7-9	N/A	57-65	N/A	145-294	N/A
HD	1	9	N/A	60	N/A	272	N/A
HE	1	7	N/A	28	N/A	113	N/A
HF	5	5-7	6	29-51	35	111-241	149
HG	1	8	N/A	42	N/A	214	N/A
HI	6	7	7-8	30-47	41	144-230	186
HJ	4	6-8	7	29-44	37	125-198	173
HL	1	6	N/A	40	N/A	163	N/A
HM	3	6-8	7	35-44	43	158-206	180
HN	1	8	N/A	44	N/A	198	N/A
HO	1	8	N/A	35	N/A	130	N/A
HQ	1	7	N/A	30	N/A	116	N/A
HS	1	8	N/A	38	N/A	170	N/A
HV	4	5-8	7	25-52	42	107-237	187

4.3.1.2 Community composition of the unhealthy oropharynx microbiome

The five main phyla (in terms of abundance) found in the healthy samples (*Actinobacteria, Bacteroidetes, Firmicutes, Fusobacteria* and *Proteobacteria*) were also present in unhealthy samples (cold and viral samples) and antibiotic treated samples (Appendix 8) but the abundances between the different health conditions were different. There was a decrease in *Firmicutes* in 9 participants (n=12) and an increase in *Proteobacteria* in 9 participants (n=12) when going from a healthy to cold state (Figure 4.1A) whereas the viral group had a decrease in *Firmicutes* in 3 participants (n=4) and an increase in *Proteobacteria* in 2 participants (n=4) (Figure 4.1B). Antibiotic treated samples on the other hand showed similar abundances of *Firmicutes* in comparison to the healthy samples, but 2 participants (n=3) did show an increase in the phylum *Actinobacteria* (Figure 4.1C). Overall, when looking at the healthy and unhealthy group there was a significant increase in the abundance of the phylum *Proteobacteria* (*P* = 0.002) and a significant decrease in *Bacteroidetes* (*P* = 0.05) in unhealthy samples (Table 4.2). No differences between the healthy and unhealthy samples were found in the abundances of *Firmicutes, Actinobacteria* or *Fusobacteria*.

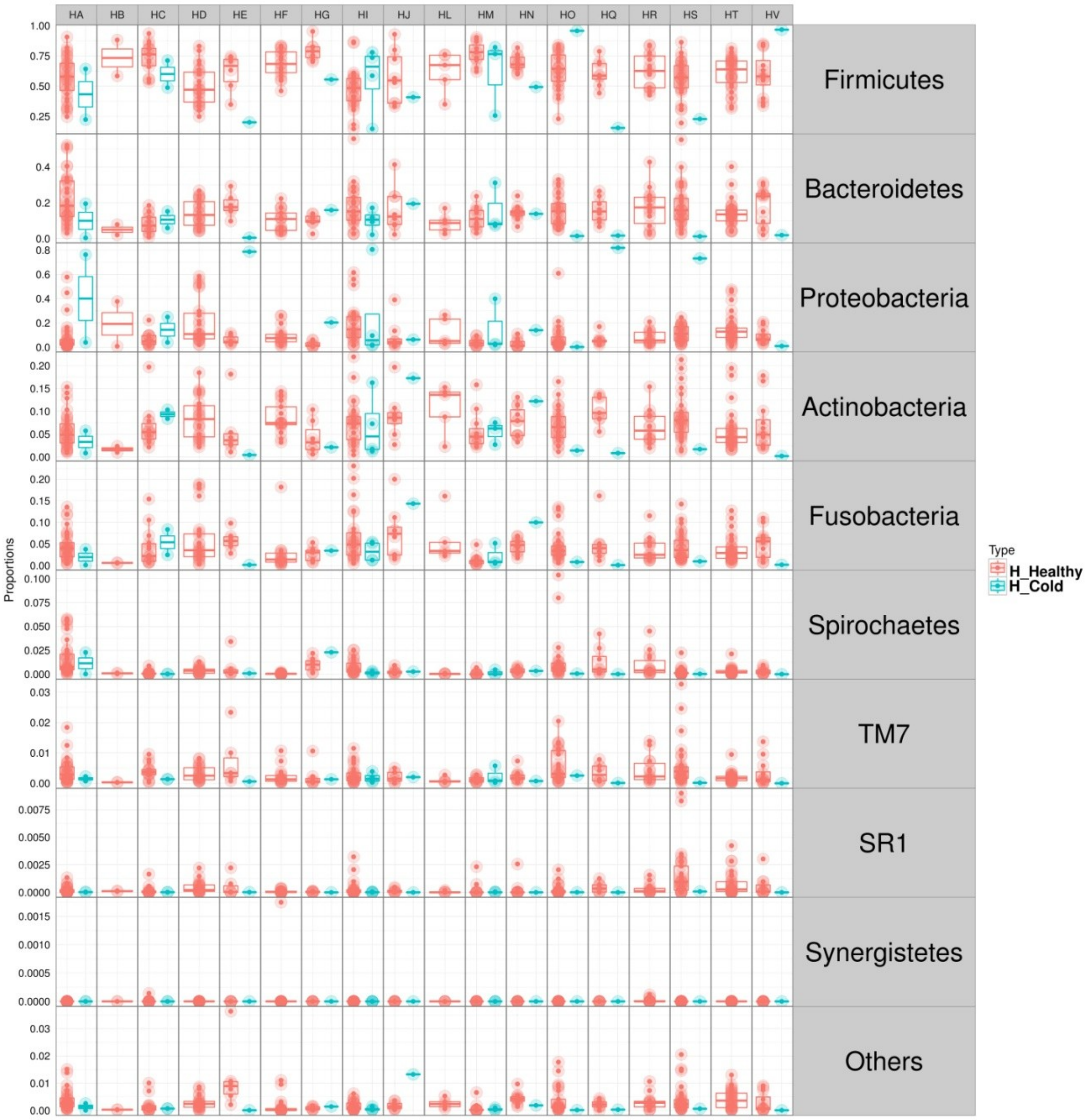


Figure 4.1A – Box plot showing the most abundant phyla (n=9) (the rest pooled in the category 'Others') and the median abundance in each participant in healthy and cold samples.

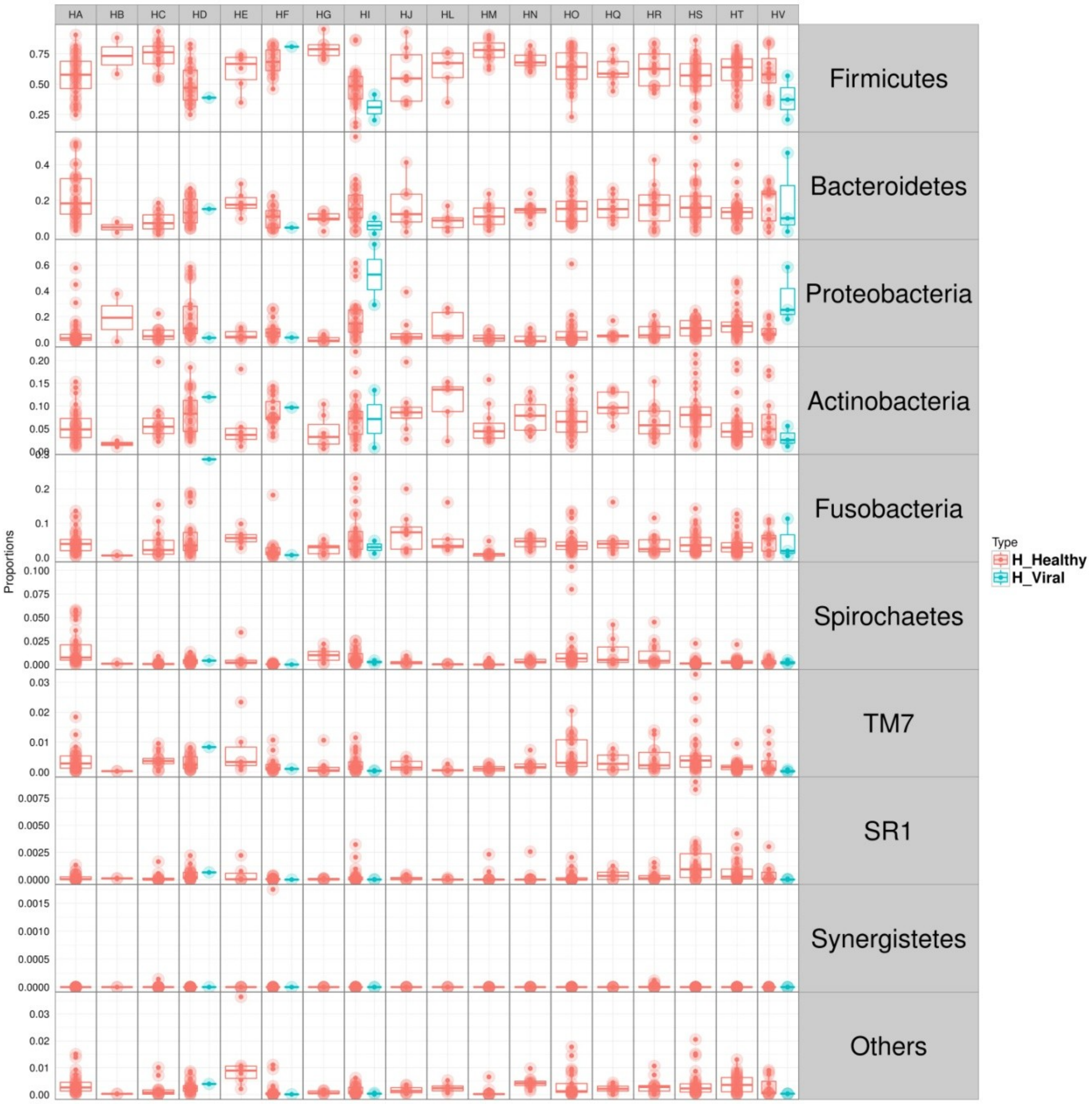


Figure 4.1B – Box plot showing the most abundant phyla (n=9) (the rest pooled in the category 'Others') and the median abundance in each participant in healthy and viral positive samples.

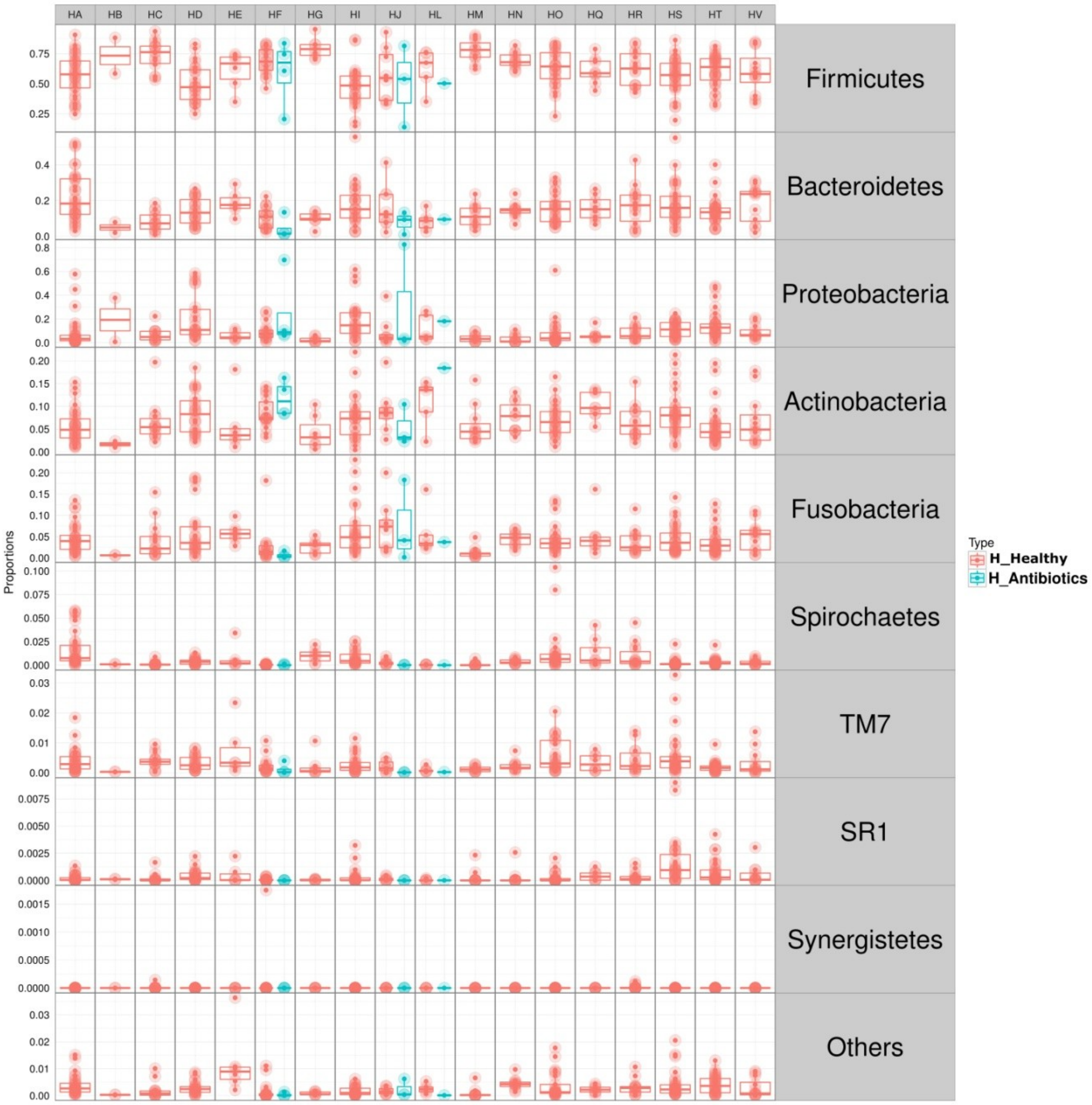


Figure 4.1C – Box plot showing the most abundant phyla (n=9) (the rest pooled in the category 'Others') and the median abundance in each participant in healthy and antibiotic treated samples.

Table 4.2 – Summary of the parameter estimates of the linear mixed model (LMM) investigating the abundances (response variable) of the five most abundant phyla in unhealthy samples compared to healthy samples (reference category). Significant *P* values are shown in bold.

Unhealthy samples	**Estimate**	**Std. Error**	**df**	**t value**	***P* value**
Firmicutes	0.028704	0.211088	313	0.136	0.891925
Proteobacteria	0.9015	0.293	310.4	3.077	**0.0023**
Bacteroidetes	-0.552464	0.281813	307.71	-1.96	**0.0509**
Actinobacteria	-0.089296	0.274397	313	-0.325	0.745
Fusobacteria	-0.420947	0.292411	311	-1.44	0.151

At genus level, the most abundant genera in both healthy and unhealthy samples (cold and viral) and antibiotic treated samples were *Streptococcus*, *Prevotella* and *Veillonella*, but there were marked increases in specific genera such as *Pseudomonas* in comparison to the healthy samples (Appendix 9). When comparing the healthy and cold samples (Figure 4.2A) there was a decrease in abundances of *Streptococcu*s and *Prevotella* in 8 and 9 participants respectively (n=12). The viral group had 3 participants where the abundance of *Streptococcus* decreased (n=4) and an increase in *Neisseria* and *Haemophilus* was observed in 1 and 3 participants respectively (n=4) (Figure 4.2B). For samples from subjects that had undergone antibiotic treatment (Figure 4.2C), 2 participants showed decreases in *Streptococcus* (n=3) and 2 participants had increases in the genus *Actinomyces* (n=2). When looking at the healthy and unhealthy samples overall, there was a significant decrease in the abundance of *Prevotella* in unhealthy samples (*P* = 0.012) with no significant difference in the abundance of *Streptococcus* and *Veillonella* (*P* = 0.808 & *P* = 0.385) (Table 4.3). Therefore, unhealthy communities had the same genera present, but the communities differed by having different abundances of genera.

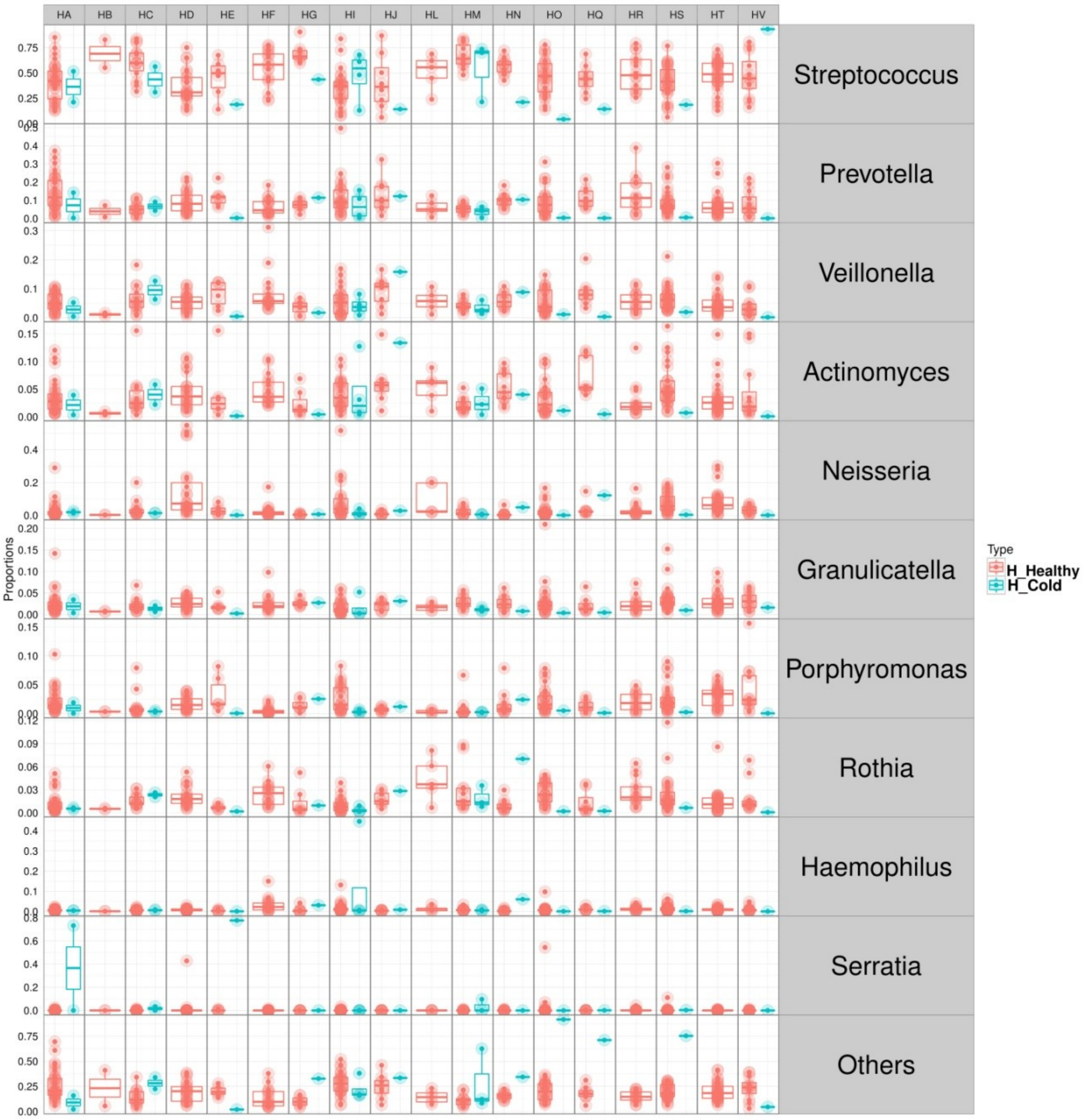


Figure 4.2A – Box plot showing the most abundant genera (n=10) (the rest pooled in the category 'Others') and the median abundance in each participant in healthy and cold samples.

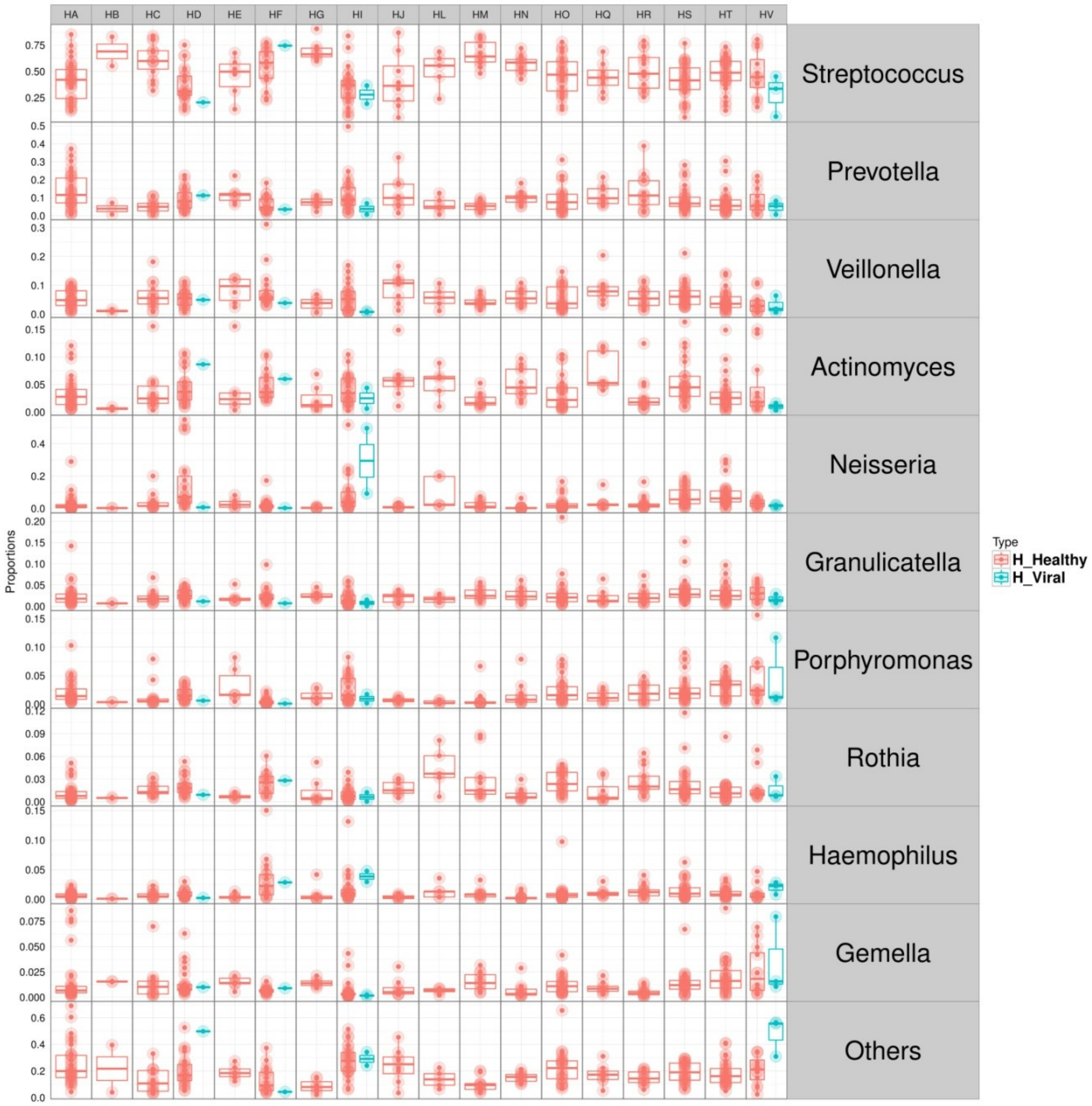


Figure 4.2B – Box plot showing the most abundant genera (n=10) (the rest pooled in the category ‘Others’) and the median abundance in each participant in healthy and viral positive samples.

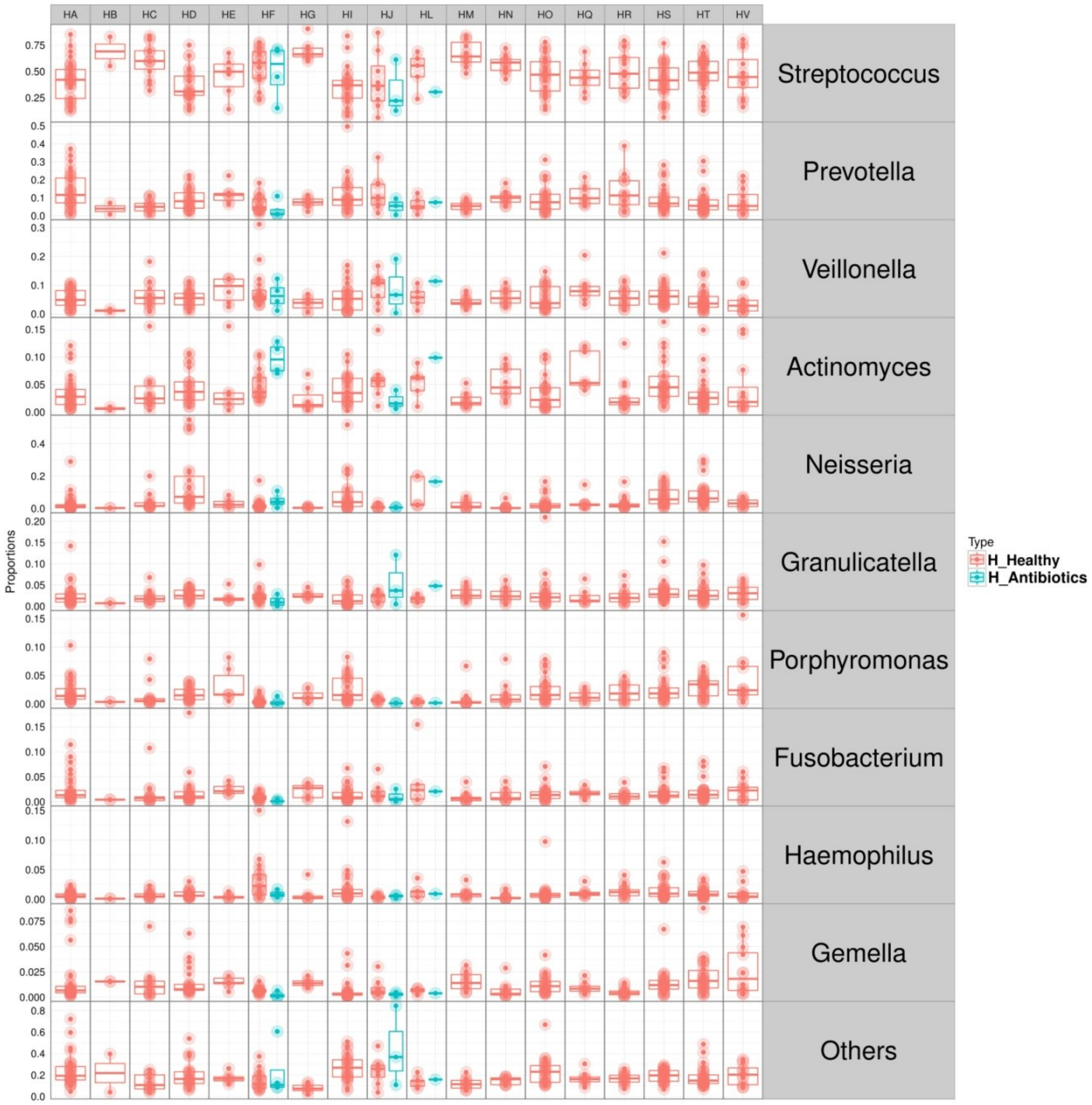


Figure 4.2C – Box plot showing the most abundant genera (n=10) (the rest pooled in the category 'Others') and the median abundance in each participant in healthy and antibiotic treated samples.

Table 4.3 – Summary of the parameter estimates of the linear mixed model (LMM) investigating the abundances (response variable) of the five most abundant genera in unhealthy samples compared to healthy samples (reference category). Significant *P* values are shown in bold.

Unhealthy samples	**Estimate**	**Std. Error**	**df**	**t value**	***P* value**
Streptococcus	-0.05	0.21	311.00	-0.24	0.81
Prevotella	-0.77	0.31	312.00	-2.53	**0.012**
Veillonella	-0.25	0.29	311.00	-0.87	0.39
Serratia	0.23	0.38	313	0.59	0.55
Pseudomonas	0.23	0.38	311	4.75	**<0.001**

OTUs that were the most abundant in healthy samples were also present in the unhealthy samples (cold and viral - Figure 4.3A & B) and antibiotic treated samples (Figure 4.3C) but at differing abundances in participants. For example, participant HM showed an increase in *Streptococcus salivarius* in the cold samples, whereas participant HJ's cold samples showed a decrease in *Streptococcus salivarius*. Therefore healthy and unhealthy samples had differing abundances of OTUs that varied in participants.

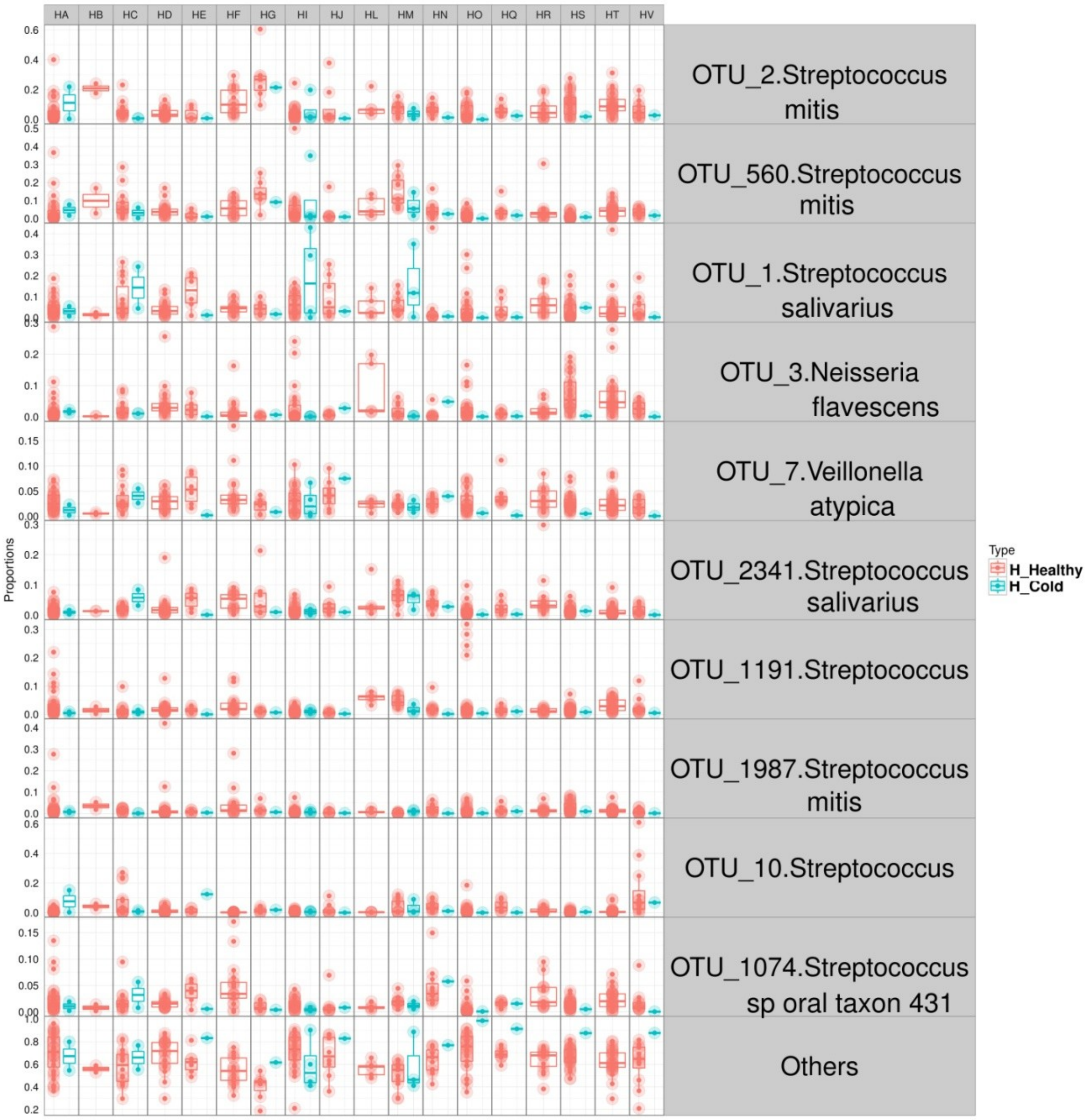


Figure 4.3A – Box plot showing the most abundant operational taxonomic units (OTUs) (n=10) with the rest pooled in the category 'Others' and the median abundance in each participant in healthy and cold samples.

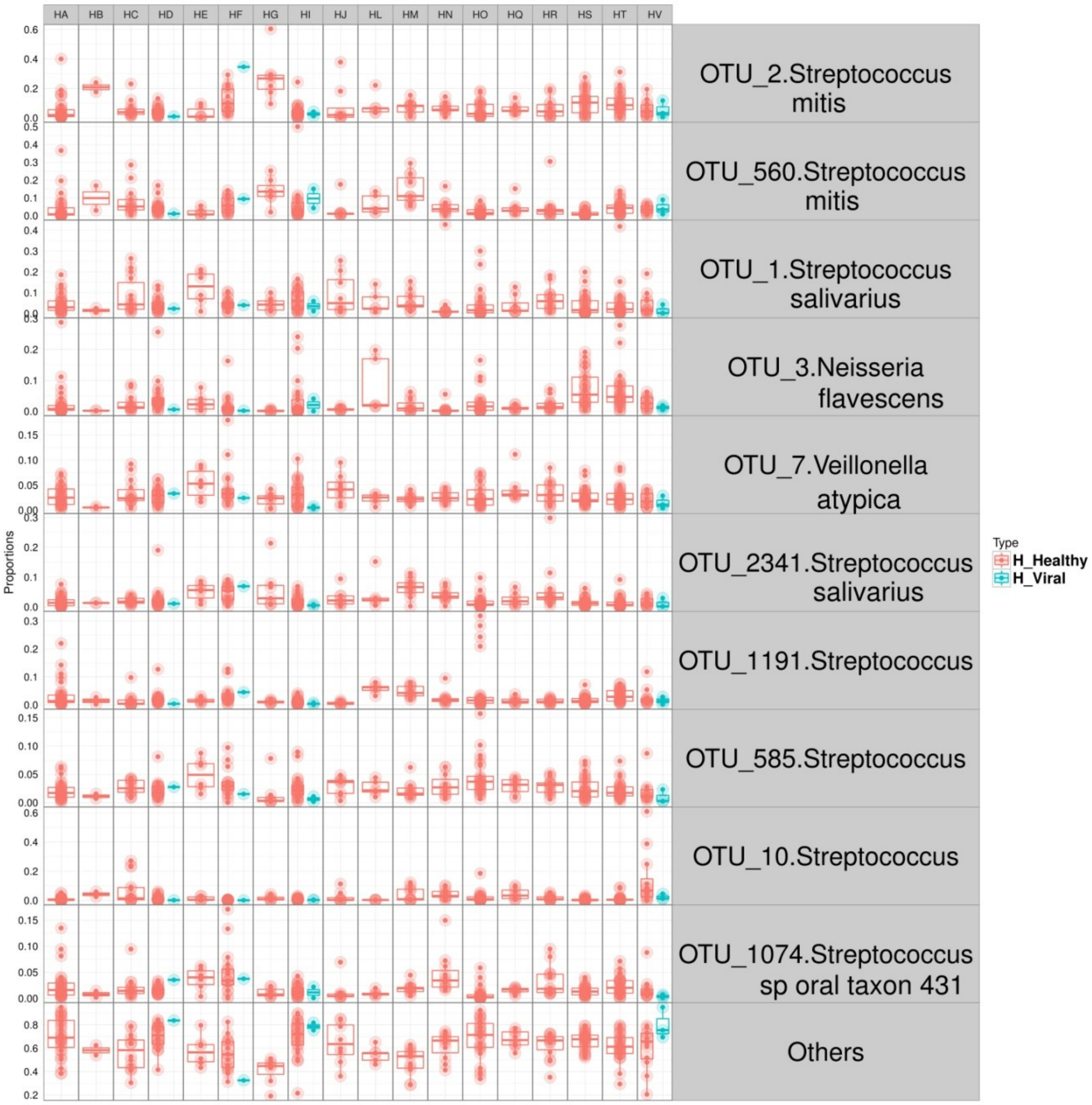


Figure 4.3B – Box plot showing the most abundant operational taxonomic units (OTUs) (n=10) with the rest pooled in the category 'Others' and the median abundance in each participant in healthy and viral positive samples.

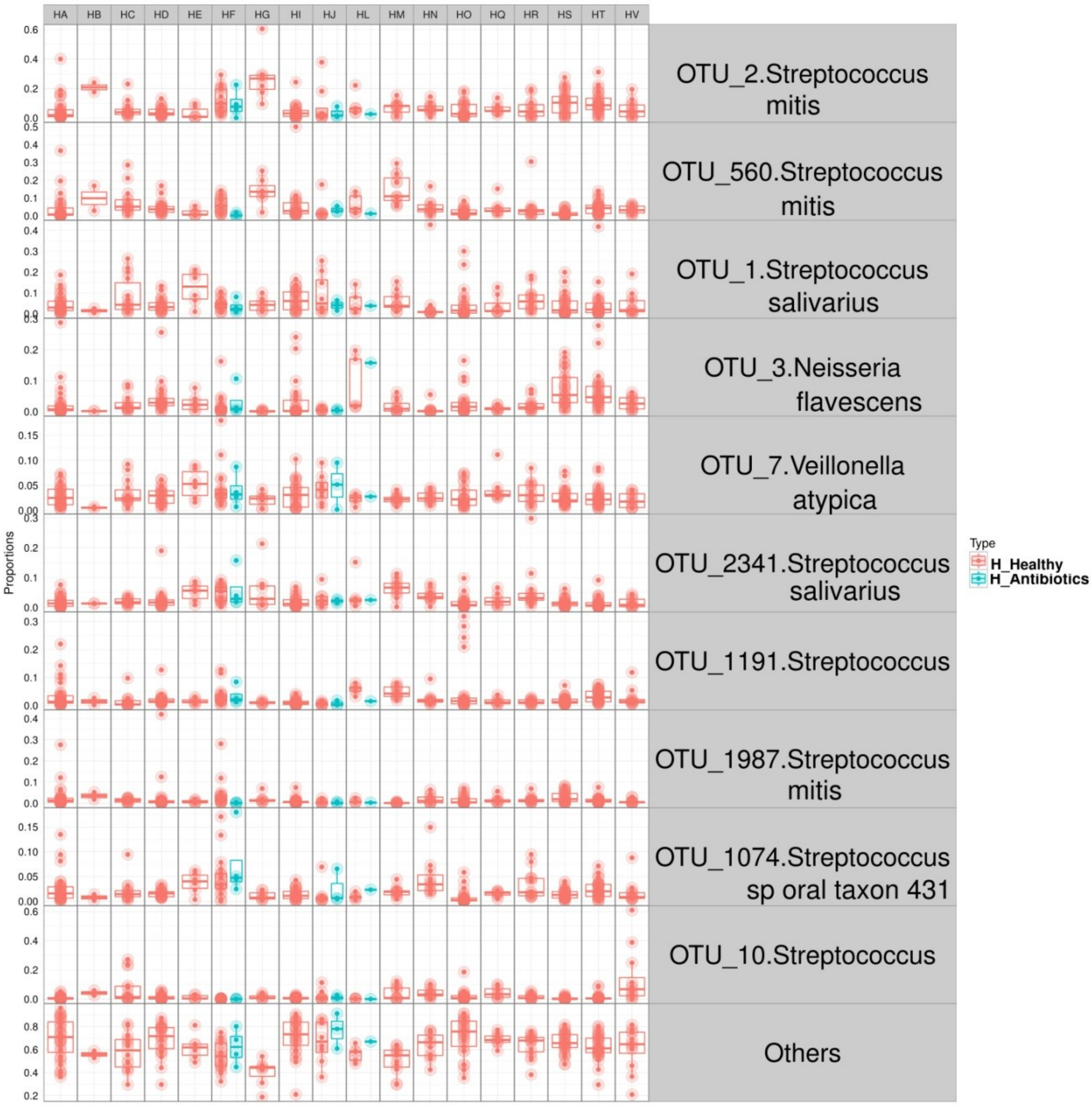


Figure 4.3C – Box plot showing the most abundant operational taxonomic units (OTUs) (n=10) with the rest pooled in the category ‘Others’ and the median abundance in each participant in healthy and antibiotic treated samples.

4.3.2 Local contributions to beta diversity

The changes in community composition between the different health conditions in non-smoking participants were shown through time plots displaying the local contributions to beta diversity (LCBD) (Figure 4.4). Most unhealthy samples belonging to the cold group did have greater dysbiosis than other samples (shown by significant *P* values) suggesting that these samples were unique in community composition. Changes in community composition were shown in antibiotics (n=3, 38% of all antibiotics samples), cold (n=9, 47% of all cold samples) and viral samples (n=3, 43% of all viral samples) showing that community composition was significantly changed in these samples. Some participants such as HC or HN also produced unhealthy samples (in this case categorised as cold) but these samples seemed to have a similar community composition to the healthy samples. To determine the specific changes in community composition of the unhealthy samples and antibiotic treatment, further investigation was done to show the most significant changes in the abundance of OTUs in their unhealthy group category (Table 4.4). Change from a healthy to a cold state was associated with an increase in the abundance of specific *Staphylococcus* OTUs and *Serratia* with a decrease in *Streptococcus* OTUs and *Prevotella*. Antibiotic treatment resulted in a decrease of many OTUs including *Prevotella* and *Veillonella* with an increase in *Enhydrobacter and Actinomyces* whereas viral infections were associated with an increase in the abundance of specific *Moraxella* OTUs with decreases in *Acinetobacter* and *Veillonella* OTUs. Viral infection was also associated with increases in *Haemophilus* and *Neisseria* OTUs.

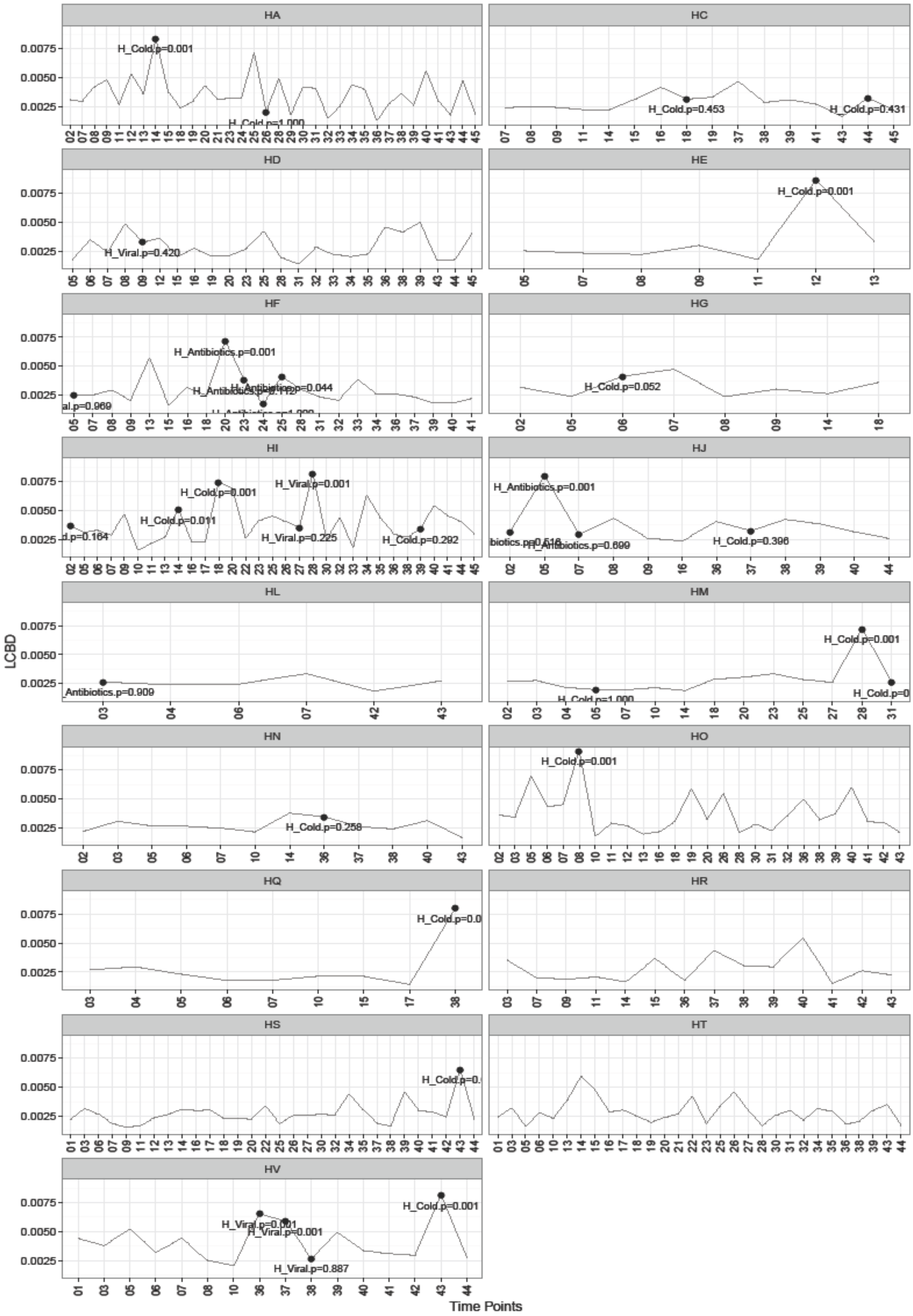


Figure 4.4 - Participant time plots at operational taxonomic unit (OTU) level showing local contributions to beta diversity (LCBD) of samples in regards to health status. Significant *P* values are shown for unhealthy samples with the greatest changes in community structure. Time points refer to the week of when a sample was handed in. Participant HB is excluded due to not having enough samples.

Table 4.4 – The most significant operational taxonomic units (OTUs) (*P* adjusted values < 0.01) in terms of increasing abundance present in the different health groups. Only the 15 most significant OTUs were displayed for the healthy group due to the large numbers of significant OTUs, with the cold and viral samples only having 10 OTUs significantly increasing in abundance.

Most significant OTUs (*P* adjusted value < 0.01) increased in different health states			
Healthy	**Cold**	**Viral**	**Antibiotics**
OTU_53 *Prevotella* OTU_132 unclassified_Prevotella OTU_433 *Streptococcus* OTU_64 *Porphyromonas* OTU_4 *Staphylococcus* OTU_111 Unclassified OTU_52 *Porphyromonas* OTU_56 *Granulicatella* OTU_75 *Fusobacterium* OTU_104 *Veillonella* OTU_53 *Prevotella* OTU_81 *Prevotella* OTU_130 Unclassified OTU_113 *Tannerella* OTU_76 *Acinetobacter*	OTU_751 *Staphylococcus* OTU_1069 *Staphylococcus* OTU_965 *Staphylococcus* OTU_428 *Serratia* OTU_19 *Neisseria* OTU_86 *Sneathia* OTU_1255 *Corynebacterium* OTU_218 *Treponema* OTU_14 *Serratia* OTU_1202 Unclassifed_ *Clostridiales*	OTU_1877 Unclassified_ Pasteurella OTU_926 Unclassified_Neisseria OTU_1542 TM7_genera_incertae _sedis OTU_*283* *Capnocytophaga* OTU_1845 *Moraxella* OTU_1936 *Acinetobacter* OTU_29 *Moraxella* OTU_621 Unclassified_ Pasteurellaceae OTU_17 *Neisseria* OTU_847 *Haemophilus*	OTU_69 *Enhydrobacter* OTU_1871 *Actinomyces*

4.3.3 Comparison of diversity in healthy and unhealthy communities

A reduced diversity (in terms of both species richness and Shannon H index) at OTU level was shown in unhealthy samples (Figure 4.5) with greatest reductions in diversity occurring between the healthy and cold group (richness = *P* < 0.001, Shannon H diversity = *P* < 0.001, Simpson index = *P* < 0.001) and between healthy and antibiotics group (richness = *P* < 0.001, Shannon H diversity = *P* < 0.001, Simpson index = *P* = 0.007). Therefore, the healthy group were the most species rich and diverse followed by viral, cold and antibiotics, showing that the cold and antibiotics group are low diversity communities. The viral group had similar species richness and diversity compared to the healthy group.

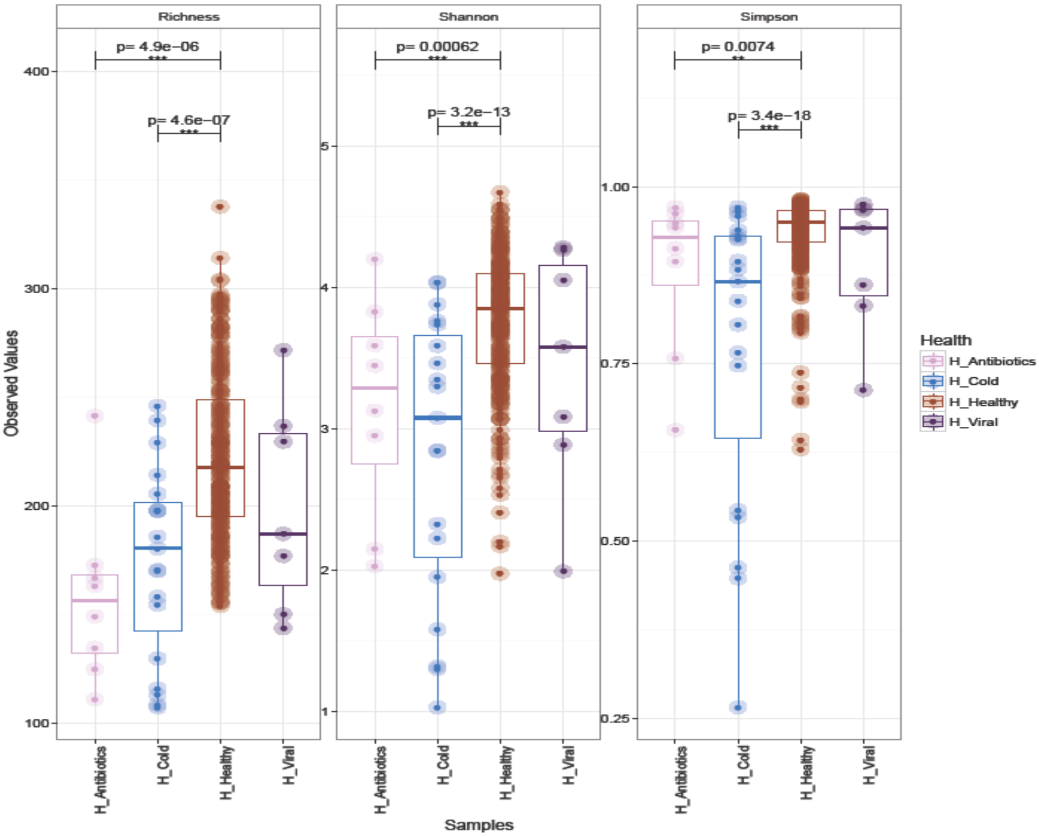


Figure 4.5 – Alpha diversity measures calculated for all samples at operational taxonomic unit (OTU) level. Samples are categorised according to health status and significant test results are done through pair-wise ANOVA with results shown as significant *P* values.

4.4 Discussion

Samples from unhealthy subjects (suffering from a cold or viral infection) and from those receiving antibiotic treatment showed changes in community composition at the three taxonomic levels in comparison to the healthy samples; specifically in the abundances of taxa. A particularly prominent pattern was *Firmicutes* dominance in healthy and unhealthy samples, but unhealthy samples had increases in abundance in *Proteobacteria*. However, increases in abundance in the phylum *Proteobacteria* cannot be solely responsible for explaining disease or infection as there were also various healthy samples where *Proteobacteria* was the most dominant phylum. This could reflect the variation in taxa abundance present in participants as other studies have also shown the healthy oropharynx microbiome to be *Proteobacteria* dominant in abundance (Charlson et al., 2011). However, in a given participant, this ratio was more extreme in unhealthy samples where there was a higher representation of *Proteobacteria* in comparison to their healthy samples. At genus level there was a significant difference in the abundance of *Prevotella* in unhealthy samples, but there did not seem to be a significant difference in the abundance of *Streptococcus* in unhealthy samples, even though other studies have noted a decrease in abundance in *Streptococcus* during respiratory diseases such as chronic obstructive pulmonary disorder that directly affects the oropharyngeal community (Park et al., 2014). As *Streptococcus* was the most abundant genus in healthy and unhealthy samples, this may account for why no significant differences in abundance were observed in respect to health status. *Streptococcus* is a genus that contains various commensal and pathogenic species, so there may have been a decrease in the commensal OTUs and an increase in pathogenic *Streptococcus* OTUs. However, as only some of the *Streptococcus* OTUs could be identified to OTU level, it was difficult to distinguish between the commensal and pathogenic *Streptococcus* OTUs and how they varied in abundance in healthy and unhealthy samples.

Specific health categories also showed changes in community composition showing that antibiotic treatment and colds affect the oropharynx microbiome in different ways. Healthy communities had greater abundances of OTUs belonging to the genera *Streptococcus* and *Prevotella* which were reduced in cold samples.

The cold samples had increases in abundance of specific genera and OTUs showing dysbiosis and changes in community composition in comparison to the healthy samples. Dysbiosis was also observed in the samples from subjects with viral infection, with increases in abundance of specific genera such as *Neisseria* and *Haemophilus*, but these communities were more similar to healthy communities. This suggests that viral infection may increase the abundance of specific OTUs. Antibiotic treatement resulted in a decrease in OTUs common in healthy samples, with exception to a few OTUs belonging to the genus *Actinomyces* that increased in abundance. Therefore, the specific changes in community composition in different health groups (and participants) needs to be further investigated to show exactly how a healthy community changes during a specific disease, infection or disturbance.

In terms of diversity, healthy communities were the most diverse and antibiotics treated ones the least which was consistent with other studies as observed in the microbiome of the GI tract (Jakobsson et al., 2010). However, participants did show variation in healthy samples with changes in diversity and richness, sometimes on a weekly basis, but the dominant members of the community such as *Streptococcus*, *Prevotella* and *Veillonella* were always present. Even though highly diverse communities may be indicative of health, (as healthy communities overall had the highest diversity and species richness) there were some cases of healthy samples having low diversity communities, even though the overall diversity from the participant was not considered low compared to the rest of the participants. But from this sampling overall, significant differences in diversity were observed between the specific health categories. The diversity in different health states needs further investigation and should also take into account each individual participant to determine what OTUs still remain and what OTUs are removed in all healthy samples of high and low diversity communities.

As in most microbiome studies, the causalities are challenging to determine. Do changes in the microbiome drive changes in health status or are they merely consequences of it? There were clear cases in participants reporting cold related symptoms and cold samples showing an increase in abundance in a specific genus or a decrease in abundance in *Streptococcus*. To conclude whether ill health or disease in this case was caused by a specific genus still remains challenging due to lack of daily sampling and metadata collection. However, this study does give insight into the changes that occur in the oropharynx microbiome during infections and disturbances, which is a starting point in determining if changes in the oropharynx are a cause or effect of a certain disease or condition.

4.5 Conclusions

Unhealthy samples had distinct community structures in comparison to healthy samples from non-smokers. These changes included increases in abundance of *Proteobacteria* at phylum level and decreases in abundance of *Prevotella* at genus level. Healthy communities were also the most diverse communities, with the cold and antibiotics samples having the least diversity. This shows that the oropharynx microbiome is affected by infections and antibiotics resulting in altered and low diversity communities.

5 Investigating the stability of the oropharynx microbiome in non-smoking participants

5.1 Introduction

Understanding the factors underlying the structure and composition of microbial communities in individuals is challenging due to interpersonal variation and fluctuations in composition, especially during disease and early development. The first step towards understanding the symbiotic relationships between microbes in the oropharynx with their hosts is to characterise the baseline healthy microbiome and the differences associated with disease as observed in Chapters 3 & 4. This improved understanding of the desired compositional states of the healthy microbiome will then determine which features, when disrupted, are associated with disease. However, the natural complexity of the microbiome, and the presence of intra- and inter-subject variability, further complicates the definition of what a "desired" state may look like for a population or an individual. But understanding these differences between individuals could potentially result in personalised restoration of the microbiome as a future clinical treatment.

Variability in microbial community structure between individuals may arise from natural processes (colonisation history), but factors such as diet, lifestyle, environmental and host changes also play a role (David et al., 2014). A study by Bogaert et al., 2011 showed that variation in healthy individuals is increased at genus and species level making it hard to define a core microbial population due to the diversity of OTUs present between individuals, but also due to the fact that variation occurred between the different seasons. The extent of variation of microbial communities within and between individuals over a period of time is still under investigation, but nevertheless very important as it determines the true microbial components of a community, as well as identifying the microbes that are not regular members of the community. From this, it is possible to explore how ones microbiome differs from another whilst examining the rich diversity of the community, and investigating whether these differences are natural changes or a result of a specific disturbance. It is especially important to understand natural variation over time to understand the stability of the microbiome and whether instability increases the risk of pathogen susceptibility.

Stability can be described as the ability of communities to withstand disturbances as well as the similarity of communities in terms of taxon presence and abundance. Stability can be measured by microbiome time series projects which have the advantage of recording specific metadata and linking microbial dynamics to host behavior.

Understanding the variation in the healthy oropharynx has great importance. Further exploration of the community composition of the oropharynx microbiome improves the ecological understanding of these communities (Fierer et al., 2012). Understanding the relationships between microbes of the upper respiratory tract (URT) during perturbations is anticipated to provide insights into the pathogenesis of URT infections. It will also aid understanding of the effects of stability and resilience on these communities and the process of recovery in the oropharynx. Therefore, this chapter explores the extent of variation in microbial communities in non-smoking participants. This was investigated in different health states whilst observing the changes that occur to the oropharynx microbiome before, during and after disturbances to determine the stability and resilience in individual microbiomes. The objectives of this chapter can be broken down into the following questions: What is the variation in oropharyngeal communities? How does the community change before, during and after a disturbance? How stable is the microbiome in non-smokers and how quickly can the microbiome recover? The hypothesis is that non-smoking participants have a stable oropharynx microbiome that can recover quickly from disturbances.

5.2 Methods

5.2.1 Exploring variation in the oropharyngeal community

All statistical analysis was performed in R (version 3.1.2). To assess the variation of the oropharynx microbiome between the different health groups, a NMDS plot using Bray-Curtis distance with variance ellipses (*Chapter 2.11.2*) was produced using Vegan package (version 2.4.0) (Oksanen, 2013). A NMDS plot at OTU level was produced showing similarity of community composition in samples with regards to health status. The variability in microbial community structure in regards to health status was investigated at OTU level using betadisper() in Vegan (version 3.1.2) and these distances were used to quantify the extent of variation in each health group through a *betadisper* box plot (Bray-Curtis distance). This measured the distance of each individual sample to that group's centroid (mean) allowing for comparison between the different health groups. Samples that had the greatest distance away from the group centroid were considered to have a different community structure as opposed to samples with a shorter distance to the centroid. ANOVA was performed on *betadisper* distances where the distance of each individual sample to the centroid (mean) of each different health group was assessed and the means compared to determine if variation changed according to health status. Tukey's HSD test was used as a post-hoc test after ANOVA to determine which groups differed in variation.

5.2.2 Exploring the stability of the oropharynx microbiome

Using distances from *betadisper*, individual participant graphs were plotted to show the distance of each sample to that participant's centroid (mean). These stability plots allowed comparison in distances of all samples within and between participants which could determine the degree of change in the microbial community during the sampling period. This would also determine whether fluctuations could be caused by disturbances such as changes in health status. Overall this would give an indication of the stability of a participant's microbiome as peaks (greater distances) over the sampling period can be identified. A participant with higher peaks (corresponds to how different the microbiome is at different time points) is thought to show a more unstable microbiome compared to a participant with smaller consistent peaks. For each participant, the difference in distance of the individual sample to that participant's centroid was plotted, making these plots comparable between participants. Community stability was quantified by calculating the coefficient of variation (ratio of standard deviation to the mean) for each individual participant using the distances from *betadisper*.

The temporal stability in communities was statistically tested to observe if the community structure changed before, during and after a cold/viral disturbance through a linear mixed model (LMM) using distances from *betadisper*. This would determine if the community changed only during symptoms or if changes occurred before symptoms were present. Sampling week (to accommodate for potential temporal trends) was fitted as a fixed effect and participant ID as a random effect using lme4 (version 1.1-9) and MASS (version 7.3-44) packages. This would also give an indication of how quickly the community would respond to the disturbance and if changes were still apparent to the microbiome after the disturbance had cleared and symptoms were no longer present.

To determine a possible relationship between stability and diversity, the coefficient of variation was tested against diversity variables (species richness) using spearman correlation to determine any correlation between the overall stability of each participant and mean species richness of their communities. The resilience (resistance to disturbances) was also tested by using the number of cold/viral samples against species richness to characterise its resilience to perturbations (*Chapter 2.11.3*).

5.3 Results

5.3.1 Variation in the oropharynx microbiome in health and disease

The variation in communities between the different health groups in non-smokers was explored (Figure 5.1A). Healthy samples clustered together reasonably tightly although dispersion of some samples was observed, the differences of which could be due to natural variation from within and between participants. As there were uneven samples sizes between the different health categories, accurate investigation of the similarity in community structure in different health states was challenging. However from visual observation, the unhealthy samples were more disperse in that they were located further away from each other in comparison to the healthy samples that were clustered closer together; this shows that unhealthy groups overall were more variable in community structure compared to the healthy group. This was shown in the *betadisper* box plot, displaying which health group had the most variation (Figure 5.1B). From this box plot the cold samples showed to have the most variation between samples due to the dispersion of samples observed. The samples from the cold group were shown to be variable with a greater dispersion within samples compared to the healthy group. This would suggest samples from participants with the cold had greater changes in community structure compared to the healthy samples, as well as each cold sample having distinct communities.

ANOVA testing showed that the extent of variation changed in different groups ($P < 0.001$) with significant differences between the healthy and cold groups ($P < 0.001$) but not between the healthy and viral groups ($P = 0.5$), and healthy and antibiotics group ($P = 0.6$). Therefore this showed that variation of microbial community structure is affected by health status, with the cold group having the most variability and changes in community structure between samples.

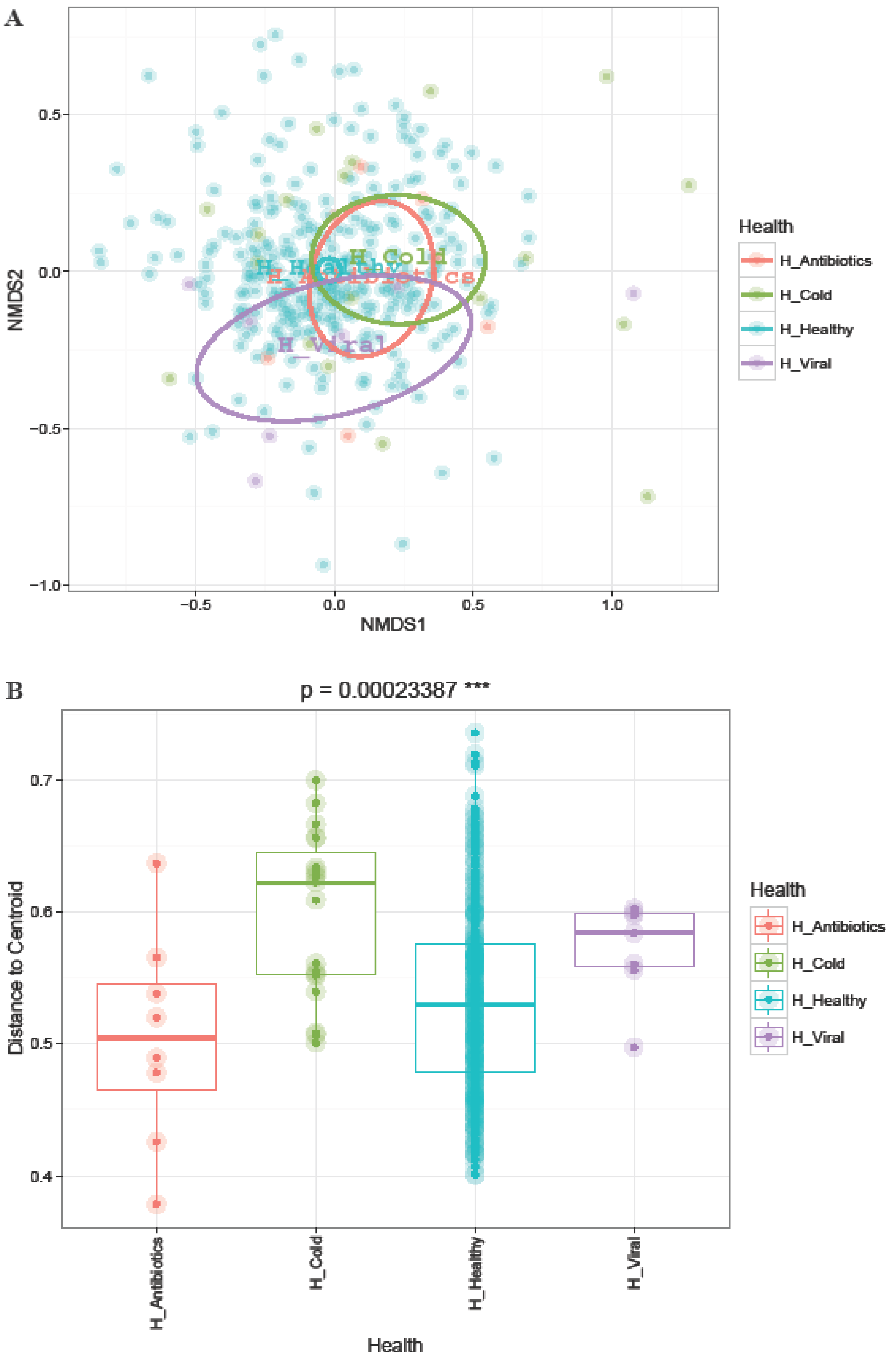


Figure 5.1 - Variation in microbial community structure in healthy and unhealthy groups at operational taxonomic unit (OTU) level. The health status is indicated by different colours in the key. Figure 5.1A shows the variance ellipse and clustering from the non-metric multidimensional scaling (NMDS) plot in each health group and Figure 5.1B shows the median and distribution of distance (using *betadisper*) in samples from the centroid of each group. The *P* value (ANOVA) shows a significant difference in variability between groups.

5.3.2 Changes in the community structure before, during and after a disturbance

5.3.2.1 Individual responses to disturbances

The changes in community structure from disturbances for participant HI was observed through a NMDS plot (Figure 5.2A). This plot showed clusters of healthy samples that were considered stable with similar community composition and little changes in the communities on a weekly basis. However outliers were observed away from the cluster, some of which could be accounted for through changes in health status as participants recorded symptoms of illness for some of these samples. Participant HI had various cold symptoms over the duration of sampling with 2 samples of a viral infection (Rhinovirus). A taxa plot showing community composition in participant HI (Figure 5.2B) revealed a single cold sample showing an increase in abundance in the genus *Haemophilus*; there was also an increase in *Neisseria* during Rhinovirus infection. However viral samples looked similar to the healthy samples, indicating that in this case viral infections did not have a major impact on the microbiome. There seemed to be a distinct change in community structure during disturbances from the cold samples (negative for viral infections) with little changes in community structure pre-disturbance. However, after a cold, the microbial community recovered quickly in that it returned back to a normal healthy state (within a week) where the post-disturbance sample returned back to a baseline representing health.

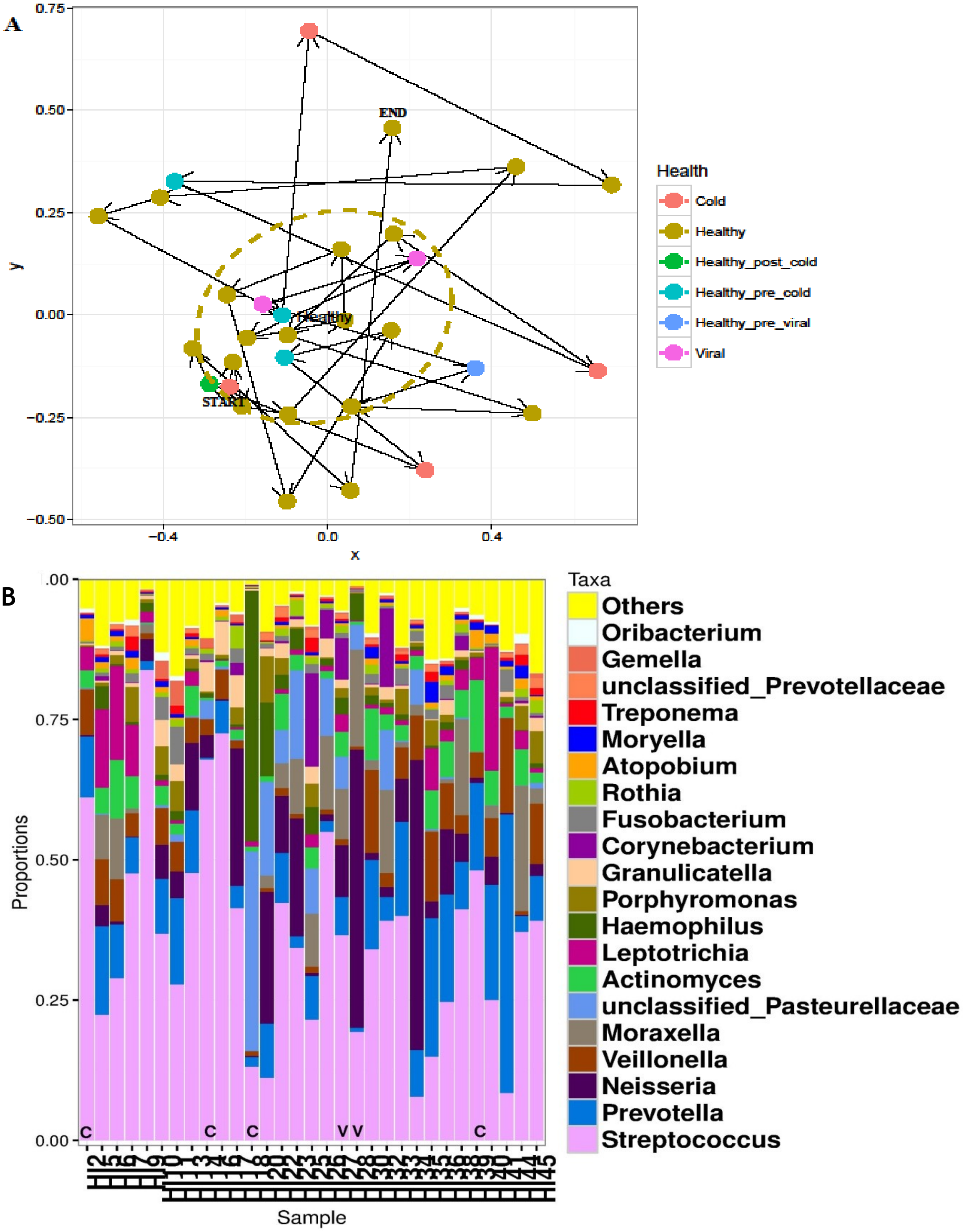


Figure 5.2 - Non-metric multidimensional scaling (NMDS) plot at operational taxonomic unit (OTU) level from participant HI (Fig. 5.2A) showing similarity of all samples in regards to health status. Figure 5.2B shows the taxa plot displaying the relative abundance of the top 20 most abundant genera. Week numbers are represented as sample numbers with cold (C) and viral (V) samples shown. Only the top 20 most abundant genera were chosen to give a clear visualisation of the most abundant taxa members in the community.

5.3.2.2 Individual responses to antibiotic treatment

The effects of antibiotic treatment on the oropharynx microbiome were also investigated through NMDS plots. Participant HF (Figure 5.3A) was on a month long prescription of tetracycline antibiotics for treatment of acne whereas participant HJ (Figure 5.3B) was on a 6 week prescription of erythromycin, also for treatment of acne. In addition, participant HF suffered from a viral infection (Respiratory Syncytial Virus) which showed a similar community composition to the healthy samples. For participant HF there was a divide in samples between the healthy and antibiotic samples, which also occurred for participant HJ. As participants were given different antibiotics, it is likely that specific members of the microbiome of each participant were affected differently by the antibiotic treatment. This is shown in the NMDS plots where participant HF had all antibiotic treated samples within the healthy centroid, whereas participant HJ had all antibiotic treated samples outside the centroid. This shows that antibiotic treatment affects participants very differently. These changes were easier to observe when looking at the individual participant's taxa plots where increases in genera *Pseudomonas* and *Actinomyces* were found for participant HF (Figure 5.3C) and increases in *Actinomyces* and *Acinetobacter* for participant HJ (Figure 5.3D). However, some antibiotic treated samples were similar in composition to the healthy samples and this was seen in both participants. The first sample from participant HF on antibiotic treatment (HF20) was dramatically altered while the remainder of the antibiotics samples showed a similar community composition to the healthy samples. In participant HJ, the first antibiotics sample (HJ2) was more similar in community composition to the healthy samples, in comparison to the second sample received (sample HJ5) where a distorted community structure was observed. The next sample received (HJ7) showed an increase in the genus *Streptococcus* (which is a dominant member of the healthy oropharyngeal community) showing that this sample also resembled a healthy community. This sample was also similar in community composition to sample HJ8, which was a healthy sample received one week after antibiotics use. This data showed that antibiotic treatment did impact the microbiome, but the microbiome was able to restore itself whilst on treatment and recover quickly after treatment (usually within a week) where the community returned back to a state similar to samples obtained pre-treatment.

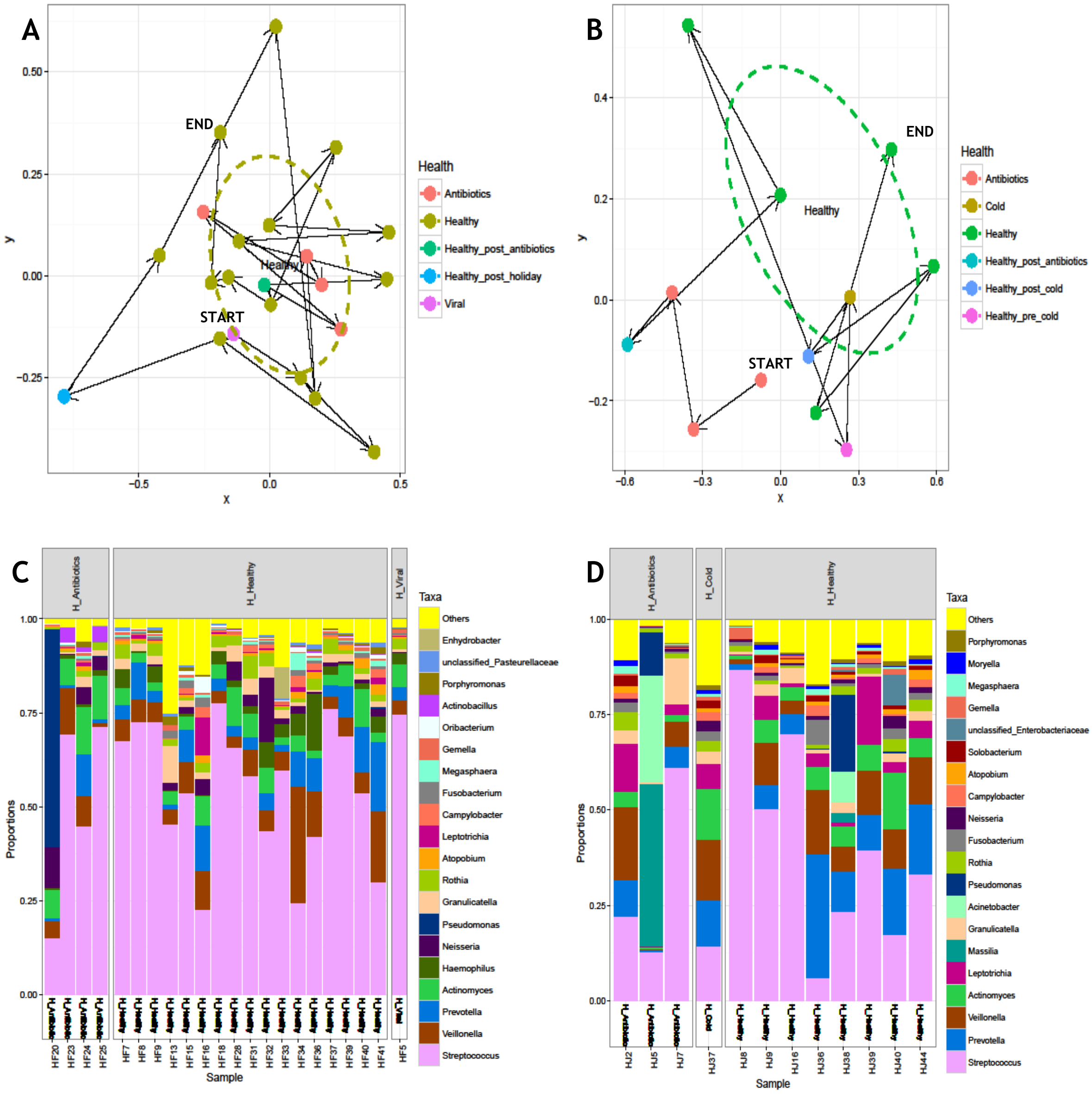


Figure 5.3 – Non-metric multidimensional scaling (NMDS) plots of participants HF (Fig. 5.3A) & HJ (Fig. 5.3B) at operational taxonomic unit (OTU) level demonstrating the degree of variability over the sampling period (variance ellipses are calculated from just the healthy samples). In both graphs, the starting and end point are marked and the arrows show the direction of sampling on a weekly basis. Health status for each plot is depicted through different colours in the legend. Taxa plots display the relative abundance of the top 20 most abundant genera in Participants HF (Fig. 5.3C) and HJ (Fig. 5.3D).

5.3.3 Stability of the oropharynx microbiome

Stability plots for each participant were produced to determine fluctuations throughout the sampling period (Figure 5.4). It was found that some participants such as participant HC were more consistent in their sampling as each sample had a similar distance which reflected that there was a similar change in community structure on a weekly basis. Other participants such as participant HS had far more deviations with greater peaks suggesting greater changes in community structure in comparison to other samples, which not all could be accounted for by changes in routine or health status. Therefore, in this case it was concluded that participant HC had a more stable microbiome than participant HS. The value of stability for each participant was obtained by calculating the coefficient of variation (also shown in Figure 5.4) which reflects the variation from the long-term mean and so is considered a suitable summary statistic as an indication of how stable participants' oropharyngeal communities are. The coefficient of variation for participants were variable, with some participants having higher values indicating more changes in community structure on a weekly basis which could be regarded as a more unstable state. In regards to these values, participant HG had the highest coefficient of variation which could be regarded as having a more unstable microbiome (with greater changes in the microbial community structure between weeks and therefore greater variation overall). In contrast, participant HC was considered to show the most stable microbiome due to the lowest coefficient of variation and similar minimal changes in the microbial community on a weekly basis (there was not much change in microbial community structure over the sampling process). As the majority of participants had reasonably low coefficients of variations with similar distances, the healthy oropharynx microbiome for each participant was considered stable, as there did not seem to be major differences in community structure within participants on a weekly basis and most deviations in community structure (not all) could usually be linked to a change in health status.

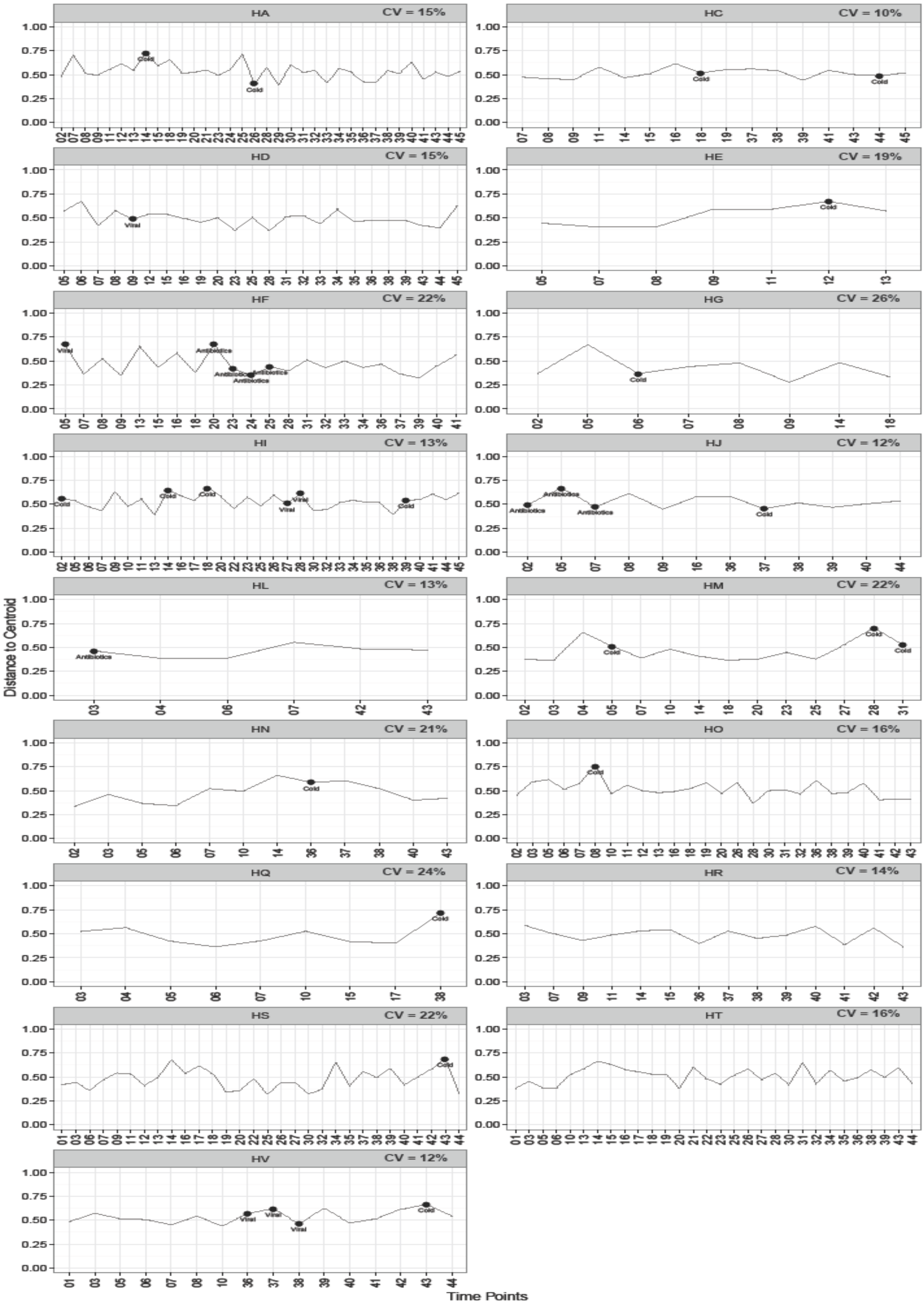


Figure 5.4 - Individual stability plots for each participant showing the distance of each sample (from *betadisper*) to the participant's group centroid as well as coefficient of variation (CV) values. Disturbances in health status are shown by either antibiotics, cold or viral labels (all other samples are considered healthy) with the time line of swabbing showing the weeks of when a sample was submitted. Participant HB was omitted due to not having enough samples.

5.3.3.1 Investigating the temporal stability before, during and after a disturbance

To observe if the community structure changed before, during and after a cold/viral disturbance, a LMM was performed using the distances from *betadisper*. The results showed that there was a significant difference in community composition in communities one week before the disturbance (altered state due to the cold and positive viral samples) (*P* < 0.001) and during disturbance (*P* < 0.001), but no significant changes were observed in the community one week after the disturbance (*P* = 0.3266) (Table 5.1). This demonstrates that changes to the bacterial community were apparent one week before symptoms, with communities returning back to normal one week after symptoms were no longer present. This highlights the quick recovery from the disturbance. From the stability plots the communities showed strong resilience in returning rapidly towards the long-term mean composition i.e. shorter distance to the centroid showing the oropharynx microbiome to be resilient in that it responds and recovers quickly from a disturbance.

Table 5.1 – Parameter estimates from a linear mixed model (LMM) where distance (from the *betadisper* stability plots at operational taxonomic unit (OTU) level) was tested against infection variables. Healthy samples are the reference category. Significant *P* values are shown in bold.

Status of infection	**Estimate**	**Std. Error**	**df**	**t value**	***P* value**
One week before symptoms	0.05329	0.01428	298.6	3.733	**0.000226**
During symptoms	0.06548	0.0122	303	5.368	**<0.0001**
One week after symptoms	0.0207	0.02107	299.6	0.982	0.326694

5.3.3.2 Is diversity linked to stability?

The diversity (as measured by species richness and Shannon H index) was shown to differ in each participant, especially at lower taxonomic levels. A change in diversity has been linked to a change in health status, but the link between diversity and stability remains uncertain. To determine a possible relationship, the coefficient of variation was tested against diversity variables (Table 5.2) using spearman correlation to determine any correlation between the overall stability of each participant and mean diversity (species richness) of their communities. The results from the spearman correlation and significance testing for stability (coefficient of variation) and each diversity variable are as follows: $CV_{meanphyla}$ (-0.1672, P = 0.5211), $CV_{meangenera}$ (-0.4064, P = 0.1054), $CV_{meanOTU,}$ (-0.053, P = 0.8398), $CV_{number\ of\ cold/viral\ disturbances}$ (-0.0463, P = 0.8599). There was no obvious relationship between the diversity and stability of the microbiome of participants at the three taxa levels. Therefore, from this sampling, diversity is not a factor to drive stability as some participants that had high diversity values also had high coefficients of variation which could be seen as having more variable and unstable microbiomes.

Table 5.2 – Mean species richness for all participants and the numbers of disturbances identified as used in the diversity/stability test. Participant HB is excluded from this table due to not having enough samples.

Participant	Mean number of phyla	Mean number of genera	Mean number of OTUs	Number of disturbances/antibiotics throughout sampling weeks (n= 45)		
				Cold	Viral	Antibiotics
HA	8	49	243	2	0	0
HC	8	46	212	2	0	0
HD	9	52	250	0	1	0
HE	8	43	210	1	0	0
HF	7	40	184	0	1	4
HG	8	40	204	1	0	0
HI	8	45	218	4	2	0
HJ	8	41	184	0	0	3
HL	8	47	211	0	0	1
HM	7	39	190	3	0	0
HN	8	43	216	2	0	0
HO	8	43	206	1	0	0
HQ	8	45	226	1	0	0
HR	9	50	234	0	0	0
HS	9	45	218	1	0	0
HT	9	48	235	0	0	0
HV	8	43	195	1	3	0

5.4 Discussion

The results show that variation in the microbial community structure in the oropharynx microbiome occurred within and between participants over the timescale of sampling. In most participants, the changes in the microbiome over time were usually small fluctuations around a relatively stable microbial community. Each participant had a distinct oropharynx microbiome in that there was high beta diversity between participants, with abundances of taxa varying within and between participants (Fierer et al., 2012). This could also directly contribute to each participant having its own stability pattern; however variation did occur within participants over time. Variation in bacterial community structure is expected due to changes in host and environmental conditions as well as external factors. As each participant’s oropharyngeal community varies over time due to lifestyle, health, age, diet and culture, it is crucial to discriminate between the normal perturbations of the human microbiome and changes in response to a disturbance. It is this intra- and inter-subject variability that makes it more difficult to determine a core microbiome, especially as not all healthy samples had similar abundances or presence of specific OTUs. Even though the majority of healthy samples had a dominant member of the community (*Streptococcus* or *Prevotella* at genus level), healthy samples were variable in terms of prevalence and abundance of taxa and community diversity on a weekly basis. This makes it challenging to give an accurate definition of what a healthy or desired state is for the oropharyngeal community.

Viral infection (Rhinovirus and Respiratory Syncytial Virus) were not associated with changes to the microbiome. However, due to the small number of samples from subjects with viral infection, this needs to be further explored, especially due to conflicting results published from other studies where viruses do impact the oropharynx microbiome (Allen et al., 2014). There is the possibility that the cold group did contain viruses in some samples, but these viruses may not have been picked up from the respiratory screen because only a selection of viruses were tested as well some participants not handing in a viral swab when symptoms appeared. Therefore cold samples that were negative for viruses tested in the respiratory screen could either be other viruses not detected or could be samples that had symptoms occurring from the external environment such as irritants or pollen levels. Regardless, the cold samples did show changes to the microbiome, with the majority of changes occurring to the community structure before and during the symptoms. Even though sampling occurred weekly, changes to the microbiome were observed one week before the symptoms occurred. The reasons for this may include time required to overcome the existing microbial community; this includes the time required for the bacteria to multiply to the level required to produce disease and this multiplication stage may not result in any symptoms. This pattern was evident for most participants when suffering a cold, where the community seemed to be disrupted one week before symptoms were noted. One sample had increases in bacteria such as *Neisseria* whereas other samples had increases in *Haemophilus* which were also present during the cold. However this will also depend on the type and severity of infection and how accurate participants were in recording symptoms as participants may only have recorded symptoms on sampling days and not have recorded any symptoms observed in between sampling days.

Not surprisingly, antibiotic treatment altered the oropharynx microbiome and typically resulted in low diversity communities, even when the site of action for the antibiotics was not the oropharynx. This was also observed in other studies (Jakobsson et al., 2010) showing that antibiotic treatment does result in low diversity communities. Antibiotic treatment eliminates specific groups of bacteria (due to antimicrobials targeting a limited range of bacteria) and can therefore be expected to be changing the microbial populations. One study (Santiago et al., 2014) also reported the impact of antibiotic treatment on the gut microbiome resulting in increased microbial load specifically gram negative bacteria showing the effects of antibiotic treatment on the microbiome are more complex than previously thought. Post-antibiotic treatment, the oropharynx microbiome recovered quickly showing that it was resilient in that the diversity and abundances of taxa were restored again to a state similar to pre-treatment. Resilience is usually measured in terms of taxonomic composition; however, there should also be consideration in measuring function before, during and after a disturbance, as even though changes in taxa occur, this may be even more important to show that the community's role has not changed after a disturbance (Jakobsson et al., 2010). Therefore despite disturbances affecting the microbiome, the microbiome in all participants was considered stable in that even though fluctuations in the community structure were apparent, these fluctuations were generally minor with the dominant taxa still present. Microbiomes of other body sites such as the skin in healthy participants have also shown to be stable (Oh et al., 2016) showing that stable microbiomes may be an indicator of health. During specific disturbances the community structure changed, but the microbiome was resilient in that it recovered from these disturbances quickly, usually within a week.

The diversity-stability relationship was also assessed across participants and some participants' communities were more stable than others. A more diverse community is expected to be more resistant to invasion by pathogens (Konopka,

2009) and therefore more stable. As some participants had high coefficients of variation, as well as high species richness values, no correlation in this case was found between diversity and stability. However, various studies in macroecology have reported that diverse communities tend to be the most stable ones (Loreau & de Mazancourt, 2013) (Lozupone et al., 2012), so this still remains unknown for microbiome studies, especially due to the small sample size of unhealthy communities in this study. Therefore, the role of diversity in different health states is complex and challenging, especially as it is still unknown if a low diverse state is the cause or consequence of the disease, and how this affects stability. Even though the oropharynx microbiome was found to be stable, determining the stability of the oropharynx microbiome over a longitudinal period of time was a challenge as samples were not present every week as participants did miss some weeks of swabbing. Participants may have also inaccurately reported symptoms due to forgetting, being rushed or reporting the wrong symptoms. This could explain why some samples had very distinct, different communities for which there were no obvious explanations. Overall though, it was found that each participant had a distinct oropharynx microbiome that was considered stable and resilient to disturbances at the genus and OTU level.

5.5 Conclusions

This chapter shows that healthy participants have stable and resilient oropharynx microbiomes. When faced with a disturbance the microbiome becomes altered with changes in abundances in taxa. However, the microbiome was quick to recover from these disturbances and return back to a normal and healthy state usually within a week. Healthy samples from participants were dominated by few taxa, with other taxa residing at lower abundances, but there was also variation in the abundances in the dominant taxa within and between participants. By understanding the extent of variation in healthy participants and bacterial abundances before, during and after infection, it may be possible in the future to identify respiratory diseases (from further investigations and controlled studies) and possibly restore health through investigating how to manipulate microbial populations.

6 Comparing the community structure and stability of the oropharynx microbiome of non-smoking to smoking participants

6.1 Introduction

Smoking has been associated with many risks; exposure to cigarette smoke results in changes in the host's environmental conditions, disruption in the body's natural defence mechanisms and impaired or reduced host immune responses against infections (Van Zyl-Smit et al., 2014). There is also an increased risk of respiratory tract infections (Bagaitkar et al., 2008) through disruption of commensal bacteria potentially providing an opportunity for colonisation and growth of pathogenic microorganisms.

The effect of smoking on any microbiome is a still a topic under investigation; the introduction of next-generation sequencing (Petrosino et al., 2009) is changing this by enabling quick in-depth analysis of communities from various body sites. Studies using this technology have started exploring the effects of smoking on the oropharynx microbiome as the oropharynx is one of the first sites of contact for cigarette smoke. Results have already shown distinct communities between smokers and non-smokers (Charlson et al., 2010). The healthy oropharyngeal community consists of *Firmicutes* and *Bacteroidetes* as the dominant phyla, and *Streptococcus*, *Prevotella* and *Veillonella* the most abundant genera (Lemon et al., 2010). In contrast and comparison with non-smokers, the dominant phyla in the oropharyngeal communities of smokers are *Actinobacteria* and *Bacteroidetes* with increases in the abundance of the genera *Megasphaera* and *Fusobacterium* as well as pathogenic *Streptococcus* species (*S. pneumoniae*). In addition, certain commensal *Streptococcus*, *Prevotella* and *Peptostreptococcus* species have a reduced abundance in smokers' oropharynx microbiomes (Charlson et al., 2010). Various pathogens isolated from smokers communities have been associated with diseases such as periodontitis (Zeller et al., 2014), tonsillitis (Bagaitkar et al., 2008), chronic obstructive pulmonary disorder (Erb-Downward et al., 2011) and tuberculosis (Van Zyl-Smit et al., 2014).

The changes in the oropharynx microbiome associated with smoking could also affect the stability properties of the community; longitudinal sampling would ideally show how the community changes over a defined period of time and in response to fluctuations. These changes in a healthy community structure may affect the overall stability of the microbiome, something that is relatively unexplored. These changes are important as they may determine the outcome or recovery from a disturbance or infection, as well as be responsible for making communities more susceptible to infections. There have been no studies exploring the longitudinal effects of smoking on the microbiome on a weekly basis and recovery from infections in comparison to non-smokers. This would determine if communities from smokers have specific bacteria responsible for increased susceptibility to infection, and also if smoking results in unstable communities. Therefore this chapter aims to determine the distinct differences in microbial community structure between healthy participants (non-smokers) and smokers. The objectives for this chapter are as follows: to compare the oropharynx microbiome of non-smoking healthy participants to smokers, to determine the changes that occur to the oropharyngeal community during a disturbance in smokers, to determine if healthy participants have more stable microbiomes compared to smokers and to determine if smokers have a longer recovery time from a disturbance (from the cold and viral samples only) compared to healthy participants. The hypothesis is that smokers will have a changed microbial community structure in comparison to non-smoking participants; the smoker’s microbiome will be unstable with increased differences in community structure in samples from the same participant and the community will take longer to recover from a disturbance.

6.2 Methods

6.2.1 Initial analysis

In total, 490 oropharyngeal samples were obtained from 30 participants; 313 samples from non-smoking participants (n=18) and 177 from smokers (n=12). Samples from the non-smokers and smokers were collected over two different years (2013 for non-smokers and 2014/2015 for smokers). From the non-smoking healthy participants’ samples, 34 samples were identified as unhealthy samples which were categorised into the following groups: cold (n=19 from 12 participants), antibiotics (n=8 from 3 participants) and viral (n=7 from 4 participants). The cold group includes samples from participants self reporting symptoms of a cold but were negative for the respiratory screen at Gartnavel hospital as described in section *Chapter 2.2.1*. The antibiotics group included any participants undergoing antibiotic treatment and the viral group included only symptomatic samples with confirmed viruses from the respiratory screen. For the smokers, 32 samples were identified as unhealthy samples which were categorised into two groups, cold (n=14 from 6 participants) and antibiotics (n=18 from 2 participants). There were no positive viral samples from the smokers group. The rest of the smoker's samples were considered as healthy samples – a healthy sample from a smoker is classified as one without any symptoms of disease or any changes in the normal routine of the participant.

6.2.2 Microbial community composition

Statistical analysis was performed in R software (version 3.1.2). Where appropriate, the abundance data were normalised (McMurdie & Holmes, 2014) before specific analyses. Linear mixed models (LMM) were constructed through lme4 package (version 1.1-9) using log transformed abundances on all samples from non-smoking participants and smoking participants (healthy and unhealthy) to determine differences in abundance of certain taxa between smokers and non-smokers. Week was used as a fixed effect as eventhough sampling for non-smokers and smokers occurred in two different years, the weekly sampling procedure that occurred in non-smokers and smokers was the same. Participant ID was used as a random effect.

Microbial compositional structure was assessed using non-metric multidimensional scaling plots (NMDS) to determine the differences in community composition in regards to smoker and health status. To determine the difference in community composition between non-smoking participants and smokers, an NMDS plot was produced from only the healthy samples of non-smokers and smokers. The Bray-Curtis dissimilarity index was applied which considers bacterial taxon abundance. Unweighted UniFrac distance analysis from the Phyloseq package (version 1.17.2) was also used (McMurdie & Holmes, 2013) which takes into account the phylogenetic distances (relatedness) of taxa through presence or absence, but without accounting for their proportional representation. A covariance ellipse using Ordiellipse() and veganCovEllipse() in Vegan (version 2.4.0) was added (95% confidence interval calculated from the standard errors of samples from the mean of each group) with the centroid of the ellipse representing the group mean. Covariance for each group was calculated using cov.wt() and the shape of the ellipse was defined by the covariance within each group - the bigger the ellipse, the more variability in community structure in samples within the group. Significant difference testing for different groups at OTU level was done using PERMANOVA through adonis() in Vegan (version 2.4.0). The significant difference testing for clustering was corrected using the command strata to take into account repeated sampling from participants.

To find OTUs that are significantly different between non-smoking and smoking participants, DESeq2 package (version 1.24.0) in R (version 3.1.2) (Love et al., 2014) was used as before. This determined the specific OTUs responsible for distinguishing between a non-smoker and smoker's community by identifying the OTUs with the most significant differences in abundance between the two groups. This uses a negative binomial GLM fitting on the abundance data (OTU frequencies) and empirical Bayes to shrink OTU-wise dispersions to identify OTUs that have the log-fold changes between different conditions. Differential expressions are tested by performing a Wald test on shrunken log-fold changes and are adjusted for multiple comparisons. This results in adjusted *P* values for the most significantly different OTUs between non-smokers and smokers.

6.2.3 Assessing community diversity

Samples were rarefied to the minimum number of reads (5118) to test for alpha diversity. Alpha diversity at OTU level was investigated to determine the possible associations between non-smokers, smokers and health status in oropharyngeal community structure. The diversity indices calculated for healthy and unhealthy samples from non-smoking participants and smokers were species richness, Shannon H index and Simpson index. The aov() from Vegan (version 2.4.0) was then used to calculate pair-wise ANOVA *P* values (taking into account repeated sampling from participants) which were displayed on top of alpha diversity figures.

6.2.4 Assessing community stability

Stability plots were produced as described in *Chapter 4.2.3* using betadisper() in Vegan (version 2.4.0) where the distance to the centroid (mean) from each sample was calculated for each participant showing temporal fluctuations for non-smokers and smokers. To determine if community structure changed before, during and after a disturbance for non-smokers and smokers, a linear mixed model (LMM) was constructed. The community composition of healthy samples from smokers was compared to healthy samples from non-smokers (as well as the unhealthy samples from non-smokers and smokers) in different infection states. The LMM used distances from *betadisper* to determine if there were differences in community structure between non-smokers and smokers.

6.3 Results

6.3.1 Comparison of healthy communities from non-smoking participants and smokers

6.3.1.1 Initial analysis of smokers' samples

At the end of the sampling period, 177 smoker's samples were received in total, with 145 designated as healthy. The number of samples (healthy and unhealthy) received from each smoking participant is shown in Table 6.1.

Table 6.1 – Metadata shown for smokers samples including the total number of samples (healthy and unhealthy) and the range in taxa numbers (with the median shown in brackets) received from each smoking participant.

Participant	**Age**	**Sex**	**Total samples**	**Number of years smoking**	**Average number of cigarettes smoked per week**	**Phylum**	**Genus**	**OTU**
SA	40	M	23	20	30	5-9 (9)	11-59 (46)	97-347 (221)
SB	19	F	3	5	120	8-8 (8)	42-46 (42)	183-199 (184)
SC	19	F	23	2	6	6-9 (8)	32-53 (44)	136-254 (208)
SD	19	F	8	1	60	5-8 (6)	28-46 (37)	127-225 (156)
SE	19	F	1	5	35	N/A	N/A	N/A
SF	19	F	19	4	30	6-9 (7)	30-51 (41)	120-260 (179)
SG	33	M	25	15	70	7-9 (8)	29-58 (53)	158-295 (201)
SH	30	F	25	13	12	7-9 (9)	31-58 (44)	101-313 (223)
SI	30	F	24	10	25	6-9 (8)	29-59 (46)	126-279 (222)
SK	19	F	10	5	60	8-9 (9)	40-52 (47)	172-279 (235)
SL	19	F	10	3	10	8-9 (8)	41-54 (46)	209-286 (234)
SM	19	F	6	3	30	7-8 (8)	45-52 (50)	208-236 (228)

The taxonomic profiling of samples from the oropharynx of individual smoking participants identified with RDP classifier revealed 5 to 9 phyla, 11 to 59 genera and 97 to 347 assignments at OTU level which was similar to non-smokers (5 to 10 phyla, 20 to 70 genera and 140 to 340 OTUs). The species richness at OTU level for non-smokers and smokers is shown in Figure 6.1.

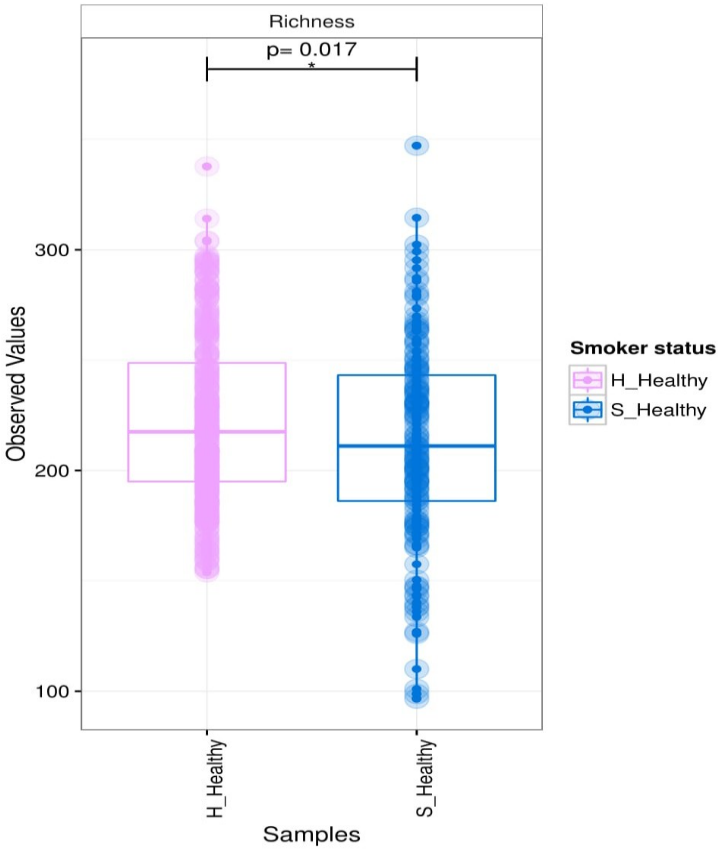


Figure 6.1 - Box plot showing species richness at operational taxonomic unit (OTU) level for healthy non-smokers (H_Healthy) and healthy smokers (S_Healthy) showing there is a significant difference in richness between the 2 groups as observed by the significant *P* value. The significant difference testing for clustering was corrected due to having repeated sampling from participants.

6.3.1.2 Community composition of healthy samples from non-smoking participants and smokers

The most abundant taxa from healthy samples from smoking participants included *Firmicutes* (mean proportion of the whole sample ± SEM = 50% ± 2%), *Bacteroidetes* (18% ± 1%), *Proteobacteria* (13% ± 2%), *Actinobacteria* (13% ± 1%) and *Fusobacteria* (3% ± 1%) (Appendix 10 & Appendix 11). However, comparison of abundances from healthy (smoker) and healthy (non-smoker) samples showed smokers to have significant increases in abundance of all phyla apart from *Fusobacteria* (Table 6.2) showing that community differences in abundance between smokers and non-smokers were present at the phylum level.

Table 6.2 – Linear mixed model (LMM) parameters investigating the abundances (response variable) of certain phyla in healthy samples from smokers compared to healthy samples from non-smokers (reference category). Significant *P* values are shown in bold.

Smokers samples	**Estimate**	**Std. Error**	**df**	**t value**	***P* value**
Firmicutes	0.3350	0.1115	424	3.0060	**0.002**
Proteobacteria	0.6834	0.2946	307	2.3200	**0.02**
Bacteroidetes	0.7542	0.1949	25.9	3.8700	**<0.001**
Actinobacteria	0.9916	0.1809	27	5.4830	**<0.001**
Fusobacteria	-0.0746	0.2845	22.3	-0.2620	0.796

At genus level, healthy samples from smoking participants had the most dominant genera of *Streptococcus* (39% ± 2%), *Prevotella* (12% ± 1%) and *Actinomyces* (5% ± 1%) (Appendix 12 & Appendix 13). When comparing abundances of healthy samples (smokers) to healthy samples (non-smokers), smokers had significant increases in abundance in the genera *Streptococcus* (*P* = 0.005) and *Prevotella* (*P* = 0.001) (Table 6.3).

Table 6.3 – Linear mixed model (LMM) parameters investigating the abundances (response variable) of certain genera in healthy samples from smokers compared to healthy samples from non-smokers (reference category). Significant *P* values are shown in bold.

Smokers samples	**Estimate**	**Std. Error**	**df**	**t value**	***P* value**
Streptococcus	0.3130	0.1129	424	2.7720	**0.005**
Prevotella	0.7884	0.2137	25.4	3.6890	**0.001**
Veillonella	0.3341	0.1957	25.1	1.7080	0.1

The most abundant OTUs belonged to *Streptococcus* species again reflecting the general abundance of their phylum, *Firmicutes* (Appendix 14). The most abundant OTUs identified included *Streptococcus mitis*, *Streptococcus salivarius* and *Streptococcus parasanguinis* which were also the most dominant OTUs in the healthy samples from non-smoking participants.

To determine the overall difference in community composition between non-smoking participants and smokers, an NMDS plot was produced from only the healthy samples from non-smokers and smokers, where samples from each group clustered separately. This was characterised by differences in abundance of OTUs using Bray-Curtis distance (Figure 6.2A) and presence and absence of different types of OTUs using unweighted UniFrac distance (Figure 6.2B). Significant differences were observed for both distances (Bray-Curtis, $P < 0.001$; unweighted UniFrac, $P < 0.001$) showing that non-smoking participants and smokers differ in the abundances of OTUs and in the presence and absence of specific OTUs.

These changes at OTU level were investigated to determine whether non-smokers and smokers have a different composition of OTUs in their oropharynx microbiomes. Smokers have increased abundances of opportunistic and pathogenic microorganisms as shown in Table 6.4 which shows the 10 most significant OTUs in terms of differing abundance found between the healthy samples (from non-smoking participants and smokers). Smokers had increased abundances of specific OTUs that were potentially pathogenic, and decreased abundances of certain OTUs including commensals such as *Neisseria oralis*.

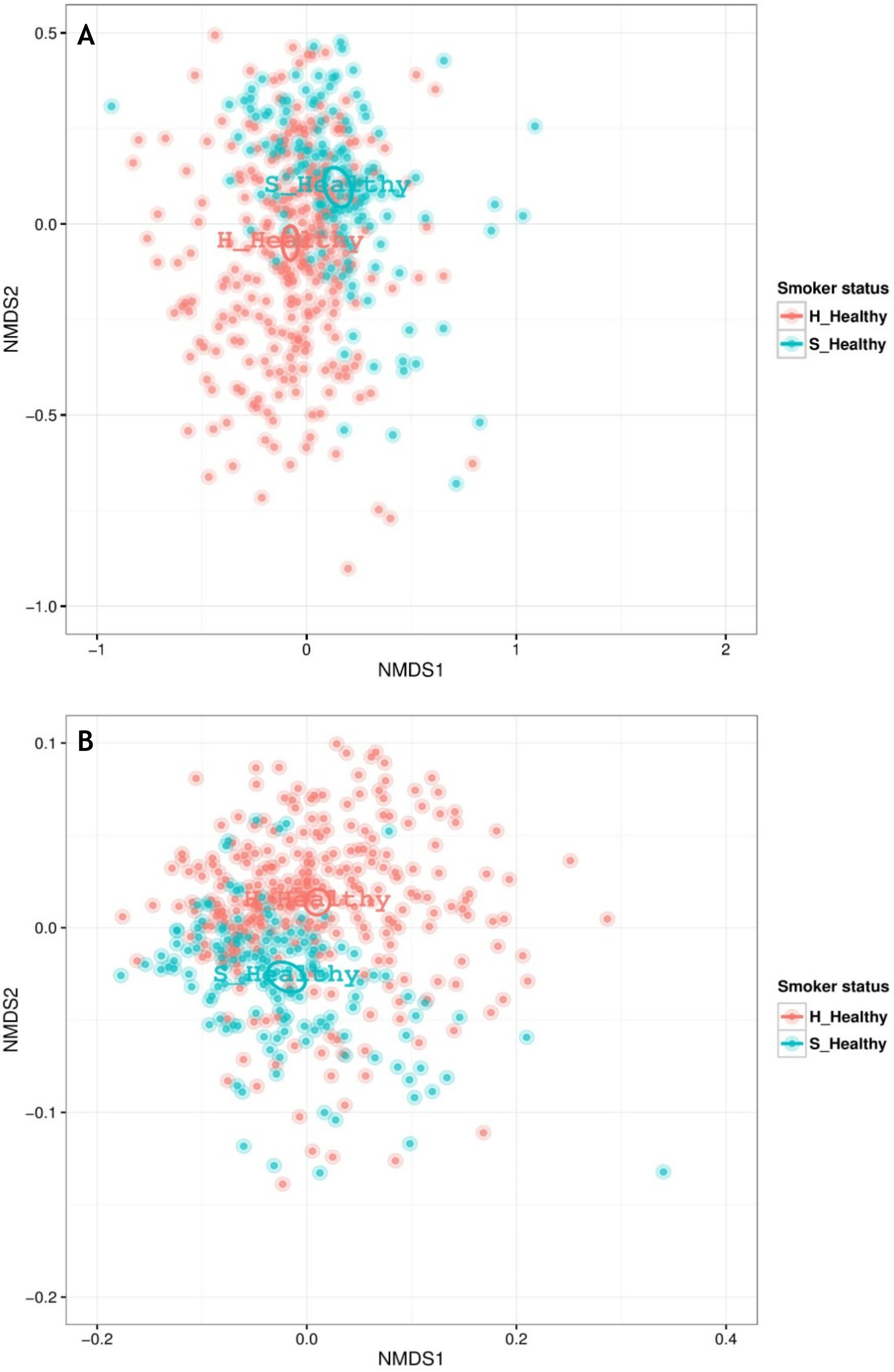


Figure 6.2 – Non-metric multidimensional scaling (NMDS) plots at operational taxonomic unit (OTU) level showing microbial community compositions of only the healthy samples in regards to smoker status. Variance ellipses were added by calculating the covariance for each group was calculated by cov.wt() and the shape of the ellipse was defined by the covariance within each group. Figure 6.2A uses Bray-Curtis distance whereas Figure 6.2B uses unweighted UniFrac phylogenetic distances.

Table 6.4 - The 10 most significant operational taxonomic units (OTUs) in terms of differing abundances found between the healthy samples from non-smoking participants and smokers. The abundance of the OTUs in smokers (whether increasing or decreasing) is shown through representation of arrows.

OTU	**Abundance in smokers**	**Adjusted *P* value**	**Description**
OTU_216 *Porphyromonas gingavalis* strain W83	↗	<0.001	Major pathogen in periodontitis (Nelson et al., 2003) - smoking increases risk of periodontal disease (Zeller et al., 2014)
OTU_60 *Streptococcus agalactiae*	↗	<0.001	Commensal and pathogen involved in sepsis and pneumonia (Tettelin et al., 2002)
OTU_66 *Streptococcus pyogenes* strain M1	↗	<0.001	Pathogen that causes tonsillitis (Bagaitkar et al., 2008), pharyngitis and scarlet fever (Ferretti et al., 2001)
OTU_82 *Enterococcus faecalis*	↗	<0.001	Some strains highly resistant to antibiotics known as vancomycin resistant *Enterococci* (Kristich & Rice, 2009) Linked to oral cancer through increased release of hydrogen peroxide (Boonanantanasarn et al., 2012) Also found in periodontitis (Wang et al., 2012)
OTU_192 *Bifidobacterium longum*	↗	<0.001	Commensal with some strains used as a probiotic in food and drinks (Sugahara et al., 2015)
OTU_17 *Neisseria oralis*	↘	<0.001	Healthy commensal found in the oral tract
OTU_8 *Chryseobacterium*	↗	<0.001	Common bacteria found in water and environmental sources
OTU_857 *Fusobacterium necrophorum subsp. funduliforme*	↗	<0.001	Present in the oropharynx in healthy individuals but has been involved in tonsillitis (Jensen et al., 2007)
OTU_13 *Corynebacterium propinguum*	↗	<0.001	Present in the oropharynx but has been involved in lower respiratory tract infections (Díez-Aguilar et al., 2013)
OTU_29 *Moraxella nonliquefaciens*	↘	<0.001	Usually a commensal, can become pathogenic (Marrs, 2016)

6.3.2 Comparison of unhealthy communities from non-smoking participants and smokers

6.3.2.1 Community similarity in healthy and unhealthy samples from non-smokers and smokers

For both non-smoking participants and smokers, the communities changed when there was a disturbance in health status as observed through NMDS plots at Bray-Curtis distance (Figure 6.3A) and unweighted UniFrac distance (Figure 6.3B). For non-smoking participants the community structure was altered when changing from a healthy to unhealthy state (Bray-Curtis, $P = 0.005$; unweighted UniFrac, $P = 0.02$). Colds were more associated with changes in abundances of OTUs whereas antibiotic treatment resulted in changes of the presence or absence of specific OTUs. For smokers, a change in health status also resulted in a change in community structure, but this was only significant for changes in abundance rather than presence or absence of OTUs when using a value of $P < 0.1$ for significance (Bray-Curtis, $P = 0.07$; unweighted UniFrac, $P = 0.6$). These changes were more apparent in the antibiotics group, as the cold samples from smokers showed similar community compositions to the healthy samples from smokers. Samples also clustered according to smoker and health status; healthy communities from non-smokers and smokers had tight clustering, whereas samples from the unhealthy groups (with the exception of cold samples from smokers) were more spread out from each other and had more variability in community structure compared to samples representing healthy states. However the antibiotic samples from both non-smoking participants and smokers seemed to have a similar community composition compared to the other samples showing that the effects of antibiotic treatment on non-smoking participants and smokers were similar.

For the smoker’s samples only, a change in community structure was not observed in regards to how many cigarettes smokers smoked per week on average. This was apparent in both NMDS plots using Bray-Curtis and unweighted UniFrac distances (Bray-Curtis, $P = 0.9$; unweighted UniFrac, $P = 0.8$) where communities did not cluster according to the number of cigarettes smoked per week.

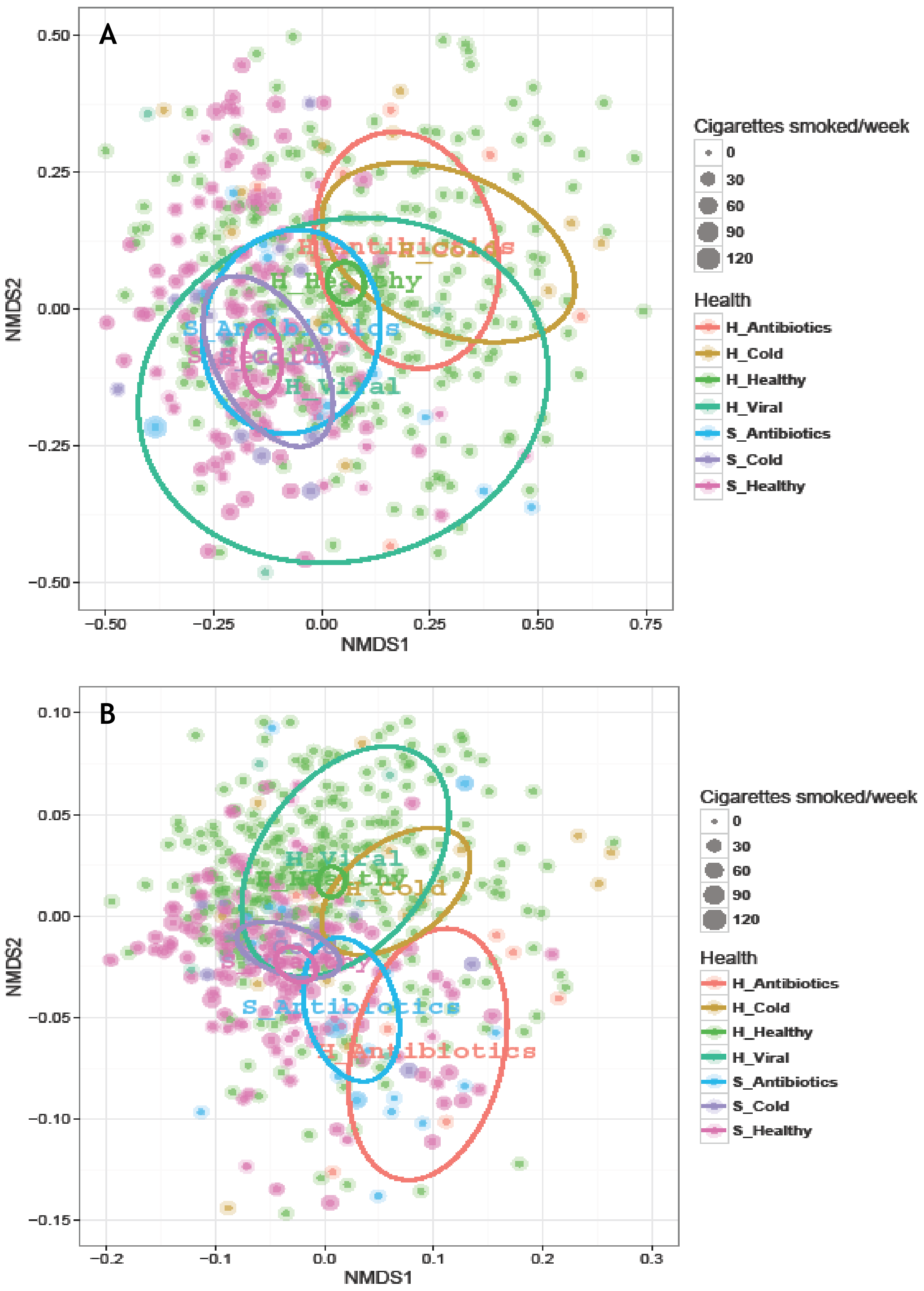


Figure 6.3 – Non-metric multidimensional scaling (NMDS) plot using Bray-Curtis dissimilarity index (Fig. 6.3A) showing differences in abundances of operational taxonomic units (OTUs) in regards to smoker and health status and the number of cigarettes smoked per week in relation to the size of circles – larger circles denote a greater number of cigarettes smoked weekly as opposed to smaller circles (smoker samples only). Figure 6.3B uses unweighted UniFrac phylogenetic distances showing presence and absence of OTUs in regards to smoker and health status and the number of cigarettes smoked per week (smoker samples only).

The specific differences in the unhealthy samples from non-smoking participants and smokers were investigated. The community composition in unhealthy samples from smokers showed a decrease in *Firmicutes* and *Proteobacteria* (compared to unhealthy samples from non-smokers) but these changes in abundance were not significant (Table 6.5).

Table 6.5 – Linear mixed model (LMM) parameters investigating the abundances (response variable) of certain phyla in unhealthy samples from smokers compared to unhealthy samples from non-smokers (reference category).

Unhealthy samples	**Estimate**	**Std. Error**	**df**	**t value**	***P* value**
Firmicutes	-0.1102	0.2882	66	-0.3820	0.7030
Proteobacteria	-0.5650	0.5634	17.71	-1.0030	0.3290
Bacteroidetes	0.7738	0.4048	66	1.9110	0.0600
Actinobacteria	0.4699	0.5600	14.71	0.8390	0.4149
Fusobacteria	-0.5936	0.4817	66	-1.2320	0.2220

Unhealthy samples from smokers had *Streptococcus* (38% ± 4%), *Prevotella* (15% ± 1%) and *Serratia* (8% ± 4%) as the most abundant genera (Appendix 15). However smokers showed no significant differences in the abundances in *Streptococcus* but a significant increase in *Prevotella* (*P* = 0.013) was observed in comparison to the unhealthy samples from non-smoking participants (Table 6.6).

Table 6.6 – Linear mixed model (LMM) parameters investigating the abundances (response variable) of certain genera in unhealthy samples from smokers compared to unhealthy samples from non-smokers (reference category). Significant *P* values are shown in bold.

Unhealthy samples	**Estimate**	**Std. Error**	**df**	**t value**	***P* value**
Streptococcus	-0.04024	0.31260	4.52	-0.129	0.903
Prevotella	1.18553	0.46705	66	2.538	**0.013**
Veillonella	-0.17916	0.52267	12.29	-0.343	0.737

6.3.2.2 Comparison of diversity in healthy and unhealthy communities from non-smoking participants and smokers

The diversity of samples in regards to smoker and health status was investigated using alpha diversity measures to observe changes in alpha diversity in regards to smoker and health status (Figure 6.4). In non-smoking participants at OTU level, significant differences were observed using pair wise ANOVA for richness and diversity measures. This was observed between the healthy and cold samples (species richness: $P < 0.001$, Shannon H index: $P = 0.01$, Simpson index: $P <0.001$) and healthy and antibiotics samples (species richness: $P < 0.001$, Shannon H index: $P = 0.001$, Simpson index: $P = 0.007$) showing that healthy samples (from non-smokers) had the greatest richness and diversity and unhealthy samples had the least richness and diversity. Looking at the smoker's samples alone at OTU level, the healthy and cold samples from smokers had no significant differences in species richness, Shannon H diversity or Simpson index showing that these groups were similar in terms of richness and diversity. Significant differences were observed in species richness between the antibiotics and cold samples at OTU level ($P = 0.003$) with samples obtained during antibiotic treatment having reduced richness compared to the cold samples but there were no significant differences in Shannon H or Simpson diversity. A significant difference was also observed in Shannon H diversity between the antibiotics and healthy samples from smokers at OTU level ($P = 0.003$) with antibiotics having reduced diversity, but not species richness or Simpson diversity. Overall this shows at OTU level the healthy samples from non-smoking healthy participants were more diverse than the healthy samples from smokers (species richness: $P = 0.01$, Shannon H index: $P <0.001$, Simpson index: $P < 0.001$), but the cold samples from smokers was shown to be the most rich and diverse. Diversity was impacted by health status where antibiotics use resulted in lower diversity in both non-smoking participants and smokers, whereas cold samples were associated with reduced richness and diversity in non-smokers but increased richness and diversity in smokers.

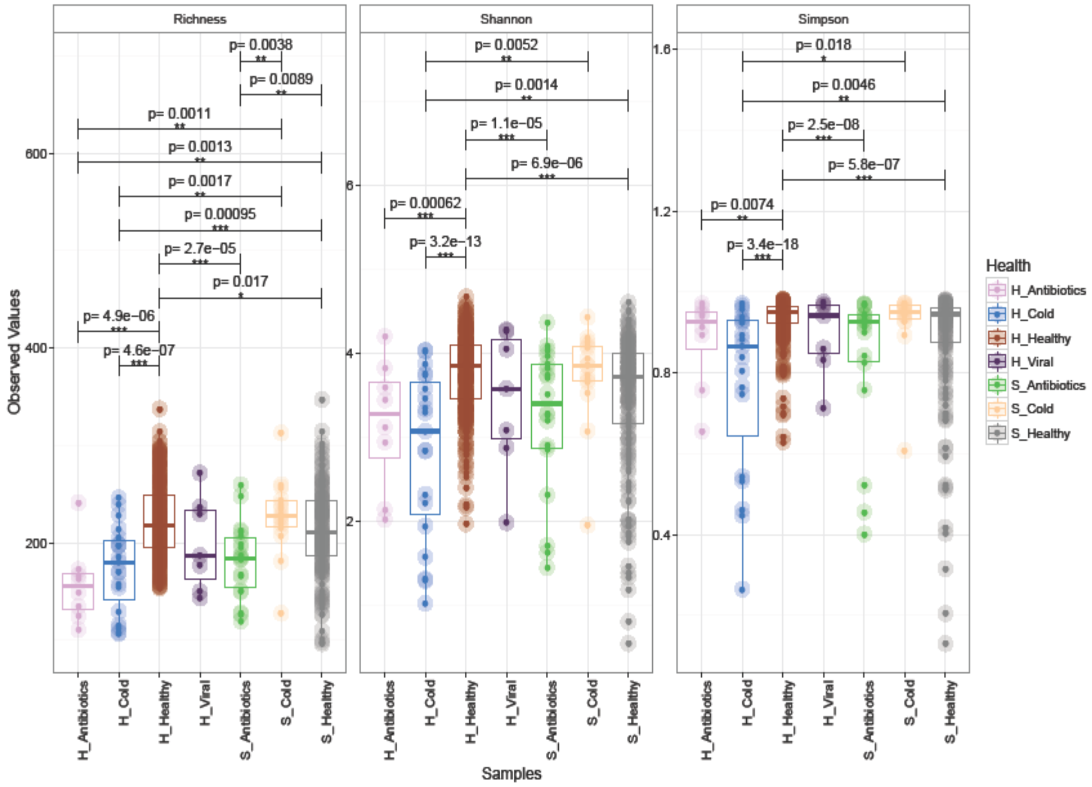


Figure 6.4 – Alpha diversity measures at operational taxonomic unit (OTU) level. Samples are categorised by health and smoker status. Significance testing was performed using pair wise ANOVA with significant *P* values shown above.

6.3.3 Comparing the stability of the oropharynx microbiome in non-smoking participants and smokers

6.3.3.1 Stability of the oropharynx microbiome in non-smoking participants and smokers

A combined stability plot for non-smoking participants and smokers showed the fluctuations of the oropharynx microbiome on a weekly basis and the changes in community structure during a disturbance (Figure 6.5). As determined before, non-smoking participants had disturbances that could be related to changes in health status, from which they would recover quickly, showing that the microbiome was resilient *(Chapter 5)*. Smokers also had changes in health status which were determined by greater peaks – more so for samples collected during antibiotic treatment than for samples collected during colds. However, visually there seemed to be greater peaks present in smokers which were not related to reported changes in health status or routine. The coefficient of variation gave a numerical stability representation for each participant. The coefficient of variations determined for the non-smoking participants ranged from 12% to 26%

and 9% to 29% for smokers. The variability in community structure was apparent within and between participants from both groups (although visually there seemed to be greater variability between smoking participants). Instead, it was concluded that each participant had a stable microbiome regardless of smoker status, as even though smokers had an altered microbial community when compared to non-smoking healthy participants, the microbiome was still relatively stable over the weeks of sampling and the coefficient of variation values for participants from both groups were generally low.

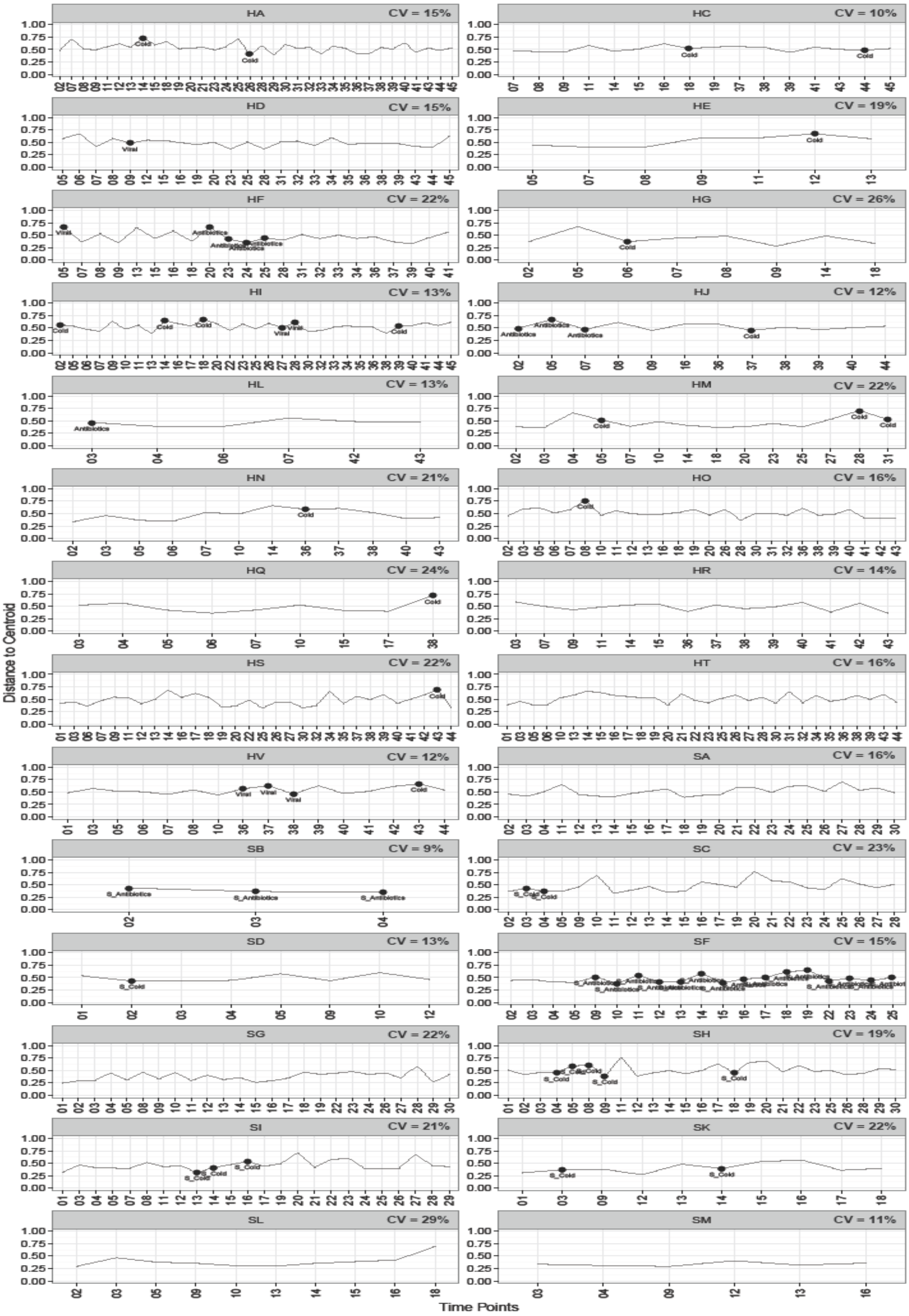


Figure 6.5 - Stability plots for all participants using distances from *betadisper* – non-smoking healthy participants are identified as starting with H, whereas smokers start with S. Time points indicate the week of when a sample was handed in and peaks refer to altered communities as identified by having a greater distance away from the centroid. Participant's HB and SE are omitted due to not having enough samples.

6.3.3.2 Is smoking associated with changes in community resilience after a disturbance?

An LMM was constructed using the distances from *betadisper* showing that there were differences in community structure between non-smoking participants and smokers (Table 6.7). There were significant differences between the healthy communities from smokers and non-smoking participants (*P* = 0.05*)*, one week before symptoms (*P* < 0.001) and during symptoms (*P* < 0.001) when compared to non-smoking participants. There were no significant differences when comparing the healthy smoker samples to any other smoker's samples during the different health states. Therefore, for the smoker's samples the time required for recovery from a cold could not be investigated.

Table 6.7 – Linear mixed model (LMM) parameters comparing the healthy samples from smokers (reference category) to different health states in smokers and non-smoking participants. The response variable is the distances from *betadisper* which represents differences in community structures. Significant *P* values are shown in bold.

Status of infection	**Estimate**	**Std. Error**	**df**	**t value**	***P* value**
Smokers: One week before symptoms	0.015	0.026	447.3	0.566	0.571
Smokers: During symptoms	-0.019	0.019	454.5	-1.047	0.296
Smokers: One week after symptoms	0.017	0.026	447	0.638	0.524
Healthy	0.025	0.013	28.8	1.977	**0.057**
Healthy: One week before symptoms	0.078	0.019	124.6	4.126	**<0.001**
Healthy: During symptoms	0.09	0.017	89.9	5.167	**<0.001**
Healthy: One week after symptoms	0.043	0.025	266.1	1.692	0.092

6.4 Discussion

The oropharynx microbiome of a smoker is distinct to the oropharynx microbiome of a non-smoking participant. Similar studies have also identified these differences with increased abundances of *Fusobacteria* and *Actinobacteria* at phylum level (Charlson et al., 2010). A higher percentage of *Firmicutes*, *Actinobacteria*, *Bacteroidetes* and *Proteobacteria* was observed in smokers, with increased abundances of potentially pathogenic microorganisms such as *Porphyromonas gingavilis*, *Streptococcus pyogenes* and *Fusobacterium necrophorum*, all of which have been implicated in oral and respiratory tract diseases such as periodontitis and pharyngitis (Camelo-Castillo et al., 2015) (Zeller et al., 2014). Smoking seems to distort healthy microbial communities through changing abundances of bacteria; this can be through either a reduction in commensal bacteria or overgrowth of opportunistic pathogens (Wu et al., 2016), which may affect the overall structure and functioning of the community. Increased abundances of potentially pathogenic OTUs were associated with certain oral and respiratory tract diseases like periodontitis (Nelson et al., 2003) and pharyngitis (Bagaitkar et al,. 2008).

The smoker’s community was distorted during a disturbance, but more from antibiotics treatment rather than a cold. The community structure changed in both non-smoking participants and smokers on antibiotic treatment showing that antibiotics use results in a disruption of the microbiome which was also seen in other studies (Jakobsson et al., 2010). For samples collected when cold symptoms were present, non-smoking participants displayed a change in community structure when shifting from a healthy to cold status, but this was not observed for smokers. From this it can be assumed that smokers have a permanent altered state (which may even be more stable than non-smokers) with higher abundances of pathogenic microorganisms, and so their healthy samples are similar in community composition to the cold samples.

The diversity of the oropharyngeal microbiome of non-smoking participants and smokers was also investigated and showed healthy samples from non-smokers to be significantly more diverse than the healthy samples from smokers in terms of species richness and Shannon H and Simpson diversity. The cold samples from the smokers were also similar in diversity to the healthy samples from the smokers; this again confirms the similarity in community composition between these two groups. The cold samples from smokers were shown to be the most diverse overall at OTU level when compared to all other groups from smokers and non-smoking participants and this could be due to increased transient and unknown bacteria – other studies have also reported smokers' communities to be more diverse than non-smokers (Charlson et al., 2010) with these studies suggesting that smokers may have greater resilience than non-smokers. However other studies have reported smokers to have less diverse communities when sampling in the oral tract (Camelo-Castillo et al., 2015) suggesting the role of diversity in smokers in the oropharynx and other body sites requires further investigation. Regardless, these results showed that smokers had differing diversity values (and OTU abundances) compared to non-smoking participants.

Participants from both groups were found to have different microbial community compositions that were stable as communities' had similar compositions on a weekly basis regardless of smoker status as confirmed through the stability plots. The coefficient of variation values in the non-smoking participants ranged from 12% to 26% and 9% to 29% in smokers, showing that the values were relatively similar; therefore it could not be stated that smokers have a more unstable microbiome as previously hypothesised. Non-smoking participants had high resilience in that they recovered quickly from a disturbance (either antibiotic treatment) however this could not be investigated for the smokers as the participants that were undergoing antibiotic treatment were still continuing treatment at the end of the sampling period so it could not be determined how quickly communities recovered again. The community structure of non-smoking participants changed one week before a disturbance but did not for smokers; this suggests smokers have an altered but relatively stable state that is not changed during a disturbance. However, even if distinct changes did occur during infection, it would not be possible to determine whether the pathogen was present before the participants recorded symptoms, or if the microbiome was disturbed prior to the infection which facilitated the disease. As the majority of smoking participants had been smoking for a number of years (the average of years smoking prior to this study was 7) it could well be that initial smoking may make the microbiome less stable at first, but could eventually settle to an altered stable state. The effects of initial smoking on the oropharynx microbiome were not investigated in this study.

This sampling showed the distinct changes in the bacterial community structure of the oropharynx in non-smoking participants and smokers. Sampling for smokers was limited to a total period of 6 months (as compared to a total sampling period of 9 months for non-smoking participants). A smaller sample size for the smokers made it more difficult to perform longitudinal sampling and determine stability patterns as there were fewer samples to investigate for each participant. Interestingly though, a similar number of unhealthy samples was received from smokers (n=32) and non-smokers (n=34) in a shorter time frame suggesting that smokers are more vulnerable or susceptible to illness. However, no viral infections were confirmed for the smokers, even though it may be that that cold samples were positive for viruses that were not picked up in the respiratory screen. This could be due to weekly sampling missing the onset of the viral infection, but also due to participants, as the number of viral swabs received from smokers was low compared to the number of viral swabs received from non-smoking participants. This shows that the effect of viral infection on the smoker’s oropharynx microbiome still requires investigation.

6.5 Conclusions

These findings identify the characteristic patterns of microbial communities in smoking and non-smoking participants. Specific OTUs were found to have increased abundances in the smokers group and these were *Porphyromonas gingavilis*, *Streptococcus pyogenes* and *Streptococcus agalactiae* which are all pathogens involved in oral and respiratory tract infections. Smokers and non-smoking participants also had different responses to deviations of the healthy state; non-smokers had high resilience with quick recovery, whereas smokers did not display changes to their microbiome during periods of a disturbance, suggesting a permanent altered state. Even though variability in community structure occurred within and between all participants, each participant’s microbiome was still stable regardless of smoker status. These results indicate that smokers do have stable but altered microbiomes compared to non-smoking participants.

7 Determining the function of the oropharynx microbiome

7.1 Introduction

As amplicon-based markers are widely used for microbiome studies, the prediction of the functional capabilities of these communities from 16S rRNA data sets would be extremely useful. For instance, by investigating the functions associated with microbial community structures it is possible to establish whether presence of certain taxa affects microbiome function, as well as explore functions associated with microbial communities of different health conditions (Shreiner et al., 2016). This not only provides information about the structure and general function of the community, but also investigates to what extent microbial variation between people might be associated with variation in its function. If this is observed, then this opens up the possibility for developing therapeutic tools where microorganisms can be used to restore or alter communities and thereby affect their functions in disease scenarios. Therefore investigating and understanding the functions associated with microbial communities are now becoming a very important area of research for microbiome studies. Improved methods for this have recently become available using the Tax4Fun package in R, (Aßhauer et al., 2015) which enables prediction of the functional community profile of 16S rRNA gene data by linking 16S rRNA gene sequences to already identified functions of sequences from genomes based on a minimum 16S rRNA sequence similarity.

Several authors have explored the functions of the bacterial communities of the oropharynx by comparing it to some diseased state. Castro-nallar et al., (2015) explored the microbial and functional diversity between a control and schizophrenic group. There were significant differences between the abundance of specific taxa, with an increase in the abundance of lactic acid producing bacteria in the oropharynx of schizophrenics. The taxonomic changes in communities resulted in altered expression of pathways in the control and schizophrenic group. Control groups had significantly increased expression levels of pathways involved in energy metabolism such as ATP synthesis which was lowered in schizophrenic patients. The schizophrenic group had significantly increased expression levels of pathways involved in environmental information processing such as glutamate transport, which had reduced expression levels in the control group. Schizophrenia has been linked to disturbances in the neurotransmitters glutamate and dopamine (Moghaddam & Javitt, 2012), suggesting that bacterial communities may also influence or exacerbate symptoms of this condition by having increased expression levels of these pathways in their communities.

However, the functional roles of the oropharynx microbiome are still not well defined highlighting the need for more studies focusing on both taxonomic composition and functional diversity of the oropharynx in healthy participants. This information can then be used in disease scenarios, comparing how functional profiles of the oropharynx differ during upper respiratory infections, disturbances such as antibiotic treatment as well as lifestyle factors like smoking. Smoking has been shown to result in altered community structures in the oropharynx when compared to non-smoking participants (*Chapter 6*), yet there is still very little information on whether smoking affects the functional roles of the oropharynx microbiome.

This chapter will explore the predicted functions associated with the 16S rRNA gene from the oropharynx of healthy and unhealthy samples from non-smokers, as well as a comparison of just the healthy samples from non-smokers and smokers. The objectives of this chapter are to address the following questions: what predicted oropharyngeal functions are associated in non-smoking participants, and do smokers have changed oropharyngeal functions compared to non-smokers.

7.2 Methods

All samples were processed as described in the methods chapter *(Chapter 2)*. To predict functions associated with oropharynx samples, functional profiles of 16S rRNA gene sequences were identified using the Tax4Fun package (version 0.3.1) (Aßhauer et al., 2015) in R (version 3.1.2) which links 16S rRNA gene sequences with the functional annotation of sequenced prokaryotic genomes using a nearest neighbour identification based on a minimum sequence similarity. This involves (i) annotating the representative sequences of the OTUs against the SILVA database, (ii) transforming the annotated 16S rRNA profile to its equivalent KEGG (Kyoto Encyclopedia of Genes and Genomes) taxonomic profile using a precomputed association matrix, (iii) normalising the abundance of KEGG organisms by 16S rRNA copy number, and (iv) combining the normalised KEGG abundance profile with precomputed functional profiles of KEGG organisms (obtained with UProC; Meinicke, 2015) to predict the functional profile of the microbial community under study. This generates a relative abundance of KEGG orthology (KO) identifiers associated with each sample depending on matches of the representative sequence from each OTU to KEGG organisms. All prokaryotic KEGG organisms are available in Tax4Fun for SILVA SSU Ref NR database release 115 and KEGG database release 64.0. The 200 most abundant predicted KO identifiers were selected for the comparisons of the conditions of interest. Statistically significant differences in relative abundances of predicted functions between conditions were estimated using Kruskal-Wallis tests with Benjamini- Hochberg correction for false discovery rate, with a significance level of $P < 0.05$. For this study, 16S rRNA gene sequences were used to predict functional community profiles between healthy and unhealthy samples from firstly non-smoking participants and then between the healthy samples from non-smoking participants and smokers. Figures show the top 20 most significant results as displayed on a log transformed scale for visualisation.

7.3 Results

7.3.1 Predicted functions associated with the healthy oropharynx microbiome

From previous chapters it has been determined that different community compositions exist in the different health groups of non-smoking participants *(Chapters 3 & 4)*. These changes in community composition could potentially affect the functions associated with the communities. The predicted functional properties of oropharyngeal communities from the different health groups (as shown as the abundances of KO pathways/enzymes) were explored. The significant differences in abundances of KO's between the health groups showed which predicted functions were associated when going from a healthy to unhealthy state (as described by cold and viral samples) and a disturbed state (antibiotics samples). When observing the 20 most significantly different predicted functions between the healthy and cold samples, higher abundances in

KO's were seen in the enzymes/pathways associated with the healthy samples (Figure 7.1), showing that these enzymes are present and required in healthy and unhealthy samples. Specific pathways that were significantly different between the healthy and cold samples include the pathway K06610 MFS transporter which is involved in sugar transport across membranes (Pao et al., 1998) and K06969 23S rRNA (cytosine1962-C5)-methyltransferase which is an enzyme involved in ribosome production (Purta et al., 2008) (Table 7.1). The lower abundance of KO's associated with enzymes/pathways in the cold samples may be an effect of this kind of disturbance/infection with cold samples having lower abundances of OTUs as opposed to baseline abundances of OTUs in the healthy samples.

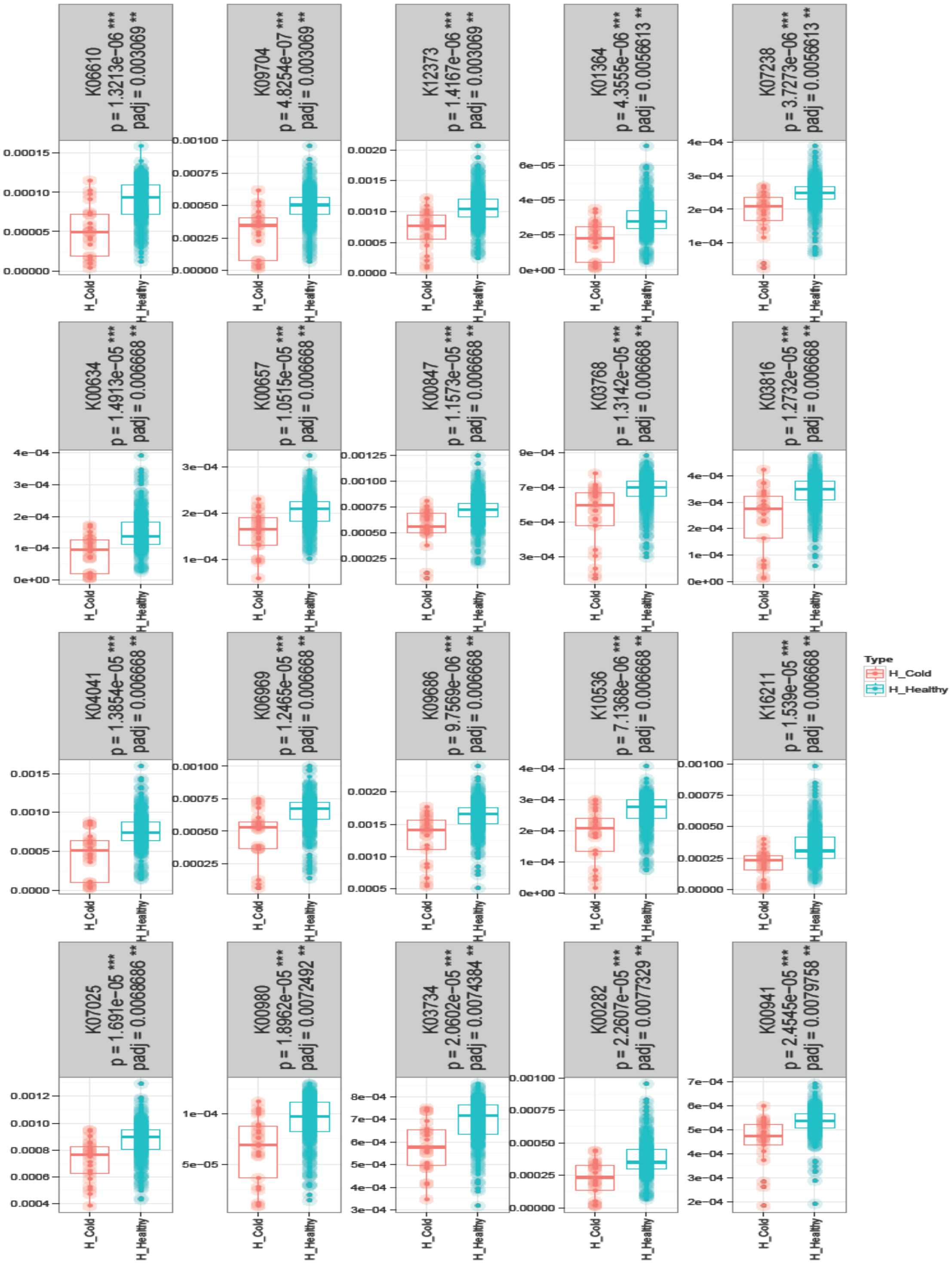


Figure 7.1 - Predicted functional profiles showing the top 20 most significantly different relative abundances of KO's (kyoto encyclopedia of genes and genomes orthology) shown by *P* values between healthy (blue) and cold (red) samples from non-smokers. Box plots are plotted showing the median (and the 25th and 75th percentiles) abundance of KO for each health group.

Table 7.1 – The kyoto encyclopedia of genes and genomes orthology (KO) numbers and description of the enzymes/pathways identified in healthy and cold samples as shown in Figure 7.1.

KO numbers	Pathways
K06610	MFS transporter, SP family, inositol transporter
K09704	uncharacterized protein
K12373	hexosaminidase
K01364	streptopain
K07238	zinc transporter, ZIP family
K00634	phosphate butyryltransferase
K00657	diamine N-acetyltransferase
K00847	fructokinase
K03768	peptidyl-prolyl cis-trans isomerase B (cyclophilin B)
K03816	xanthine phosphoribosyltransferase
K04041	fructose-1,6-bisphosphatase III
K06969	23S rRNA (cytosine1962-C5)-methyltransferase
K09686	antibiotic transport system permease protein
K10536	agmatine deiminase
K16211	maltose/moltooligosaccharide transporter
K07025	putative hydrolase of the HAD superfamily
K00980	glycerol-3-phosphate cytidylyltransferase
K03734	FAD:protein FMN transferase
K00282	glycine dehydrogenase subunit 1
K00941	hydroxymethylpyrimidine/phosphomethylpyrimidine kinase1

There were also significant differences in abundance of KO enzymes/pathways between the healthy and viral samples (Figure 7.2). There were various KO's that had higher abundance in the viral samples than the healthy samples such as the enzyme K00007 D-arabinitol 4-dehydrogenase (Table 7.2) which is involved in fructose metabolism. From the enzymes identified in Figure 7.2 & Table 7.2, only K00002 alcohol dehydrogenase (involved in degradation of aromatic compounds) and K00075 UDP-N-acetylmuramate dehydrogenase (carbohydrate metabolism) had increased abundance in the healthy samples. However, viral samples also had significant increases in abundance in other KO enzymes not shown in Figure 7.2 or Table 7.2. This included the magnesium and cobalt transport protein CorA which has been implicated in virulence during infection (Kersey et al., 2012) and therefore had a reduced abundance in healthy samples. In comparison to the microbiome from the cold samples, the microbiome from viral samples had an increase in abundance in KO enzymes and pathways required for everyday processes but also in enzymes/pathways involved in virulence.

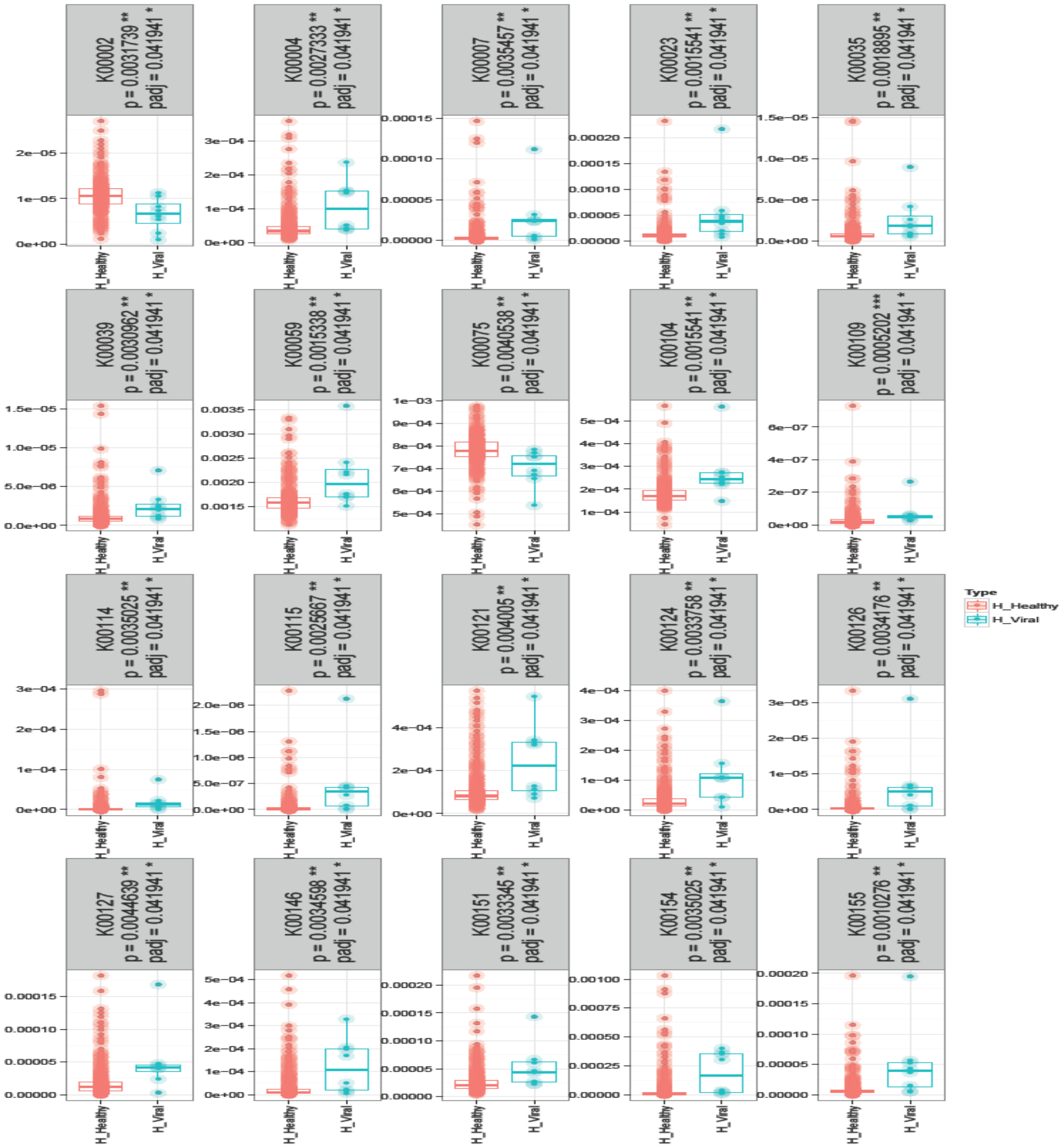


Figure 7.2 - Predicted functional profiles showing the top 20 most significantly different relative abundances of KO’s (kyoto encyclopedia of genes and genomes orthology) shown by *P* values between healthy (red) and viral (blue) samples from non-smokers. Box plots are plotted showing the median (and the 25th and 75th percentiles) abundance of KO for each health group.

Table 7.2 - The kyoto encyclopedia of genes and genomes orthology (KO) numbers of the enzymes/pathways identified in healthy and viral samples as shown in Figure 7.2. Asterisks represent KO enzymes/pathways that had the highest abundance in viral samples.

KO Numbers	**Pathways**
K00002	alcohol dehydrogenase
K00004*	butanediol dehydrogenase / meso-butanediol dehydrogenase / diacetyl reductase
K00007*	D-arabinitol 4-dehydrogenase
K00023*	acetoacetyl-CoA reductase
K00035*	D-galactose 1-dehydrogenase
K00039*	ribitol 2-dehydrogenase
K00059*	3-oxoacyl-[acyl-carrier protein] reductase
K00075	UDP-N-acetylmuramate dehydrogenase
K00104*	glycolate oxidase
K00109*	2-hydroxyglutarate dehydrogenase
K00114*	alcohol dehydrogenase (cytochrome c)
K00115*	glucose dehydrogenase (acceptor)
K00121*	S-(hydroxymethyl)glutathione dehydrogenase / alcohol dehydrogenase
K00124*	formate dehydrogenase iron-sulfur subunit
K00126*	formate dehydrogenase subunit delta
K00127*	formate dehydrogenase subunit gamma
K00146*	phenylacetaldehyde dehydrogenase
K00151*	5-carboxymethyl-2-hydroxymuconic-semialdehyde dehydrogenase
K00154*	coniferyl-aldehyde dehydrogenase
K00155*	NAD-dependent aldehyde dehydrogenases

Antibiotic treatment on the other hand was associated with a reduced abundance in KO enzymes/pathways that had higher abundances in the healthy samples (Figure 7.3). The enzymes/pathways identified in both groups of samples (Table 7.3) are involved in everyday processes such as membrane transport (K11077 mannopine transport system permease protein). But the reduction in all pathways in the antibiotic treated samples, an example being the production of flagella for bacteria (K02383 flagellar protein FlbB) is perhaps due to the antibiotic treatment eliminating groups of bacteria. This showed that in comparison to the predicted baseline abundance of these KO enzymes/pathways in the healthy samples, antibiotic treatment resulted in reduced abundance of pathways, most likely a result of antibiotic treatment killing bacterial populations.

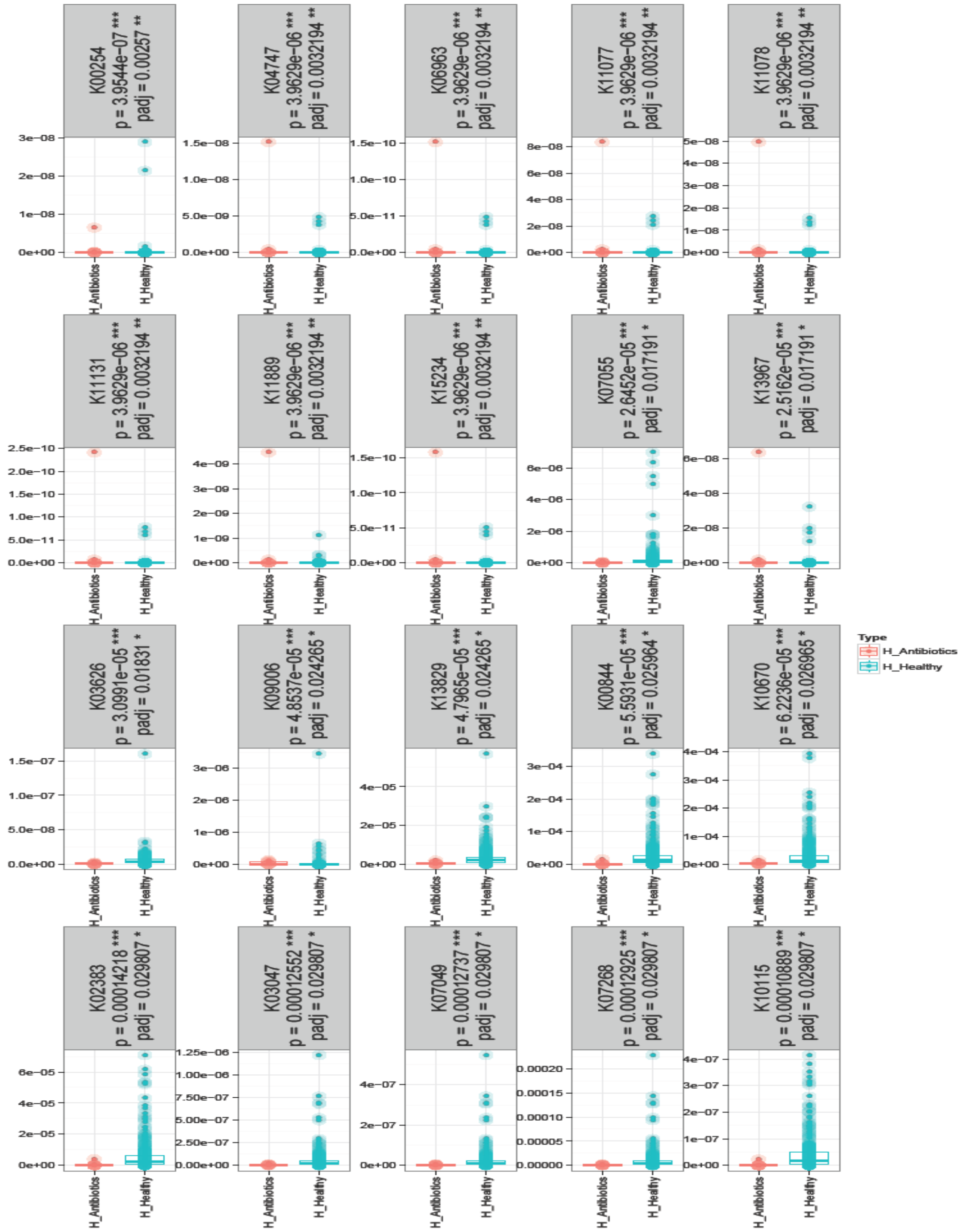


Figure 7.3 - Predicted functional profiles showing the top 20 most significantly different relative abundances of KO's (kyoto encyclopedia of genes and genomes orthology) shown by *P* values between healthy (blue) and antibiotic treated samples (red) from non-smokers. Box plots are plotted showing the median (and the 25th and 75th percentiles) abundance of KO for each health group.

Table 7.3 - The kyoto encyclopedia of genes and genomes orthology (KO) numbers of the enzymes/pathways identified in healthy and antibiotic treated samples as shown in Figure 7.3.

KO numbers	Pathways
K00254	dihydroorotate dehydrogenase
K04747	nitric oxide reductase NorF protein
K06963	tRNA acetyltransferase TAN1
K11077	mannopine transport system substrate-binding protein
K11078	mannopine transport system permease protein
K11131	H/ACA ribonucleoprotein complex subunit 4
K11889	type VI secretion system protein ImpN
K15234	citryl-CoA lyase
K07055	tRNA wybutosine-synthesizing protein 2
K13967	N-acetylmannosamine-6-phosphate 2-epimerase / N-acetylmannosamine kinase
K03626	nascent polypeptide-associated complex subunit alpha
K09006	uncharacterized protein
K13829	shikimate kinase / 3-dehydroquinate synthase
K00844	hexokinase
K10670	glycine reductase
K02383	flagellar protein FlbB
K03047	DNA-directed RNA polymerase subunit D
K07049	TatD-related deoxyribonuclease
K07268	opacity associated protein
K10115	maltooligosaccharide transport system permease protein

Overall these results show that various KO enzymes and pathways predicted from the 16S rRNA gene are shared between healthy, unhealthy and antibiotic treated samples from non-smoking participants. Disturbances affected the functions of the oropharynx microbiome in different ways, by having reduced abundances of KO enzymes/pathways in cold samples, increased abundances in viral samples (as well as having increased functions involved in virulence) and reduced abundances in antibiotic treated samples.

7.3.2 Do smokers have changed functions in the oropharynx?

Analysis of the top 20 most significantly different KO's between the healthy samples from non-smoking participants and smokers are shown in Figure 7.4. This showed that non-smoking participants and smokers had the same pathways identified such as roles in amino acid or carbohydrate metabolism, but smokers had increased abundances of KO's. However, smokers were found to have reduced abundances of certain functions (Table 7.4) compared to those of non-smoking participants. This included pathways for production of single-stranded DNA specific exonucleases (K07462) and the DNA mismatch repair protein MutL (K03572) which is not shown in the 20 most significant results. Both of these enzymes play a role in correcting errors after DNA replication and are vital for DNA repair (Skaar et al., 2002). Therefore, smokers had reduced abundances in KO pathways that were involved in DNA repair.

Although smoking resulted in increased abundances in pathogenic bacteria *(Chapter 6),* it also affected the functions involved in bacterial pathogenesis and survival by having reduced abundances of KO enzymes/pathways involved in virulence and resistance. An example being the K06158 ATP binding cassette which is involved in virulence such as transportation of toxic molecules (Davidson et al., 2008) where there was reduced abundances in smokers. This was also seen in the copper resistance phosphate response regulator CusR which provides natural resistance to copper which is toxic to bacteria (not shown in the 20 most significant results), with reduced abundance levels being observed in smokers (Munson et al., 2000). Therefore these results suggest smoking can actually influence the replication and survival of bacteria during colonisation by reducing abundances of some virulence factors which may affect the outcome of infection and pathogenesis.

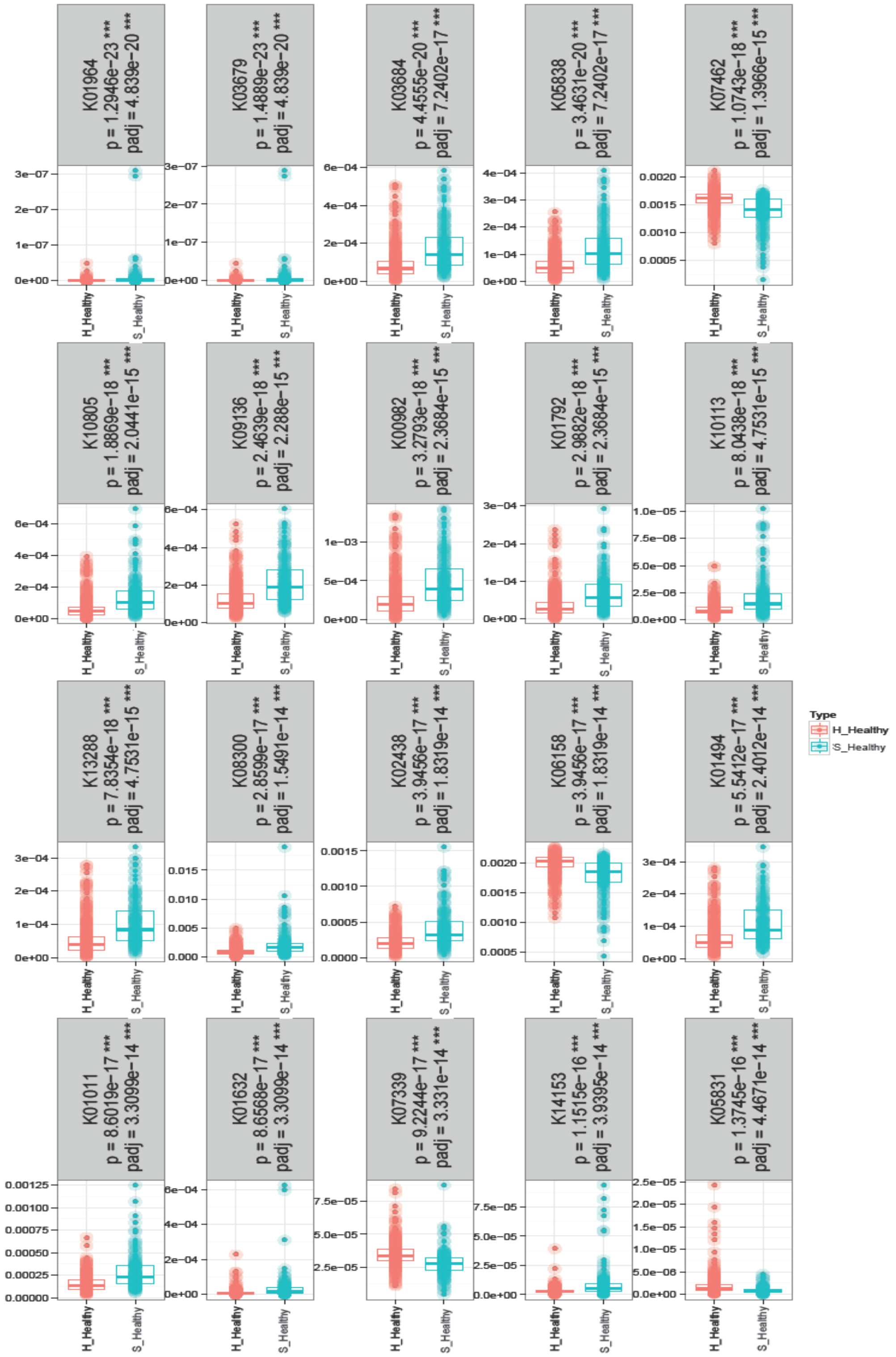


Figure 7.4 - Predicted functional profiles showing the top 20 most significantly different relative abundances of KO's (kyoto encyclopedia of genes and genomes orthology) shown by *P* values between healthy samples from non-smokers (red) and smokers (blue). Box plots are plotted showing the median (and the 25th and 75th percentiles) abundance of KO for both groups.

Table 7.4 - The kyoto encyclopedia of genes and genomes orthology (KO) numbers of the enzymes/pathways identified in only the healthy samples from non-smoking participants and smokers as shown in Figure 7.4. Asterisks represent KO enzymes/pathways that had increased abundances in the healthy samples from smokers in comparison to the healthy samples from non-smoking participants.

KO numbers	Pathways
K01964	acetyl-CoA/propionyl-CoA carboxylase
K03679	exosome complex component RRP4
K03684*	ribonuclease D
K05838*	putative thioredoxin
K07462	single-stranded-DNA-specific exonuclease
K10805*	acyl-CoA thioesterase II
K09136*	ribosomal protein S12 methylthiotransferase accessory factor
K00982*	glutamate-ammonia-ligase adenylyltransferase
K01792*	glucose-6-phosphate 1-epimerase
K10113*	maltooligosaccharide transport system substrate-binding protein
K13288*	oligoribonuclease
K08300*	ribonuclease E
K02438*	glycogen operon protein
K06158	ATP-binding cassette, subfamily F, member 3
K01494*	dCTP deaminase
K01011*	thiosulfate/3-mercaptopyruvate sulfurtransferase
K01632*	fructose-6-phosphate phosphoketolase
K07339	mRNA interferase HicA
K14153*	hydroxymethylpyrimidine kinase
K05831	LysW-gamma-L-lysine/LysW-L-ornithine carboxypeptidase

7.4 Discussion

This study has shown the oropharynx microbiome to have various predicted functions involved in metabolism, protein synthesis and DNA repair. Metagenomic analysis has shown a wealth of data reporting how functions from bacterial communities in the GI tract promote a healthy state in the host

(Gerritsen et al., 2011). This includes bacteria in the GI tract having metabolic activities that lead to the production of important nutrients such as short-chain fatty acids, vitamins and amino acids, which humans are unable to produce themselves (Wong et al., 2006). Therefore the bacterial communities in the oropharynx may also interact with the human host not just in defence but also in metabolic or immune response activities.

Healthy and unhealthy samples from non-smoking participants showed largely the same predicted functions, however, there were functions associated with virulence and pathogenesis in the unhealthy samples which may be a result of higher abundances of pathogenic bacteria. Smokers also had significant differences in KO's or predicted functions compared to non-smoking participants. As cigarette smoke contains carcinogens, toxins and oxidants, it is expected that this would disrupt microbial communities through inducing cellular damage and host changes such as inflammation which could change functions (Bagaitkar et al., 2008). This study showed reduced abundances in the predicted functions of DNA synthesis and repair mechanisms in bacteria from smokers (compared to non-smoking participants) showing that smoking does affect the oropharynx microbiome in terms of pathogenic microorganisms and the functions associated. Cigarette smoke has been linked to increased DNA mutations and DNA abnormalities in buccal cells (Tan et al., 2008) suggesting that the oropharyngeal microbial populations in smokers may also have reduced functions of DNA repair (which may arise due to increased DNA mutations) especially as enzymes such as MutL are extremely important in preventing mutations from becoming permanent in dividing cells (Li, 2008). However there were numerous predicted functions present in both groups that are necessary for everyday processes but in general there was an increased abundance in smokers.

The purpose of predicting functions of the oropharynx microbiome is to determine if differences in functional diversity of the microbiome can contribute to specific diseases. Associations have already been reported in schizophrenia where schizophrenic patients' pathways were significantly involved in environmental information processing whereas controls had higher proportions of pathways involved in energy metabolism (Castro-Nallar et al., 2015). However to truly understand the functional differences between health and disease, there must first be an investigation and understanding in whether the changed state of the microbiome is the cause or effect of the disease. This can only be determined through longitudinal sampling over a range of healthy and diseased subjects to explore community composition and transcriptomics to measure function. Once this has been established, functional differences may then be used for the development of biomarkers in health or disease.

The advantages of using the Tax4Fun package is that 16S rRNA sequencing is more cost-efficient than whole-genome shotgun sequencing, especially for initial exploration into the functions of the community. The disadvantage is that it only predicts the functional profile related to what is available in the reference database, therefore prediction is limited. Other disadvantages also incluce strain differences within OTUs and no actual measurement of expression. As this is based on predicting functional profiles, the validity of the functional profiling has not been investigated; therefore the coverage of taxonomic assignments needs to be investigated to check the reliability of the predictions. In order to gain a better representation of the functions associated with any microbiome, metagenomics and transcriptomics would need to be performed which would identify functions associated with genes from bacteria as well as the other microorganisms recovered from the oropharynx. It would also be interesting to look at participants who had recently started smoking (as all smokers participants had been smoking for at least over a year) and to determine when exactly the functions began to change. This would involve doing longitudinal sampling again for non-smoking participants and smokers and doing functional profiles for each sample on a weekly basis to determine how long it takes for changes to occur and if all participants' functional profiles change in the same way. Longitudinal sample would also give an estimate to how long recovery would take for a changed microbiome to return back to normal after stopping smoking. However, initial exploration of the predicted functions associated with the healthy oropharynx microbiome show these communities to be involved in various pathways. Disruption of these communities (either through illness or smoking) do show changes in these functions, suggesting respiratory infection or smoking results in changed functions.

7.5 Conclusions

Non-smoking participants had increased abundances of KO's (predicted functions) that were reduced in the unhealthy samples or in antibiotic treated samples. Smokers also had different functional profiles including reduced abundances of KO pathways involved in DNA repair, suggesting that bacterial communities in smokers were prone to DNA mutations. Therefore the microbiome of non-smoking communities has various predicted functions that are changed during disturbances including smoking and antibiotic treatment.

8 Conclusions and future work

The purpose of this PhD project was to develop the ecological knowledge and understanding of bacterial communities isolated from the oropharynx of non-smoking participants and smokers. Understanding these dynamics will provide opportunities to improve knowledge and information regarding the community structure of the healthy oropharynx microbiome and the changes that occur during respiratory infections and disturbances. More specifically, this thesis addressed the following questions:

What is the community composition of the healthy oropharynx microbiome in non-smokers?
How does the community change during a disturbance and antibiotic treatment?
How stable is the community and how does this change over a defined period of time?
How does the oropharynx microbiome of a non-smoker compare to a smoker's microbiome?
What functions are predicted to be associated with the oropharynx microbiome in non-smokers and smokers?

Overall, this PhD project has shown that the healthy oropharynx microbiome of a non-smoker consists of a diverse community of taxa that is similar at phylum and genus level between participants. The changes that occur during a disturbance are distinct and require further investigation in how the microbiome affects the host and health status.

8.1 Conclusions

8.1.1 Characterisation of the healthy and unhealthy oropharynx microbiome in non-smokers

Investigation into the microbiome of the oropharynx showed the healthy oropharynx microbiome to be *Firmicutes* and *Streptococcus* dominated at phylum and genus level, respectively. Participants had broadly similar bacterial community structures at phylum level with increasing differences at genus and OTU level. Healthy communities overall were the most diverse, with samples taken when participants had a cold or were undergoing antimicrobial treatment having less diversity, which was consistent in other studies where diseases seemed to have the least diversity in microbiome compositions in comparison to their healthy counterparts (Lozupone et al., 2012). The microbiomes from healthy samples consisted of various bacteria existing together which were investigated through co-occurrence networks. These showed interactions of bacteria in healthy states; some competing, whilst others were mutual. In microbiomes from healthy samples all genera were involved in positive interactions, in that as one genus increased in abundance then so did another genus. These interactions may be a factor to why taxa are assembled in these proportions to maintain a homeostatic status. A disorder in these proportions may be a result of disturbed interactions between genera, either through presence or increased abundance of other genera altering the community as a whole. This may be the case in unhealthy samples, but the abundances in communites may also fluctuate in healthy and unhealthy settings so overrepresented taxa in unhealthy situations may also be a result of bacteria adapting. But, investigation of these networks in healthy samples did allow identification of the important drivers or key genera present in the community. And so, rather than focusing on determining a core microbiome in the healthy oropharynx, the project focused on identifying the most abundant phyla and genera in healthy and unhealthy communities; for the healthy communities these were *Streptococcus*, *Prevotella* and *Veillonella* at genus level. Taxa were more variable at OTU level, but there was still identification of certain OTUs such as *S. mitis* that were always present in participants. This species has previously been described as a dominant community member of the healthy oropharynx (Mitchell, 2011). Variation in OTUs within and between participants could be a result of host and lifestyle factors (David et al., 2014), but lifestyle factors such as diet, alcohol consumption or activities performed were not recorded and so this study cannot determine how much of the external environment affects and influences characterisation and variation in the oropharynx microbiome.

Characterisation of the unhealthy samples showed different community structures – the microbiome of samples that tested positive for viruses had increases in bacteria such as *Haemophilus and Neisseria* which was also observed in other studies (Hofstra et al., 2015). The microbiome of virus-negative samples collected during colds had increases in *Haemophilus, Neisseria* and *Serratia* whereas the microbiome of samples collected during antibiotic treatment had increases in *Pseudomonas* and *Actinomyces*. Overall viral infection was not associated with major changes to the oropharynx microbiome, perhaps due to the fact that there was a small sample size of viruses (n=8) received. However, whether or not a pathogen can cause infection depends on the balance of microbiome homeostasis and abundances of pathogens. This homeostasis of commensal bacteria such *S. mitis* or *S. salivarius* (determined as the most abundant OTUs in healthy participants) may keep pathogenic species such *S. pneumoniae* or *H. influenzae* residing at low levels (Santagati et al., 2012) – *H. influenzae* was shown to be present in the majority of healthy samples but increased abundances were noticeable during colds and viral infections. Therefore opportunistic or pathogenic species may already be well established in the resident microbiome at levels that will not cause an infection or disease, but do increase in abundance either before the onset of infection or during disease or infection. The loss of protective species may be the cause of this and detrimental to the microbiome resulting in an altered community. Therefore the result of studies exploring healthy and diseased scenarios identifies the presence or absence of specific bacteria that are predictive of, or the cause of disease.

8.1.2 Stability and recovery of the healthy oropharynx microbiome

Investigating the individual stability patterns of the microbiomes of non-smoking participants showed participants to have stable oropharynx microbiomes in that each participant had a community structure that was relatively similar on a weekly basis and the fluctuations (as observed from the stability plots) that did occur between weeks were usually small. This was how stability was defined in this study; an unstable microbiome would be one that had different community structures weekly and greater fluctuations between samples from the same participants. Even though there was variation in community structure present in samples within and between participants, the oropharynx microbiome was still resilient to disturbances in that it returned towards their long-term average state within a week of recovering after disturbances; this has also been observed in other studies where oropharynx samples have been taken from the same participant after a noted period of time showing similar community structures in the two samples (Charlson et al., 2011). However there is a need to understand the immigration and emigration patterns of microbes in healthy and unhealthy states to understand the variation present in the healthy oropharynx microbiome. For example, is post-infection recovery driven by ecological forces such as migration rates and competitive exclusion or does the host play an active role in rebuilding a stable ecosystem? The microbiome is involved in resilience against pathogens, but is also involved in immune regulation and barrier defence. Therefore the host may impact the microbiome of various body sites and host-microbiome interactions are being explored most notably in the microbiome of the GI tract using mouse models (Kostic et al., 2013). However, how it impacts the assembly, stability and resilience of microbial communities in the oropharynx needs investigating. The oropharynx is constantly exposed to the external environment and is influenced by respiratory and gastrointestinal processes of which bacteria are adapted to. Therefore the crosstalk between the host and the oropharynx microbiome could be involved in influencing the outcome of a disease or infection.

Stable communities may contain keystone taxa (*Streptococcus, Prevotella and Veillonella* in the healthy samples), and if these keystone taxa are lost, this may result in abrupt changes to the community (Fierer et al., 2012). In unhealthy samples, there were increased abundances in *Haemophilus* or *Serratia* and a low diversity community. *Streptococcus* was still present but did not show any significant difference in abundance between the healthy and unhealthy samples, whereas a significant reduction in *Prevotella* was observed in the unhealthy samples. Therefore a combination of taxa may be responsible for promoting a healthy stable community. High biodiversity has shown to promote stable communities in various ecosystems ranging from plants to fish (McCann, 2000), but no relationship between bacterial diversity and stability was observed in this study. All microbiomes are subject to perturbations over the course of normal development, ageing and disease, so all this needs to be considered when determining what makes a community stable whilst taking into account the natural variation present.

Even though there were similarities in microbial composition between healthy participants, participants still had differences in abundances and presence/absence of taxa which makes it more challenging to use individual variation in finding therapeutic options and treating disease. It may be that each individual needs to be studied in health and response to a disease to determine which bacteria respond and how the microbiome recovers overall. A standardised bacterial community to restore a healthy state may not be useful if healthy individuals vary in their bacterial communities and so personalised treatments may be required.

8.1.3 Comparison of the oropharynx microbiome in non-smokers and smokers

When comparing the healthy non-smoking microbiome to a smoker's microbiome, smokers were found to have high diversity communities (similar to healthy participants) that were distinct in community structure from non-smoking healthy participants. Charlson et al., (2010) observed smokers' communities to be significantly more diverse than non-smokers in that there was greater species richness and samples had high Shannon H indexes. This could be due to smokers having more potentially pathogenic species (Bagaitkar et al., 2008) or an increase in transient species, especially as smokers were shown to have an increased abundance in *Chryseobacteria* which is a common environmental bacterium (Salter et al., 2014). Specific OTUs were found to have increased abundances in the smokers group and these were *Porphyromonas gingavilis*, *Streptococcus pyogenes* and *Streptococcus agalactiae* which are all pathogens involved in oral and respiratory tract infections (Abusleme et al., 2013) (Santagati et al., 2012). Other studies have also reported different community structures in smokers with increased abundances in specific genera such as *Megasphaera*, *Streptococcus, Veillonella* and *Actinomyces* (Charlson et al., 2010) whereas this study showed increases in *Streptococcus*, *Prevotella* and *Porphyromonas*. This shows that smoking results in altered oropharyngeal communities, and also affected individuals in different ways by altering abundances. However this may depend on the participant's individual community structure prior to smoking.

Smokers and non-smoking participants also had different responses to disturbances; non-smokers had high resilience with quick recovery whereas smokers did not display changes to their oropharynx microbiome during periods of a disturbance which suggests a permanently altered state. As a result, this permanently altered state may suggest that the smoker's microbiome is more stable during disturbances which may make them more resilient. Smokers may also have more "invaders" which colonise the oropharyngeal community increasing susceptibility to infection (Bagaitkar et al., 2008). Regardless of the different community structure compared to non-smokers, the smokers' oropharynx was still seen as stable as samples did not drastically change in community composition over a weekly basis (which was also observed in healthy non-smoking participants). However smokers may be more at risk of respiratory disease due to having higher abundances of potentially pathogenic bacteria in their oropharynx microbiome.

8.1.4 Predicted functions of the non-smoking and smoking oropharynx microbiome

The predicted functions of bacteria from the oropharyngeal communities were investigated in the different health groups in non-smoking participants and between smokers and non-smokers. Various functions were predicted from the healthy oropharynx microbiome deemed as necessary in bacteria (that are vital functions) such as protein synthesis to carbohydrate metabolism. In terms of the different health groups, healthy participants had increased abundances of KO's that resulted in higher levels of most enzymes/pathways in comparison to participants with a cold. This could be due to the healthy group having increased diversity of OTUs as well as increased abundances of *Streptococcus, Prevotella* and *Veillonella* OTUs. Microbiomes from virus-positive samples had increased abundances of certain KO enzymes/pathways that were needed for everyday processes but also involved in virulence. Antibiotics usage on the other hand resulted in reduced abundance of the majority of pathways/enzymes in comparison to the healthy groups, which may be indicative of antibiotics eliminating certain groups of bacteria as there were many OTUs that had reduced abundances in comparison to the healthy groups (Langdon et al., 2016).

The bacterial communities in smokers was associated with different predicted functional profiles including reduced abundances of KO pathways/enzymes involved in DNA repair, suggesting that bacteria from smoker's samples were prone to DNA mutations. Smoker's communities had higher abundances of pathogenic bacteria than non-smokers which may contribute to the different functional profiles observed in the smoker's communities. Further investigation is required in understanding if reduced abundances of the DNA repair pathways in smokers (and other pathways) affect health and the bacteria that influence this. Determining the predicted functions of bacteria in healthy samples from non-smokers can show what predicted functions are associated with healthy communities and which can become changed during a disturbance or smoking. Microbiome studies need to address the similarities and differences among participants in both microbial taxa and functional pathways; changes in gene expression and functions can be investigated by transcriptomics which could lead to the use of biomarkers to identify disease.

8.2 Limitations and drawbacks of this study

The strengths of this study included longitudinal analyses, a reasonable number of non-smoking participants and use of high-throughput technology; however the limitations included the modest sample size especially in unhealthy samples and sequencing depth. In this study, a cut off of 5000 reads for each sample was used which still resulted in adequate community coverage. This is important in drawing conclusions regarding species richness, diversity or relative abundance of community members detected. A larger read size would have been ideal, but regardless of this, these results still show that valid conclusions were drawn at a cut off of 5000 reads. A larger sample size would also have resulted in more samples in the healthy as well as unhealthy groups, especially as some participants only had 1 unhealthy sample. As a result the differences observed in the unhealthy microbiome may have been due to natural variation, hence the importance of collecting metadata for each sample. The ratio of males to females in smokers and non-smokers were also different, with more females participating and more samples collected from females than males. This resulted in uneven sample sizes in categories which may influence the data as the changes in the microbiome may have been observed in females rather than males due to having more samples. Having more samples would also address the differences of the normal microbiome between males and females. Microbiome studies should consider power tests and sample size to ensure there are enough samples present in categories ensuring greater statistical power to show a biological difference.

Another limitation was that it was not possible to identify all OTUs to species level through only sequencing the V1-V2 region, even though some *Streptococcus* OTUs could be named. By sequencing a larger region such as the V1-V3 region, more OTUs could be further identified, especially as there are various commensal and pathogenic bacteria present within the same genus. An example is *Streptococcus* in the oropharynx; various commensal and pathogenic *Streptococcus* species exist in the oropharynx and characterisation studies need to address which species are the most abundant in health and disease. However, sequencing larger regions also depend on the technologies capable of doing so, and the literature available supporting this. The V1-V2 region in this study was deemed acceptable due to the literature supporting use of this primer set in oropharynx studies in classifying *Streptococcus* to species level (Chakravorty et al., 2008).

External contamination may have been introduced into the samples either through the participant or from sample/sequencing processing, hence the need for strict metadata collection and negative controls for DNA extraction kits and PCR processes. Previous studies have shown that contamination can be introduced through extraction processes either through kits or water (Salter et al., 2014) amplifying taxa that are ubiquitous in the environment. This is particularly a problem from low biomass sites, where low template DNA concentration is competing for amplification with contaminating DNA (Biesbroek et al., 2012). In order to reduce this issue, specific steps were taken such as ensuring initial DNA template concentrations of 15ng/µl per sample pre PCR amplification, sequencing of negative controls and removal of taxa that were present in the negative controls from all other samples. The negative control in this study had a low number of sequenced reads (<1000) and necessary steps were followed to distinguish contaminating bacteria from actual bacteria representative of the oropharynx microbiome. Contaminating bacteria were assessed on whether they were found in the extraction controls, mock community, in high or low abundance and matched against studies listing common contaminating bacteria in microbiome studies (Salter et al., 2014). Additionally, background contaminating bacteria may also have been accounted for by qPCR but this was not done in this study. However, this only applies to contaminating bacteria from extraction kits and so there is a possibility that the participant may have introduced contaminating bacteria into the sample either through dropping the swab, hitting other surfaces in the mouth or through inefficient swabbing technique or handling of swab which may have not been reported.

An epidemiological limitation is that non-smokers and smokers were sampled in different years. Therefore some of the differences observed in the non-smokers and smoker's microbiome may be due to differences between years and sampling months. These differences could be attributed to outbreaks of influenza, temperature changes and seasonal changes such as fluctuating pollen levels. However, as other research studies reported similar findings when comparing the non-smokers and smokers microbiome, this shows that there must also be other factors other than the different sampling years responsible for these differences in microbiome composition.

In regards to stability, even though the microbiome was found to be stable in all participants, determining the stability of the microbiome over a longitudinal period was a challenge as samples were not present every week as participants did miss some weeks of swabbing. Participants were given strict swabbing instructions to reflect the oropharynx microbiome. However, they may have also inaccurately reported symptoms due to forgetting, being rushed or reporting the wrong symptoms. This could explain why some samples that had no associated symptoms of disease or illness had very distinct, different communities for which there were no obvious explanations.

Lastly, as microbiome based studies are based on observations, there is still the question of linking these observations in changes in community structure to the cause or effect of the disease. This is a limitation in most microbiome studies where there is limited information or data to determine whether the changes in the microbiome are caused by the disease or an effect of the disease. As this sampling was done on a weekly basis, it was not possible to determine the exact changes that occurred during the health disturbance (to do this, daily sampling would be required) and therefore this is a limitation of the current study. However, it should be noted that it was not possible to know the frequency of the fluctuations of the oropharynx microbiome beforehand, and a balance between what was reasonable to ask from the participants and the aim of covering a relatively long time period had to be found. Another way to approach the question of cause or effect in microbiome studies would be to use an animal model where the effects of a certain disease through inoculation and the changes in community structure can be observed. This could be a possibility especially in determining if certain taxa can cause respiratory disease and if certain key taxa can restore the oropharynx microbiome.

Therefore to improve the study there would need to be a larger sample size with more participants to ensure greater statistical power and reliability of results. There would need to be strict swabbing procedures and metadata collection to be able to link diseased samples to their symptoms. To also try and address the cause and effect of diseases on the microbiome, animal models can be used to determine how the microbiome responds to specific disturbances. Sampling should also be continued longitudinally, daily rather than weekly to try and capture the changes that occurred to the microbiome that may have been missed when sampling on a weekly basis.

8.3 Significance and wider implications of the current study

One of the key challenges in any microbiome study is to determine whether and how a given microbiome affects human health (Cho & Blaser, 2012). Microbiome studies commonly focus on characterising microbial communities in specific disease states or trying to determine the changes that occur during the course of a disease. However, demonstrating causality between the microbial variation and pathology instead of mere association is often extremely difficult as controlled experiments are not yet practical in most cases. However controlled studies showing a cause and effect response has been demonstrated through the use of faecal studies to restore the microbiome of the GI tract (Aroniadis & Brandt, 2014). In this study, the healthy oropharynx microbiome was characterised using longitudinal sampling and respiratory disturbances such as the common cold were investigated by linking back to metadata which allows establishing the timeline of events. This design allowed not only determining what would constitute a healthy oropharyngeal microbial community, but also investigating how the community changed during a disturbance and on antibiotic treatment. Therefore the cause and effect relationship was not investigated, but this is the only study (to date) that has looked at the oropharynx microbiome over a defined period of time on a weekly basis. This is also the only study to determine how the community changes before, during and after a disturbance, again providing novel information on how the microbiome changes during a disturbance, and its recovery time and resilience.

Another significant finding of this thesis was determining the changes that occurred to the microbiome of smokers in comparison to non-smokers and investigating the stability and recovery from a health disturbance of the two groups. The changes observed were present in abundances and presence/absence in taxa and function showing distinct changes in the oropharynx microbiome between smokers and non-smokers. Smokers also responded differently during a disturbance in that significant differences were not observed before, during or after a disturbance suggesting that smokers may possibly have a permanent altered microbiome. This is the only study investigating the stability of the smokers' microbiome over a defined period of time and the weekly changes that occur during a disturbance.

Therefore, the significance of this project overall is that it has provided new knowledge about the oropharynx microbiome in non-smoking participants and smokers. This knowledge provides a solid starting point for further investigations; for instance, to first develop deeper understanding on the ecological interactions between bacteria now known to be present and fluctuating in abundance, and perhaps in future, once the fluctuations of the microbiome members are understood better, this knowledge could be used to manipulate or restore a person's microbiome during or after a respiratory disease. A smoker's microbiome could also be manipulated to mimic the microbiome of a non-smoker.

This thesis may be used as increased knowledge of the oropharynx microbiome to improve diagnosis of disease states. This could result in treatments such as probiotic supplementation (with the necessary taxa) to restore microbiome balance that may be a useful and necessary treatment against respiratory disease and infections. The future work required for this project is as follows: longitudinal sampling and characterisation of the oropharynx microbiome over a long period of time (daily and weekly sampling over months to years) in healthy participants (non-smokers with no underlying disease) to understand natural variation of taxa in healthy people, daily sampling to catch progress of recovery from confirmed infections while ensuring large sample sizes of healthy and unhealthy samples and production of metagenomic profiles of the oropharynx in healthy participants and specific diseased scenarios (and altered communities such as smoking) through sequencing of all genes in the community rather than just focussing on one gene – this will also detail the viruses, archaea and eukaryotic microorganisms that also inhabit the oropharynx.

8.4 Future for microbiome studies

As microbiome studies are on the increase, there are now various new issues that need to be considered and addressed. Determining what microbes are there is no longer enough, there needs to be an understanding of the ecology of these communities (Costello et al., 2012). For example it is still uncertain whether the microbiome can help protect the host from infection or whether it can aid in developing infection. How are these communities built, what defines the structure and how do these communities change in time? Microbial ecology is important as it provides an understanding of how microbes interact with each other, if diversity affects the stability of communities and whether patterns of co-existence are observed among microbiomes and if this is indicative of health and disease. Microbial ecological theory also impacts pharmaceuticals, food production, diagnosis/treatment and industrial applications (Ursell et al., 2012). The future of investigating the healthy microbiome involves understanding natural variation in taxa and function in healthy participants (Lloyd-Price et al., 2016), how various factors such as age, sex and diet affects variation, whether observed differences are the cause or effects of a certain disease, and the potential restoration of microbial ecology - either through restoring certain taxa in healthy hosts or increasing biodiversity. Microbiome studies also need to address how to manipulate the microbiome through intervention trials. This could be through a probiotic drug which would require long-term follow up to determine how the microbiome responds to the intervention. This would determine the differences in microbiome structure before, during and after the intervention, with the necessary controls in place. Therefore there will always be a need to identify the taxonomic compositions of the microbiome, but it may also be useful and necessary to start focussing on functions and finding target pathways in healthy scenarios as well as those that have altered expression levels in diseased states.

8.5 Concluding remarks

The work presented in this thesis provides a better understanding of the oropharynx microbiome in healthy non-smokers, and how this community is affected by respiratory infections and disturbances. It showed the oropharynx microbiome to be a stable community over time, with distinct differences apparent between the oropharynx of non-smokers and smokers, in both community composition and predicted function. Using the information gathered in this thesis, alongside future research, it may be possible in future to diagnose and treat respiratory disease through analysing and manipulating the oropharynx microbiome.

Appendix 1

QIA-AMP DNA extraction protocol

1. Store swabs in transport medium during transport to the laboratory.
If swabs are not processed immediately they should be stored at 2-8°C for up to 24 hours. Any period longer than this will be require storage at -20°C.

2. All swabs should be vortexed to ensure dispersal of microorganisms from swab to fluid.

3. Remove swab and place in a 2ml microcentrifuge tube. Centrifuge swab tip at 5,000 x g or 8,000rpm to remove any remaining fluid in the tube.
After centrifugation remove swab tip and transfer 1ml of suspension fluid into the same 2ml microcentrifuge tube and centrifuge for 10 minutes at full speed (20,000 x g; 14,000rpm).

4. Suspend pellet in 180µl of enzymatic lysis buffer (20mg/ml lysozyme or 200µg/ml lysostaphin; 20mM TrisHCl, pH 8.0; 2mM EDTA; 1.2% Triton).
If there is no pellet formed or pellet formation is very small, remove as much supernatant as possible without touching the bottom of the tube and add 180µl of enzymatic lysis buffer to the same microcentrifuge tube.

5. Incubate for at least 30 minutes at 37°C.

6. Add 20µl Proteinase K and 200µl Buffer AL. Mix by vortexing for 10 seconds.

7. Incubate at 56°C for 1 hour. If using a heat block, vortex the tube for 10 seconds every 10 minutes.

8. Centrifuge for a few seconds to remove drops from inside the lid.

9. Incubate the 2ml microcentrifuge at 70°C for 10 minutes. If using a heat block vortex the tube for 10 seconds every 3 minutes to improve lysis.

10. Centrifuge for a few seconds to remove drops from inside the lid.

11. Add 200µl ethanol (96-100%) to the sample and mix by vortexing for 10 seconds.

12. Centrifuge for a few seconds to remove drops from inside the lid.

13. Carefully transfer the lysate from the 2ml microcentrifuge tube into a QIAamp Mini spin column (2ml collection tube).
Close the cap and centrifuge at 8000 x g (6000rpm) for 1 minute (if the lysate has not completely passed through the 2ml column after centrifugation, centrifuge at a higher speed until the QIAamp Mini spin column is empty).
Place the QIAamp Mini spin column in a clean 2ml collection tube and discard the tube containing the filtrate. Transfer any remaining lysate from the 2ml microcentrifuge tube and repeat as above.

14. Add 500µl of Buffer AW1 to the QIAamp Mini spin column. Close the cap and centrifuge at 8000 x g (6000rpm) for 1 minute. Place the QIAamp Mini spin column in a clean 2ml collection tube and discard the tube containing the filtrate.

15. Add 500µl of Buffer AW2 to the QIAamp Mini spin column. Close the cap and centrifuge at 8000 x g (6000rpm) for 1 minute. Place the QIAamp Mini spin column in a clean 2ml collection tube and discard the tube containing the filtrate.

16. Centrifuge at full speed (20,000 x g; 14,000rpm) for 3 minutes to dry the membrane.

17. Place the QIAamp Mini spin column in a clean 1.5ml microcentrifuge tube and discard the collection tube containing the flow-through. Carefully open the lid of the QIAamp Mini spin column and apply 50µl Buffer AE (having 2 centrifugation steps of 25µl Buffer AE and a final re-elution step increases DNA yield).

18. Close the lid and incubate at room temperature for 5 minutes. Centrifuge at full speed (20,000 x g or 14,000rpm) for 1 minute.

19. Qubit.

20. Run 10µl of DNA extract with 2µl loading buffer on a gel to check purity (best to run extractions on a 1% gel -1g agarose to 100ml TAE or TBE.)
If using Bioline mix reagents loading gel does not need to be used on the gel.

21. Run at 100v for 50 minutes.

22. Store DNA at -20°C until required.

Appendix 2
Production of an rDNA clone library

Production of PCR products:

1. Set up the following 50µl PCR reaction:

25µl Bioline PCR mix
1µl primer (forward and reverse) at 12.5pmol each
2µl DNA template
21µl water
Total Volume 50µl

2. The PCR reaction should run under the following conditions:

Initial denaturation - 95°C for 5 minutes
35 cycles of denaturation - 94°C for 1 minute
Annealing - 62°C for 1 minute
Extension - 72°C for 1 minute
Final Extension - 72°C for 10 minutes

3. Check the PCR product by agarose gel electrophoresis to ensure production of a single discrete band.

QIA gel extraction kit protocol – extraction and purification of DNA from agarose gels:

1. Cut DNA fragment from the agarose gel with a scalpel.

2. Weigh the gel slice in a colourless tube.
Add 3 volumes of buffer QG to 1 volume of gel (100mg - 100µl).
For >2% agarose gels, add 6 volumes of Buffer QG. The maximum amount of gel slice per QIAquick column is 400mg; for gel slices >400mg use more than one

QIAquick column.

3. Incubate at 50°C for 10 minutes.
Vortex every 2-3 minutes to dissolve the gel.

4. After the gel slice has dissolved completely, check that the colour of the mixture is yellow (similar to Buffer QG without dissolved agarose). If the colour of the mixture is orange or violet, add 10µl of 3M sodium acetate, pH 5.0, and mix. The colour of the mixture will turn yellow.

5. Add 1 gel volume of isopropanol to the sample and mix (if the agarose gel slice is 100mg, add 100µl isopropanol).

6. Place a QIAquick spin column in a provided 2ml collection tube.

7. To bind DNA, apply the sample to the QIAquick column and centrifuge (1 minute). Discard flow-through and place QIAquick column back in the same collection tube.

8. Add 300ml of buffer QG to QIAquick column and centrifuge for 1 minute to remove all traces of agarose.

9. To wash, add 300ml of Buffer PE to QIAquick column (stand for 2-5 minutes) and centrifuge for 1 minute.

10. Discard the flow-through and centrifuge the QIAquick column for an additional 1 minute at > 10,000 x g (13,000rpm).

11. Place QIAquick column into a clean 1.5ml microcentrifuge tube.

12. To elute DNA, add 30µl of Buffer EB (10mM TrisCl, pH 8.5) to the centre of the QIAquick membrane and centrifuge the column for 1 minute at maximum speed. For an increased DNA concentration, let the column stand for 5 minutes, and then centrifuge for 1 minute.

Production of lunia bertani broth and agar plates:

For 500ml broth:

5g – Tryptone

2.5g – Yeast agar

5g – NaCl

Add water to make a total volume of 500ml and autoclave.

When broth has cooled add 5ml of Kanamycin.

For 200ml agar plates:

Add 3g of agar in 200ml of water and autoclave.

When agar has cooled add 2ml of Kanamycin.

Flame the bottle top using a Bunsen burner.

Pour into plates.

Flame agar plates to get rid of any bubbles.

Let plates harden and store in fridge.

Performing TOPO cloning reaction:

1. Set up the following reaction:

Fresh PCR product – 0.5 - 4µl

Salt solution - 1µl

Water – add to a total volume of 5µl

TOPO vector - 1µl

Total volume - 6µl

The cloning reaction can be stored overnight at -20°C.

2. Mix the reaction gently and incubate for 30 minutes at room temperature (22°C - 23°C).

3. Place the reaction on ice ready for the next step.

Transforming cells:

1. Warm selective plates to 37°C prior to spreading.

2. Add 2µl of the TOPO cloning reaction into a vial of OneShot®Chemically competent *E. coli* and mix gently.

3. Incubate on ice for 30 minutes.

4. Heat shock the cells for 30 seconds at 42°C without shaking.

5. Immediately transfer the tubes to ice.

6. Add 250µl of S.O.C medium.

7. Cap the tube tightly and shake the tube horizontally (200rpm) at 37°C for 1 hour.

8. Spread 10-50µl from each transformation on a pre-warmed selective plate. To ensure even spreading of small volumes, add 20µl of S.O.C medium (plate two different volumes to ensure that at least one plate will have well-spaced colonies). Spread at least 5 plates in total.

9. Incubate plates at 37°C (ampicillin plate should produce colonies within 8 hours whereas kanomycin plates should be incubated overnight).

10. An efficient TOPO cleaning reaction should produce several hundred colonies. Pick 100-200 colonies for analysis.

Analysing transformants:

1. Pick 100-150 colonies for analysis.

2. Aliquot 500µl of LB kanamycin into a 96 deep well plate.

3. By the flame of a Bunsen burner, using a pipette and clean tip, pick up a single colony and inoculate a well in the deep well plate. Incubate overnight with shaking at 37°C at 200rpm.

4. Screen the colonies the next day for the correct insert.

5. Remove 200µl of each sample in the deep well plate (96 samples altogether) to be amplified using the M13F and M13R primers.

6. Set up the following PCR reaction to ensure the samples take up the vector:
Biomix – 10µl
M13F primer – 0.5µl
M13R primer – 0.5µl
Water - 7µl
DNA - 2µl
Total volume = 20µl

The reaction should run for 25-30 cycles at the following conditions:
95°C – 5 minutes
94°C – 1 minute
55°C – 1 minute
72°C – 1 minute
72°C – 10 minutes

Run a gel at 100v for 60 minutes to ensure that each sample has taken up the plasmid.

Restriction digest:

1. Do a restriction digestion on all 96 samples.

2. Set up the following reaction:
PCR reaction - 5µl
Hae 111 restriction digest – 0.1µl
Water – 8.4µl
Buffer – 1.5µl
Total volume = 15µl

3. Centrifuge tubes and incubate at 37°C for 4 hours.

4. Prepare a 2% gel for all 96 isolates (add 3µl of loading dye to the 15µl reaction). Run the gel at 100v for 90 minutes.

5. From the gel group isolates into OTUs depending on the same banding patterns. Each type of OTU will be sent for Sanger Sequencing.

6. Grow each OTU selected for Sanger Sequencing overnight in 5ml of LB and kanamycin (37°C, 200rpm for 20-24 hours).

Plasmid extractions:

1. Extract the DNA from the plasmid using the Invitrogen Purelink Quick Plasmid MiniPrep Kit.

2. Centrifuge 1–5ml of the overnight LB-culture (5 minutes at 5000rpm).

3. Add 250µl Resuspension Buffer (R3) with RNase A to the cell pellet and resuspend the pellet until it is homogeneous.

4. Add 250µl Lysis Buffer (L7). Mix gently by inverting the capped tube until the mixture is homogeneous. Do not vortex. Incubate the tube at room temperature for 5 minutes.

5. Add 350µl Precipitation Buffer (N4). Mix immediately by inverting the tube or for large pellets by vigorously shaking the tube, until the mixture is homogeneous. Do not vortex. Centrifuge the lysate at >12,000 × *g* for 10 minutes.

6. Load the supernatant from step 4 onto a spin column in a 2ml wash tube. Centrifuge the column at 12,000 × *g* for 1 minute. Discard the flow-through and place the column back into the wash tube.

7. Add 500µl Wash Buffer (W10) with ethanol to the column. Incubate the column for 1 minute at room temperature. Centrifuge the column at 12,000 × *g* for 1 minute. Discard the flow-through and place column back into the wash tube.

8. Add 700µl Wash Buffer (W9) with ethanol to the column. Centrifuge the column at 12,000 × *g* for 1 minute. Discard the flow-through and place the column into the wash tube. Centrifuge the column at 12,000 × *g* for 1 minute. Discard the wash tube with the flow-through.

9. Place the spin column in a clean 1.5ml recovery tube. Add 75µl of preheated TE Buffer (TE) to the centre of the column. Incubate the column for 1 minute at room temperature.

10. Centrifuge the column at 12,000 × *g* for 2 minutes. *The recovery tube contains the purified plasmid DNA.* Discard the column. Store plasmid DNA at 4°C (short-term) or store the DNA in aliquots at −20°C (long-term).

Long term storage:

1. Streak out the original colony on LB plates.

2. Isolate a single colony and inoculate into 1-2ml of LB.

3. Grow overnight until culture is saturated.

4. Mix 0.85ml of culture with 0.15ml of sterile glycerol and transfer to a cryovial.

5. Store at -80°C.

Appendix 3

QIAGEN QIA gel extraction kit protocol

1. Cut the DNA fragment from the agarose gel with a scalpel.

2. Weigh the gel slice in a colourless tube. Add 3 volumes of buffer QG to 1 volume of gel (100ml - 100µl). For example, add 300µl of Buffer QG to each 100 mg of gel. For >2% agarose gels, add 6 volumes of Buffer QG. The maximum amount of gel slice per QIAquick column is 400mg; for gel slices >400mg use more than one QIAquick column.

3. Incubate at 50°C for 10 minutes.

4. After the gel slice has dissolved completely, check that the colour of the mixture is yellow (similar to Buffer QG without dissolved agarose). If the colour of the mixture is orange or violet, add 10µl of 3M sodium acetate, pH 5.0, and mix. The colour of the mixture will turn to yellow.

5. Add 1 gel volume of isopropanol to the sample and mix (if the agarose gel slice is 100mg, add 100µl isopropanol).

6. Place a QIAquick spin column in a provided 2ml collection tube.

7. To bind DNA, apply the sample to the QIAquick column and centrifuge for 1 minute.

8. Discard flow-through and place QIAquick column back in the same collection tube.

9. Add 0.5ml of Buffer QG to QIAquick column and centrifuge for 1 minute to remove all traces of agarose.

10. To wash, add 0.75ml of Buffer PE to the QIAquick column and centrifuge for an additional 1 minute.

11. Discard the flow-through and centrifuge the QIAquick column for an additional 1 minute at > 10,000 x g (13,000rpm).

12. Place QIAquick column into a clean 1.5ml microcentrifuge tube.

13. To elute DNA, add 50µl of Buffer EB (10mM TrisCl, pH 8.5) to the centre of the QIAquick membrane and centrifuge the column for 1 minute at maximum speed. For an increased DNA concentration, add 30µl elution buffer to the centre of the QIAquick membrane, let the column stand for 1 minute, and then centrifuge for 1 minute.

Appendix 4

Supplementary Table 3.1

Supplementary Table 3.1 – Descriptive statistics of the most abundant phyla from normalised communities in healthy samples.

	Firmicutes	*Bacteroidetes*	*Proteobacteria*	*Actinobacteria*	*Fusobacteria*
samples	279	279	279	279	279
min	0.147	0.008	0.003	0.004	0.001
max	0.946	0.558	0.614	0.219	0.230
range	0.799	0.549	0.611	0.214	0.229
sum	169.21	43.57	29.99	19.85	12.575
median	0.612	0.142	0.063	0.063	0.032
mean	0.606	0.156	0.107	0.071	0.045
SE.mean	0.009	0.005	0.007	0.002	0.002
CI.mean (0.95)	0.019	0.011	0.014	0.005	0.004
var	0.027	0.010	0.014	0.001	0.001
std.dev	0.164	0.100	0.121	0.044	0.041
coef.var	0.271	0.640	1.130	0.619	0.919

Appendix 5

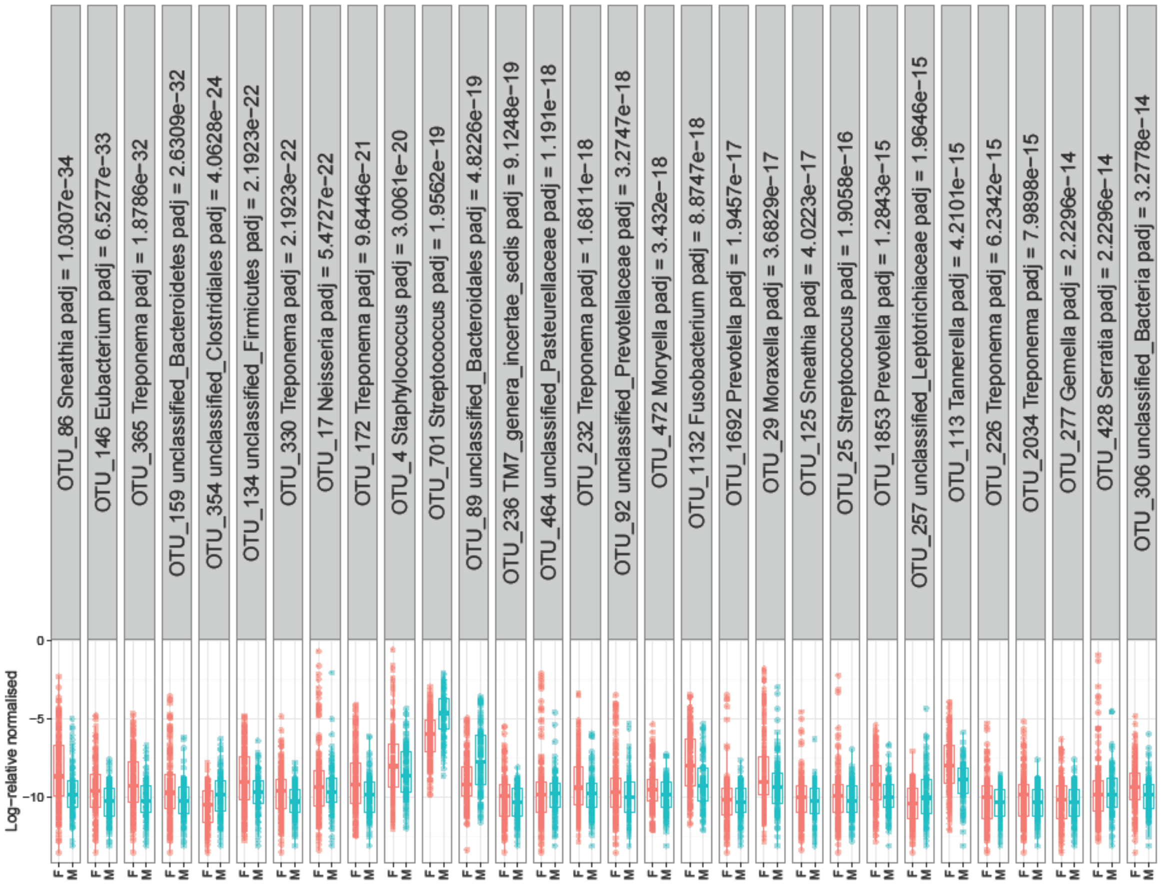
Supplementary Table 3.2

Supplementary Table 3.2 – Descriptive statistics of the top 5 most abundant genera from normalised communities in healthy samples.

	Streptococcus	*Prevotella*	*Veillonella*	*Neisseria*	*Actinomyces*
samples	279	279	279	279	279
min	0.061	0.002	0.004	0	0.002
max	0.899	0.498	0.312	0.548	0.163
range	0.838	0.495	0.307	0.548	0.161
sum	131.2	26.42	16.27	14.92	11.26
median	0.469	0.077	0.049	0.022	0.031
mean	0.470	0.094	0.058	0.053	0.040
SE.mean	0.011	0.004	0.002	0.005	0.001
CI.mean (0.95)	0.022	0.008	0.004	0.009	0.003
var	0.035	0.005	0.001	0.007	0.001
std.dev	0.187	0.075	0.041	0.084	0.033
coef.var	0.399	0.796	0.708	1.581	0.825

Appendix 6

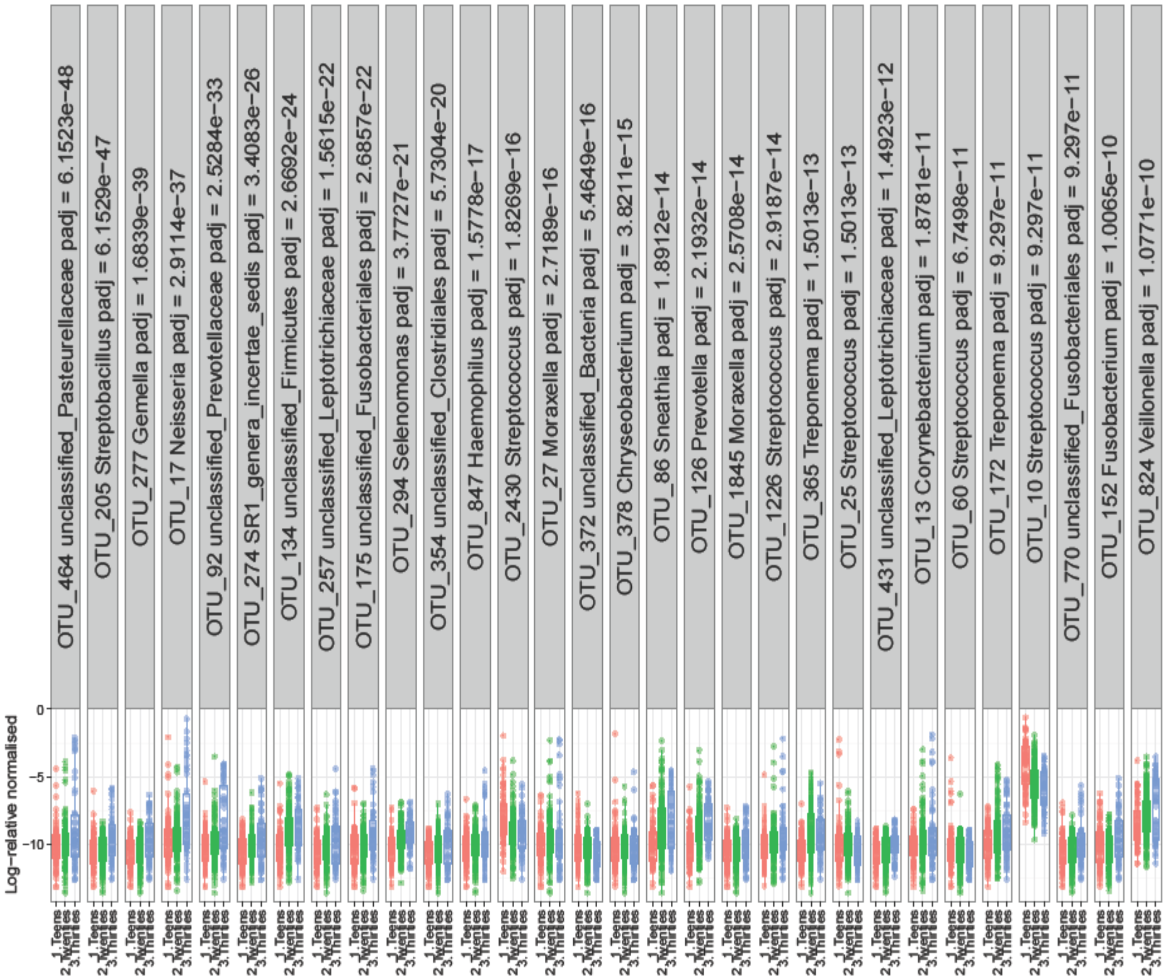
Supplementary Figure 3.1



Supplementary Figure 3.1 - Plot showing the differences in microbiome composition regarding sex in healthy samples (n=279). Negative binomial GLMs were performed using DESeq2 package to show the 30 most significant OTUs and the log relative abundance of each OTU in males (M) and females (F).

Appendix 7

Supplementary Figure 3.2



Supplementary Figure 3.2 - Plot showing the differences in microbiome composition due to age in healthy samples (n=279). Negative binomial GLMs were performed using DESeq2 package to show the 30 most significant OTUs and the log relative abundance of each OTU in three different age categories – teens, twenties and thirties.

Appendix 8

Supplementary Table 4.1

Supplementary Table 4.1 – Descriptive statistics of the top 5 most abundant phlya from normalised communities in unhealthy samples.

	Firmicutes	*Bacteroidetes*	*Proteobacteria*	*Actinobacteria*	*Fusobacteria*
samples	34	34	34	34	34
min	0.138	0.003	0.002	0.002	0.001
max	0.966	0.467	0.825	0.184	0.285
range	0.828	0.462	0.823	0.182	0.284
sum	17.44	3.305	9.351	2.308	1.419
median	0.512	0.087	0.121	0.059	0.014
mean	0.513	0.097	0.275	0.067	0.041
SE.mean	0.043	0.017	0.052	0.009	0.011
CI.mean (0.95)	0.089	0.034	0.106	0.019	0.021
var	0.065	0.009	0.093	0.003	0.003
std.dev	0.256	0.099	0.306	0.056	0.061
coef.var	0.498	1.021	1.114	0.827	1.471

Appendix 9

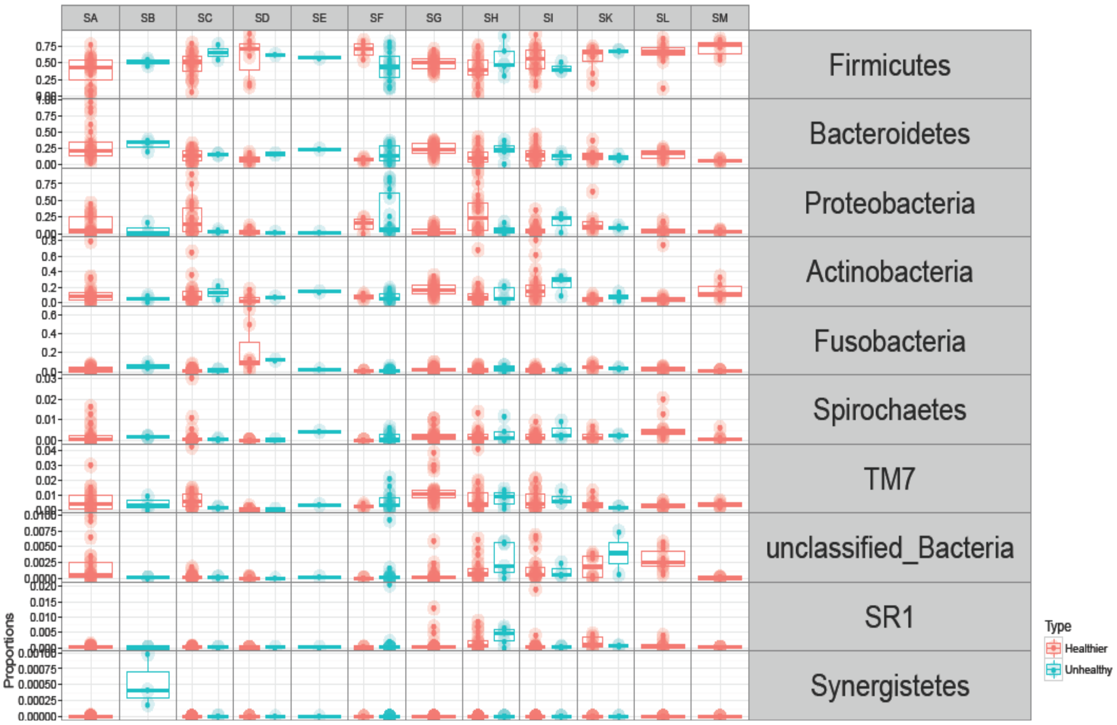
Supplementary Table 4.2

Supplementary Table 4.2 – Descriptive statistics of the top 5 most abundant genera from normalised communities in unhealthy samples.

	Streptococcus	*Pseudomonas*	*Prevotella*	*Serratia*	*Veillonella*
samples	34	34	34	34	34
min	0.043	0	0.001	0	0.001
max	0.932	0.699	0.165	0.776	0.191
range	0.889	0.699	0.164	0.776	0.189
sum	13.06	1.947	1.815	1.680	1.645
median	0.322	0.0001	0.042	0.001	0.033
mean	0.384	0.057	0.053	0.049	0.048
SE.mean	0.041	0.028	0.008	0.030	0.008
CI.mean (0.95)	0.083	0.057	0.017	0.062	0.016
var	0.057	0.027	0.002	0.032	0.002
std.dev	0.240	0.165	0.051	0.179	0.048
coef.var	0.625	2.880	0.964	3.635	0.999

Appendix 10

Supplementary Figure 6.1



Supplementary Figure 6.1 – Box plot showing the most abundant phyla (n=10) and the mean relative abundance in healthy (red) and unhealthy (blue) samples from smoking participants.

Appendix 11

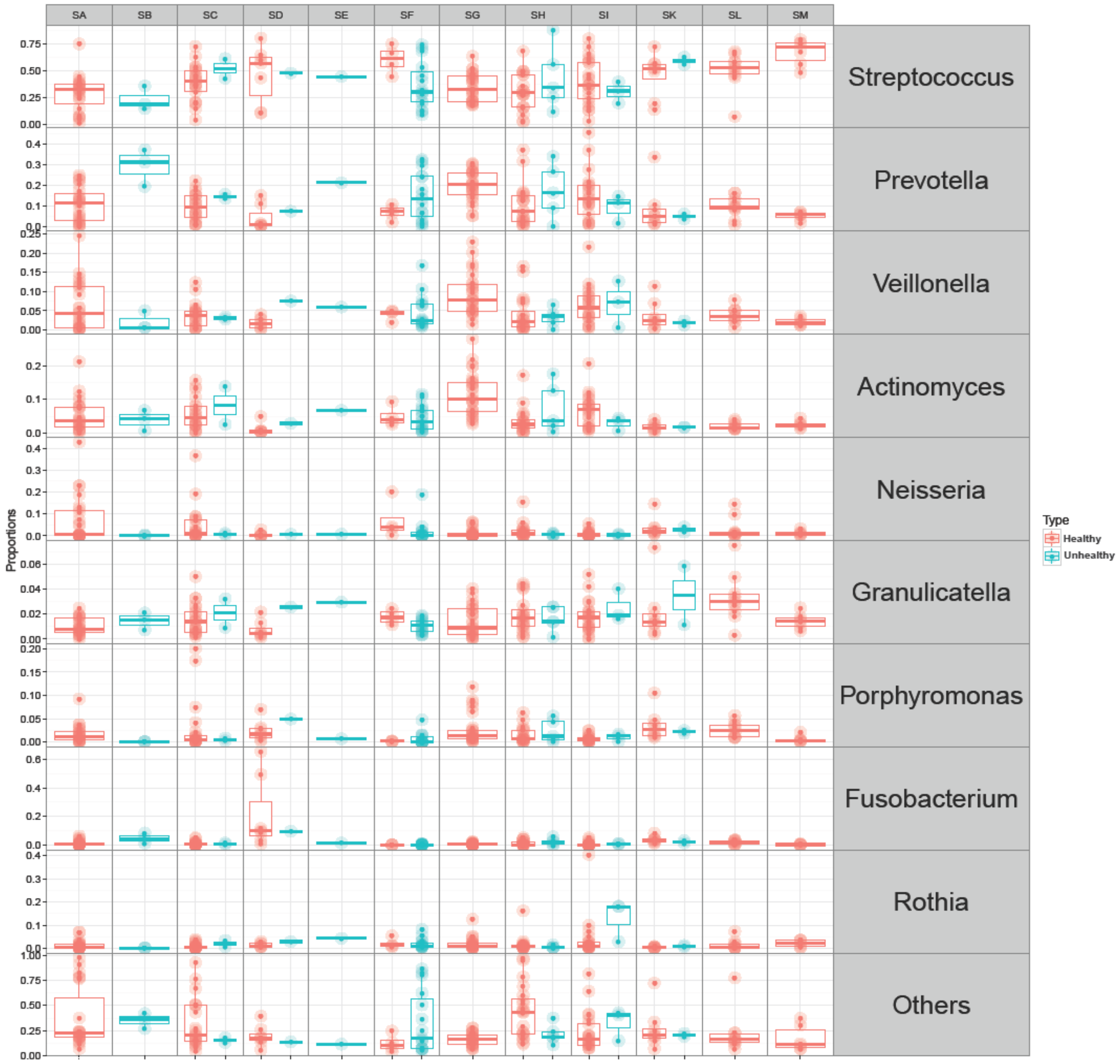
Supplementary Table 6.1

Supplementary Table 6.1 – Descriptive statistics of the most abundant phyla from normalised communities of smokers. Healthy communities are denoted by SH and unhealthy communities by SUH.

	Firmicutes		*Bacteroidetes*		*Actinobacteria*		*Proteobacteria*		*Fusobacteria*	
Health	SH	SUH	SH	SUH	SH	SUH	SH	SUH	SH	SUH
samples	145	32	145	32	145	32	145	32	145	32
min	0.012	0.126	0.011	0.002	0.003	0.008	0.001	0.001	0.001	0.001
max	0.949	0.904	0.975	0.382	0.811	0.351	0.943	0.832	0.661	0.120
range	0.937	0.778	0.964	0.379	0.808	0.343	0.943	0.831	0.661	0.120
sum	73.12	16.19	26.80	5.789	19.37	3.236	19.33	5.677	4.677	0.766
median	0.523	0.511	0.157	0.166	0.085	0.071	0.041	0.054	0.014	0.014
mean	0.504	0.506	0.184	0.180	0.133	0.101	0.133	0.177	0.032	0.023
SE.mean	0.016	0.034	0.012	0.020	0.012	0.015	0.016	0.045	0.005	0.004
CI.mean (0.95)	0.033	0.070	0.025	0.042	0.024	0.030	0.031	0.092	0.011	0.010
var	0.041	0.038	0.023	0.013	0.022	0.007	0.037	0.066	0.004	0.001
std.dev	0.203	0.196	0.154	0.116	0.151	0.085	0.194	0.257	0.071	0.028
coef.var	0.402	0.387	0.836	0.644	1.131	0.846	1.458	1.451	2.188	1.176

Appendix 12

Supplementary Figure 6.2



Supplementary Figure 6.2 – Box plot showing the most abundant genera (n=9) (with the rest pooled in the 'Others' category) and the median abundance in healthy (red) and unhealthy samples (blue) from smoking participants.

Appendix 13

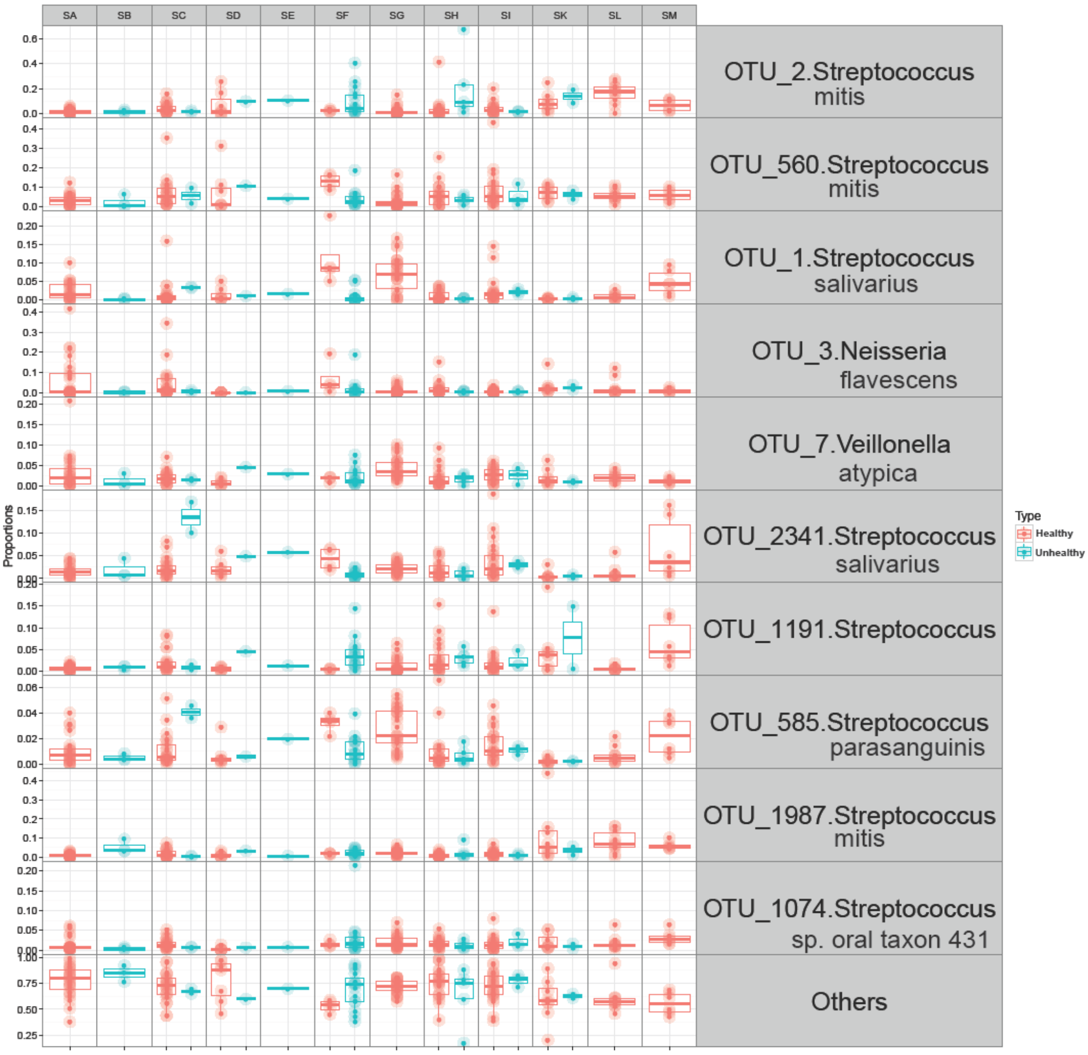
Supplementary Table 6.2

Supplementary Table 6.2 – Descriptive statistics of the top 5 most abundant genera from normalised communities in healthy samples from smokers (SH).

	Streptococcus	*Prevotella*	*Actinomyces*	*Veillonella*	*Neisseria*
Health	SH	SH	SH	SH	SH
samples	145	145	145	145	145
min	0.007	0.002	0.001	0.001	0.001
max	0.801	0.452	0.281	0.244	0.428
range	0.793	0.449	0.279	0.243	0.428
sum	56.18	17.22	8.102	7.759	4.596
median	0.387	0.108	0.035	0.041	0.007
mean	0.387	0.118	0.055	0.053	0.031
SE.mean	0.016	0.007	0.004	0.004	0.005
CI.mean (0.95)	0.031	0.015	0.008	0.008	0.011
var	0.037	0.008	0.002	0.002	0.004
std.dev	0.193	0.092	0.054	0.050	0.063
coef.var	0.498	0.778	0.976	0.938	2.012

Appendix 14

Supplementary Figure 6.3



Supplementary Figure 6.3 – Box plot showing the most abundant OTUs (n=10) (with the rest pooled in the ‘Others’ category) and the median abundance in healthy (red) and unhealthy samples (blue) from smoking participants.

Appendix 15

Supplementary Table 6.3

Supplementary Table 6.3 – Descriptive statistics of the top 5 most abundant genera from normalised communities in unhealthy samples from smokers (SUH).

	Streptococcus	***Prevotella***	***Serratia***	***Actinomyces***	***Veillonella***
Health	SUH	SUH	SUH	SUH	SUH
samples	32	32	32	32	32
min	0.086	0.001	0	0.003	0.001
max	0.878	0.372	0.795	0.176	0.168
range	0.791	0.370	0.795	0.173	0.167
sum	12.28	4.903	2.839	1.522	1.377
median	0.346	0.142	0.001	0.035	0.028
mean	0.383	0.153	0.088	0.047	0.043
SE.mean	0.036	0.019	0.041	0.007	0.006
CI.mean (0.95)	0.074	0.040	0.083	0.015	0.013
var	0.042	0.012	0.053	0.001	0.001
std.dev	0.206	0.111	0.231	0.044	0.038
coef.var	0.537	0.726	2.612	0.931	0.896

References

Abusleme L, Dupuy AK, Dutzan N, Silva N, Burleson JA, Strausbaugh LD, et al., (2013). The subgingival microbiome in health and periodontitis and its relationship with community biomass and inflammation. *The ISME journal*, 7(5), 1016–1025.

Allen EK, Koeppel AF, Hendley JO, Turner SD, Winther B & Sale MM. (2014). Characterization of the nasopharyngeal microbiota in health and during rhinovirus challenge. *Microbiome*, 2(22), 1-11.

Altschul SF, Gish W, Miller W, Myers EW & Lipman DL. (1990). Basic Local Alignment Search Tool. *J. Mol. Biol*, 215, 403–410.

Aroniadis OC & Brandt LJ. (2014). Intestinal microbiota and the efficacy of fecal microbiota transplantation in gastrointestinal disease. *Gastroenterology & Hepatology*, 10(4), 230–237.

Arumugam M, Raes J, Pelletier E, et al., (2011). Enterotypes of the human gut microbiome. *Nature*, 473(7346), 174–180.

Aßhauer KP, Wemheuer B, Daniel R, et al., (2015). Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics*, 31, 2882–2884.

Bagaitkar J, Demuth DR & Scott DA. (2008). Tobacco use increases susceptibility to bacterial infection. *Tobacco induced diseases*, 4(12), 1-10.

Bartlett JMS & Stirling D. (2003). *Protocols*. Methods in Molecular Biology, 226(2), 20-23.

Biesbroek G, Sanders EAM, Roeselers G, Wang X, Caspers MPM, Trzciński K, et al., (2012). Deep sequencing analyses of low density microbial communities: working at the boundary of accurate microbiota detection. *PloS one*, 7(3), 1-9.

Biggar KK & Storey KB. (2014). New Approaches to Comparative and Animal Stress Biology Research in the Post-genomic Era: A Contextual Overview. *Computational and Structural Biotechnology Journal*, 11, 138–146.

Blaser MJ, Chen Y & Reibman J. (2008). Does Helicobacter pylori protect against asthma and allergy? *Gut*, 57(5), 561–567.

Bogaert D, Keijser B, Huse S, Rossen J, Veenhoven R, van Gils E, et al., (2011). Variability and diversity of nasopharyngeal microbiota in children: a metagenomic analysis. *PloS one*, 6(2), 1-8.

Bolnick DI, Snowberg LK, Hirsch PE, et al., (2014). Individual diet has sex-dependent effects on vertebrate gut microbiota. *Nature communications*, 5(4500), 1-13.

Boonanantanasarn K, Gill AL, Yap YS, et al., (2012). Enterococcus faecalis enhances cell proliferation through hydrogen peroxide-mediated epidermal growth factor receptor activation. *Infection and immunity*, 80(10), 3545–3558.

Botero LE, Delgado-Serrano L, Cepeda ML, et al., (2014). Respiratory tract clinical sample selection for microbiota analysis in patients with pulmonary tuberculosis. *Microbiome*, 2(29), 1-7.

Browne HP, et al., (2016). Culturing of "unculturable" human microbiota reveals novel taxa and extensive sporulation. *Nature*, 533(7604), 543-546.

Cabrera-Rubio R, Garcia-Nunez M, Seto L, et al., (2012). Microbiome diversity in the bronchial tracts of patients with chronic obstructive pulmonary disease. *Journal of clinical microbiology*, 50(11), 3562–3568.

Camelo-Castillo AJ, Mira A, Pico A, et al., (2015). Subgingival microbiota in health compared to periodontitis and the influence of smoking. *Frontiers in microbiology*, 6(119), 1-12.

Caporaso JG, Kuczynski J, Stombaugh J, et al., (2011). QIIME allows analysis of high-throughput community sequencing data. *Nat methods*, 7(5), 335–336.

Carter BD, et al., (2015). Smoking and Mortality - Beyond Established Causes. *The New England journal of medicine*, 372, 631–640.

Castro-nallar E, Bendall M, Perez-Losada M, et al., (2015). Composition , taxonomy and functional diversity of the oropharynx microbiome in individuals with schizophrenia and controls. *PeerJ*, 3, 1–21.

Chakravorty S, Helb D, Burday M, et al., (2008). A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *J Microbiol Methods*. 69(2), 330–339.

Charlson ES, Chen J, Custers-Allen R, Bittinger K, Li H, Sinha R, et al., (2010). Disordered microbial communities in the upper respiratory tract of cigarette smokers. *PloS one*, 5(12), 1-10.

Charlson ES, Bittinger K, Haas AR, Fitzgerald AS, Frank I, Yadav A, et al., (2011). Topographical continuity of bacterial populations in the healthy human respiratory tract. *American journal of respiratory and critical care medicine*, 184(8), 957–63.

Cho I & Blaser MJ. (2012). The human microbiome: at the interface of health and disease. *Nature reviews. Genetics*, 13(4), 260–270.

Clayton, RA, Sutton G, Hinkle PS, et al., (1995). Intraspecific Variation in Small-Subunit rRNA Sequences in GenBank: Why Single Sequences May Not Adequately Represent Prokaryotic Taxa. *International Journal of Systematic Bacteriology*, 595–599.

Conlon MA & Bird AR. (2015). The impact of diet and lifestyle on gut microbiota and human health. *Nutrients*, 7(1), 17–44.

Costello EK, Stagaman K, Dethlefsen L, et al., (2012). The application of ecological theory toward an understanding of the human microbiome. *Science,* 336(6086), 1255–1262.

Csardi G & Nepusz T. (2006). The igraph software package for complex network research, *InterJournal, Complex Systems* 1695.

D'Amore R, Ijaz UZ, Schirmer M, et al., (2016). A comprehensive benchmarking study of protocols and , sequencing platforms for 16S rRNA community profiling. *BMC Genomics*, 17(55), 1-20.

David, LA, Maurice CF, Carmody RN, et al., (2014). Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, 505(7484), 559–563.

David LA, Materna AC, Friedman J, Campos-Baptista MI, Blackburn MC, Perrotta A, et al., (2014). Host lifestyle affects human microbiota on daily timescales. *Genome biology*, 15(89), 1-15.

Davidson AL, et al., (2008). Structure , Function , and Evolution of Bacterial ATP-Binding Cassette Systems. *Microbiology & Molecular Biology Reviews*, 72(2), 317–364.

Díez-Aguilar M, Ruiz-Garbajosa P, Fernandez-Olmos A, et al., (2013). Non-diphtheriae Corynebacterium species: an emerging respiratory pathogen. *European journal of clinical microbiology & infectious diseases* 32(6), 769–772.

Davidson AL, Dassa E, Orelle C, & Chen J. (2008). Structure , Function , and Evolution of Bacterial ATP-Binding Cassette Systems. *Microbiology and Molecular Biology Reviews,* 72(2), 317–364.

Distrutti E, Monaldi L, et al., (2016). Gut microbiota role in irritable bowel syndrome: New therapeutic strategies. *World journal of gastroenterology*, 22(7), 2219–2241.

Donohue I, Hillebrand H, Montoya JM, et al., (2016). Navigating the complexity of ecological stability. *Ecology Letters*, 19, 1172–1185.

Edgar RC. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature methods*, 10(10), 996–998.

Eisenstein M. (2012). The battle for sequencing supremacy. *Nature biotechnology*, 30(11), 1023–1026.

Erb-Downward JR, Dickson RP, Martinez FJ & Huffnagle GB. (2011). Analysis of the lung microbiome in the "healthy" smoker and in COPD. *PloS one*, 6(2), 1-12.

Erena MA, et al., (2013). Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods in Ecology and Evolution*, 4, 1111–1119.

Ferkol T & Schraufnagel D. (2014). The Global Burden of Respiratory Disease. *Annals of the American Thoracic Society*, 11(3), 404–406.

Ferretti JJ, McShan WM, Ajdic D, Savic DJ, et al., (2001). Complete genome sequence of an M1 strain of Streptococcus pyogenes. *Proceedings of the National Academy of Sciences of the United States of America*, 98(8), 4658–4663.

Fierer N, Ferrenberg S, Flores GE, González A, Kueneman J, Legg T, et al., (2012). From Animalcules to an Ecosystem: Application of Ecological Concepts to the Human Microbiome. *Annual Review of Ecology, Evolution, and Systematics*, 43(1), 137–155.

Fierer N & Lennon JT. (2011). The generation and maintenance of diversity in microbial communities. *American journal of botany*, 98(3), 439–48.

Gao Z, Kang Y, Yu J & Ren L. (2014). Human pharyngeal microbiome may play a protective role in respiratory tract infections. *Genomics, proteomics & bioinformatics*, 12(3), 144–150.

Gao Z, Tseng C, Strober BE, et al., (2008). Substantial Alterations of the Cutaneous Bacterial Biota in Psoriatic Lesions. *PLoS ONE*, 3(7), 1-9.

Gaspar JM & Thomas WK. (2013). Assessing the Consequences of Denoising Marker-Based Metagenomic Data. *PLoS ONE*, 8(3), 1-14.

Gazi H, et al., (2004). Oropharyngeal carriage and penicillin resistance of Neisseria meningitidis in primary school children in Manisa, Turkey. *Annals of the Academy of Medicine Singapore*, 33, 758–762.

Gerritsen J, Smidt H & Rijkers GT. (2011). Intestinal microbiota in human health and disease: The impact of probiotics. *Genes and Nutrition*, 6, 209–240.

Gevers D, Kugathasan S, Denson LA, et al., (2014). The treatment-naive microbiome in new-onset Crohn’s disease. *Cell host & microbe*, 15(3), 382-392.

Gong H, Shi Y, Zhou L, et al., (2014). Microbiota in the Throat and Risk Factors for Laryngeal Carcinoma. *Applied and Environmental Microbiology*, 80(23), 7356–7363.

Hodkinson BP & Grice EA. (2015). Next-Generation Sequencing: A Review of Technologies and Tools for Wound Microbiome Research. *Advances in wound care*, 4(1), 50–58.

Hofstra JJ, Matamoros S, van de Pol MA, et al., (2015). Changes in microbiota during experimental human Rhinovirus infection. *BMC infectious diseases*, 15(336), 1-9.

Hooper LV. (2001). Commensal Host-Bacterial Relationships in the Gut. *Science*, 292(5519), 1115-1118.

Huse SM, Ye Y, Zhou Y, et al., (2012). A core human microbiome as viewed through 16S rRNA sequence clusters. *PloS one*, 7(6), 1-12.

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409, 860-921.

Jakobsson HE, Jernberg C, Andersson AF, Sjölund-Karlsson M, Jansson JK, Engstrand L. (2010). Short-term antibiotic treatment has differing long-term impacts on the human throat and gut microbiome. *PloS one*, 5(3), 1-12.

Janda, JM & Abbott SL. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of clinical microbiology*, 45(9), 2761–2764.

Jensen A, Kristensen, LH & Prag J. (2007). Detection of Fusobacterium necrophorum subsp. funduliforme in tonsillitis in young adults by real-time PCR. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, 13(7), 695–701.

Jones WJ. (2009). High-Throughput Sequencing and Metagenomics. *Estuaries and Coasts*, 33(4), 944–952.

Joshi NA, Fass JN. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at https://github.com/najoshi/sickle.

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

Kersey CM, Agyemang PA & Dumenyo CK. (2012). CorA, the magnesium/nickel/cobalt transporter, affects virulence and extracellular enzyme production in the soft rot pathogen Pectobacterium carotovorum. *Molecular plant pathology*, 13(1), 58–71.

Khan MT, et al., (2012). The gut anaerobe Faecalibacterium prausnitzii uses an extracellular electron shuttle to grow at oxic-anoxic interphases. *The ISME journal*, 6(8), 1578–1585.

Konopka A. (2009). What is microbial community ecology? *The ISME journal*, 3(11), 1223-1230.

Kostic AD, Howitt MR & Garrett WS. (2013). Exploring host – microbiota interactions in animal models and humans. *Genes & Development*, 27, 701-718.

Kreth J, Merritt J & Qi F. (2009). Bacterial and host interactions of oral streptococci. *DNA and cell biology*, 28(8), 397–403.

Kristich CJ & Rice LB. (2009). Enterococcal Infection — Treatment and Antibiotic Resistance. *Enterococci,* 1-48.

Langdon A, Crook N & Dantas G. (2016). The effects of antibiotics on the microbiome throughout development and alternative approaches for therapeutic modulation. *Genome medicine*, 8(39), 1-16.

Lazarevic V, Whiteson K, Huse S, et al., (2009). Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. *Journal of microbiological methods*, 79(3), 266–271.

Legendre P & De Cáceres M. (2013). Beta diversity as the variance of community data: dissimilarity coefficients and partitioning. *Ecology letters*, 16(8), 951–963.

Leigh MB, Taylor L & Neufeld JD. (2010). Clone libraries. *Handbook of Hydrocarbon and Lipid Microbiology*, 3971-3985.

Lemon KP, Klepac-Ceraj V, Schiffer HK, Brodie EL, Lynch SV, Kolter R. (2010). Comparative Analyses of the Bacterial Microbiota of the Human Nostril and Oropharynx. 1(3), 4–6.

Leung RK, Zhou J, Guan W, et al., (2013). Modulation of potential respiratory pathogens by pH1N1 viral infection. *Clinical Microbiology and Infection*, 19(10), 930–935.

Li G. (2008). Mechanisms and functions of DNA mismatch repair. *Cell research,* 1(18), 85–98.

Li L & Ma Z. (2016). Testing the Neutral Theory of Biodiversity with Human Microbiome Datasets. *Scientific Reports*, 6(31448), 1-10.

Lloyd-Price J, Abu-Ali G & Huttenhower C. (2016). The healthy human microbiome. *Genome Medicine*, 8(51), 1-11.

Loreau M & de Mazancourt C. (2013). Biodiversity and ecosystem stability: a synthesis of underlying mechanisms. *Ecology letters*, 1-10.

Love MI, Huber W & Anders S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(550), 1-21.

Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R. (2012). Diversity, stability and resilience of the human gut microbiota. *Nature*, 489(7415), 220–230.

Matsuo K & Palmer JB. (2009). Anatomy and physiology of feeding and swallowing - normal and abnormal. *Phys Med rehabil Clin N Am.* 19(4), 691–707.

Marrs F. (2016). The type 4 pilin of Moraxella nonliquefaciens exhibits unique similarities with the pilins of Neisseria gonorrhoeae and Dichelobacter (Bacteroides) Journal of GeneralMicrobiology (1991), 2483–2490.

Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD. (2012). PANDAseq: paired-end assembler for illumina sequences. *BMC bioinformatics*, 13(31), 1-7.

McCann, KS. (2000). The diversity-stability debate. *Nature*, 405(6783), 228–33.

McMurdie PJ & Holmes S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PloS one*, 8(4), 1-11.

McMurdie PJ & Holmes S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS computational biology*, 10(4), 1-12.

Meinicke P. (2015). UProC: tools for ultra-fast protein domain classification. *Bioinformatics (Oxford, England)*, 31(9), 1382–8.

Mitchell J. (2011). Streptococcus mitis: walking the line between commensalism and pathogenesis. *Molecular oral microbiology*, 26(2), 89–98.

Moghaddam B & Javitt D. (2012). From revolution to evolution: the glutamate hypothesis of schizophrenia and its implication for treatment. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology*, 37(1), 4–15.

Munson GP, Lam DH, Outten FW, et al., (2000). Identification of a Copper-Responsive Two-Component System on the Chromosome of Escherichia coli K-12. *Journal of Bacteriology*, 182(20), 5864-5871.

Nelson KE, Fleischmann RD, DeBoy RT, et al., (2003). Complete Genome Sequence of the Oral Pathogenic Bacterium Porphyromonas gingivalis Strain W83. *Journal of Bacteriology*, 185(18), 5591–5601.

Nikolenko SI, Korobeynikov AI & Alekseyev MA. (2013). BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC genomics*, 14, 1-7.

Oh J, Byrd AL, Park M, et al., (2016). Temporal Stability of the Human Skin Microbiome Article Temporal Stability of the Human Skin Microbiome. *Cell*, 165, 854–866.

Oksanen J. 2013. Multivariate Analysis of Ecological Communities in R : vegan tutorial, 1-43.

Pace NR. (1997). A Molecular View of Microbial Diversity and the Biosphere. *Science*, 276(5313), 734–740.

Pao SS, Paulsen IANT & Saier MH. (1998). Major Facilitator Superfamily. *Microbiology and Molecular Biology Reviews*, 62(1), 1–34.

Park H, Shin JW, Park, SG, et al., (2014). Microbial communities in the upper respiratory tract of patients with asthma and chronic obstructive pulmonary disease. *PloS one*, 9(10), 1-11.

Pelton SI. (2012). Regulation of bacterial trafficking in the nasopharynx. *Paediatric respiratory reviews*, 13(3), 150–153.

Petrosino JF, Highlander S, Luna RA, et al., (2009). Metagenomic pyrosequencing and microbial identification. *Clinical chemistry*, 55(5), 856–866.

Price MN, Dehal P & Arkin AP. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. *PloS one*, 5(3), 1-10.

Purta E, O'Connor M, Bujnicki JM & Douthwaite S. (2008). YccW is the m 5 C Methyltransferase Specific for 23S rRNA Nucleotide 1962. *JMB*, 383, 641–651.

Quince C, Lanzen A, Davenport RJ & Turnbaugh PJ. (2011). Removing noise from pyrosequenced amplicons. *BMC bioinformatics*, 12(38), 1-40.

Ravel J, Brotman RM, Gajer P & Ma B. (2013). Daily temporal dynamics of vaginal microbiota before, during and after episodes of bacterial vaginosis. *Microbiome*, 1(29), 1-6.

Relman DA & Falkow S. (2001). The meaning and impact of the human genome sequence for microbiology. *Trends in Microbiology*, 9(5), 206–208.

Richardson DM & Pysek P. (2007). Elton, C.S. 1958: The ecology of invasions by animals and plants. London: Methuen. *Progress in Physical Geography*, 31(6), 659–666.

Robinson CJ, Bohannan BJM & Young VB. (2010). From structure to function: the ecology of host-associated microbial communities. *Microbiology and molecular biology reviews : MMBR*, 74(3), 453–76.

Rosindell J, et al., (2012). The case for ecological neutral theory. *Trends in Ecology and Evolution*, 27(4), 203–208.

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

Sanger, F. & Nicklen, S., 1977. DNA sequencing with chain-terminating. *PNAS*, 74(12), 5463–5467.

Santagati M, Scillato M, Patane F, et al., (2012). Bacteriocin-producing oral streptococci and inhibition of respiratory pathogens. *FEMS immunology and medical microbiology*, 65(1), 23–31.

Santiago A, Panda S, Khader IE, et al., (2014). Short-Term Effect of Antibiotics on Human Gut Microbiota. *PLOS one*, 9(4), 1-7.

Saraswati S & Sitaraman R (2014). Aging and the human gut microbiota-from correlation to causality. *Frontiers in microbiology*, 5(764), 1-4.

Schirmer M, Ijaz UZ, D'Amore R, et al., (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research*, 43(6), 1-16.

Schloss PD & Westcott SL. (2011). Assessing and Improving Methods Used in Operational Taxonomic Unit-Based Approaches for 16S rRNA Gene Sequence Analysis. *Applied and Environmental Microbiology*, 77(10), 3219–3226.

Segata N, Haake SK, Mannon P, Lemon KP, Waldron L, Gevers D, et al., (2012). Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome biology*, 13(6), 1-18.

Shreiner AB, Kao JY & Young VB. (2016). The gut microbiome in health and in disease. *Curr Opin Gastroenterol.*, 31(1), 69–75.

Skaar EP, Lazio MP & Seifert HS. (2002). Roles of the recJ and recN Genes in Homologous Recombination and DNA Repair Pathways of Neisseria gonorrhoeae. *Journal of Bacteriology*, 184(4), 919–927.

Stenfors LE, Bye H & Räisänen S. (2003). Noticeable differences in bacterial defence on tonsillar surfaces between bacteria-induced and virus-induced acute tonsillitis. *International Journal of Pediatric Otorhinolaryngology*, 67(10), 1075–1082.

Sugahara H, Odamaki T, Fukada S, et al., (2015). Probiotic Bifidobacterium longum alters gut luminal metabolism through modification of the gut microbial community. *Scientific reports*, 5(13548), 1-11.

Tamashiro E, Cohen N, Palmer JN, et al., (2009). Effects of cigarette smoking on the respiratory epithelium and its role in the pathogenesis of chronic rhinosinusitis. *Brazilian Journal of Otorhinolaryngology*, 75(6), 903–907.

Tan D, Goerlitz S, Dumitrescu, RG, et al.,(2008). Associations between cigarette smoking and mitochondrial DNA abnormalities in buccal cells. *Carcinogenesis*, 29(6), 1170–1177.

Tettelin H, Masignani V, Cieslewicz MJ, et al., (2002). Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V Streptococcus agalactiae. *Proceedings of the National Academy of Sciences of the United States of America*, 99(19), 12391–12396.

The Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402), 207–214.

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE *et al.* (2009). A core gut microbiome in obese and lean twins. *Nature*, 457(7228), 480–484.

Turnbaugh PJ & Gordon JI. (2009). The core gut microbiome, energy balance and obesity. *The Journal of physiology*, 587(17), 4153–4158.

Ursell LK, Metcalf JL, Parfrey LW, Knight R. (2012). Defining the human microbiome. *Nutrition reviews*, 70, 38–44.

Vincent AT, Derome N, Boyle B, et al., (2016). Next-generation sequencing (NGS) in the microbiological world: How to make the most of your money. *Journal of Microbiological Methods*, 16, 30031-30038.

Wang Q, Garrity GM, Tiedje JM, Cole JR. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology*, 73(16), 5261–5267.

Wang, QQ, Zhang CF, Chu CG & Zhu ZF. (2012). Prevalence of Enterococcus faecalis in saliva and filled root canals of teeth associated with apical periodontitis. *International journal of oral science*, 4(1), 19–23.

Warren L, Castellarin M, Freeman JD, et al., (2012). Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Research,* 299–306.

Whelan, FJ, Verschoor CP, Stearns JC, et al., (2014). The loss of topography in the microbial communities of the upper respiratory tract in the elderly. *Annals of the American Thoracic Society*, 11(4), 513-521.

Williams RJ, Howe A & Hofmockel KS. (2014). Demonstrating microbial co-occurrence pattern analyses within and between ecosystems. *Frontiers in microbiology*, 5(358), 1-10.

Willing BP, Dicksved J, Halfvarson J, et al., (2010). A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes. *Gastroenterology*, 139(6), 1844–1854.

Woese CR. (1987). Bacterial Evolution Background. *Microbiologival Reviews*, 51(2), 221–271.

Wong JMW, Souza R, Kendall CWC, et al., (2006). Colonic Health : Fermentation and Short Chain Fatty Acids. *J Clin Gastroenterology*, 40(3), 235–243.

Wu J, Peters B, Dominianni C, et al., (2016). Cigarette smoking and the oral microbiome in a large study of American adults. *The ISME journal*, 10, 2435-2446.

Yi H, Yong D, Lee K, et al., (2014). Profiling bacterial community in upper respiratory tracts. *BMC infectious diseases*, 14(583), 1-9.

Zeller I, Hutcherson JA, Lamont RJ, et al., (2014). Altered antigenic profiling and infectivity of Porphyromonas gingivalis in smokers and non-smokers with periodontitis. *Journal of periodontology*, 85(6), 837–844.

Van Zyl-Smit RN, Binder A, Meldau R, et al., (2014). Cigarette smoke impairs cytokine responses and BCG containment in alveolar macrophages. *Thorax*, 69(4), 363–370.