University of Glasgow

Busby, Joseph (2023) *Using shotgun metagenomics to explore the effects of HIV infection on the longitudinal nasopharyngeal microbiome of Malawian adults.* MSc(R) thesis.

https://theses.gla.ac.uk/83430/

Copyright and moral rights for this work are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This work cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Enlighten: Theses
https://theses.gla.ac.uk/
research-enlighten@glasgow.ac.uk

Using shotgun metagenomics to explore the effects of HIV infection on the longitudinal nasopharyngeal microbiome of Malawian adults

Joseph Busby

Submitted in fulfilment of the requirements for the
Degree of Master of Science by Research

School of Infection and Immunity
College of Medical, Veterinary and Life Sciences
University of Glasgow

September 2022

Abstract

People living with HIV are more likely to suffer from respiratory tract infections, including bacterial pneumonia, which can occur either on its own or as a secondary infection following respiratory viral infections. The nasopharynx is home to a microbial community collectively referred to as the nasopharyngeal microbiome (NPM). Many potential pathogens are carried asymptomatically in the nasopharynx in a proportion of the population, but their overgrowth and spread to the lungs can cause pneumonia. Therefore, the NPM can either repel pathogens or act as a reservoir for them to proliferate and spread to the lungs. We proposed that there may be differences in the NPM of HIV-infected and HIV-uninfected individuals, and that this might contribute to the increased risk of bacterial pneumonia associated with HIV infection.

To characterise the NPM, we carried out DNA and RNA shotgun sequencing on nasopharyngeal swab samples and processed them through a custom metagenomics pipeline that calculated the relative abundance of microbial species in each sample. We analysed the NPM of 10 HIV-infected individuals (cases) and 6 HIV uninfected-individuals (controls) at 3 timepoints: baseline, 1 month and 9 months. This study represented the first attempt to study the NPM in HIV-infected individuals.

Analysis with PERMANOVA showed that on average, NPM composition was significantly different depending on HIV status. We found that the relative abundance of the genera *Cutibacterium and Pahexavirus* were significantly lower in cases than controls, and that this was driven by bacteriophages from the *Pahexavirus* genus and their bacterial host *Cutibacterium acnes*. Diversity analysis identified that the NPM of cases was less stable over time, had lower viral richness and higher bacterial evenness. These diversity measurements could all be at least partially attributed to differing abundance of *Pahexavirus* and *Cutibacterium* between cases and controls. We proposed that *C. acnes* might act as an immunomodulator in the NPM, however this would require further study to confirm.

This study also served as a proof of concept of using shotgun metagenomics to profile the NPM, where the microbial community is much less dense than in the gut. Contaminant human sequences were a major issue for the DNA and RNA datasets that limited the scope of our analysis, even when steps were taken to deplete human sequences prior to sequencing. The compositional nature of sequence data also caused issues for analysis. The detection and quantification of relationships between microbes can be improved by quantifying the absolute abundances of microbes, instead of using their relative abundances. We proposed steps that future studies could take to further reduce human contamination and to quantify the absolute abundances of microbial species in their samples.

Table of Contents

List of Tables

List of Figures

Acknowledgements

For their continued support throughout my studies, I extend the utmost thanks to my supervisors Antonia Ho and David L. Robertson. Without their guidance and expertise this research would not have been possible. I am also very grateful to the University of Glasgow and the MRC whose funding for research and living expenses facilitated this work. I'd also like to give my appreciation to the support staff at the CVR, in particular to Fiona Graham who has always gone above and beyond in providing an exceptional level of care and support.

A special thank you to my partner Marija who has always supported me and puts up with my rants day in, day out. An additional thank you to my friends Anna, Fran, and the late Ben Stamp, who all began my research journey with me and helped shape me into who I am today.

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

Abbreviations

µl	Microlitres
16s rRNA	16s ribosomal RNA
18s rRNA	18s ribosomal RNA
AIDS	Acquired immunodeficiency syndrome
AKP	Anna Karenina Principle
ART	Antiretroviral therapy
BAL	Bronchoalveolar lavage
BCD	Bray-Curtis dissimilarity
CD4	Cluster of differentiation 4 positive T-cells
CD8	Cluster of differentiation 8 positive T-cells
CMV	Human cytomegalovirus
COPD	Chronic obstructive pulmonary disease
DNA	Deoxyribonucleic acid
DNAse	Deoxyribonuclease
g	G-Force
HAP	Household air pollution
HIV	Human Immunodeficiency Virus
IBD	Inflammatory bowel disease
ILI	Influenza-like illness
LCA	Least-common ancestor
LHMP	Lung HIV Microbiome Project
LRT	Lower respiratory tract
LRTI	Lower respiratory tract infection
ml	Millilitres
mRNA	Messenger RNA
NGS	Next-generation sequencing
NPM	Nasopharyngeal microbiome
nr/nt	NCBI non-redundant nucleotide database
OLS	Ordinary least-square
OTU	Operational taxonomic unit
PCoA	Principal coordinates analysis
PCR	Polymerase chain reaction
PERMANOVA	Permutational Multivariate Analysis of Variance
PERMDISP	Permutational analysis of multivariate dispersions
PLHIV	People living with HIV
RNA	Ribonucleic acid
RNAse	Ribonuclease

rRNA	Ribosomal RNA
SCFA	Short chain fatty acids
TB	Tuberculosis
TLR5	Toll-like receptor 5
URT	Upper respiratory tract
URTI	Upper respiratory tract infection

1 Introduction

1.1 An introduction to the microbiome

Our bodies are colonised by trillions of microorganisms: bacteria, viruses, fungi, archaea, and protists that live on almost every surface of the body including the skin, gut and respiratory tract. Collectively these organisms are referred to as the microbiome/microbiota and their set of genomes, the metagenome. There is an increasing body of data linking the human microbiome to health and disease (Man, de Steenhuijsen Piters and Bogaert, 2017; Wang *et al.*, 2017; Gilbert *et al.*, 2018). In health, the microbiome contributes to the normal physiological function of the host. It can provide supplementary metabolic functions to the host, is essential for proper immune system development, and can resist the colonisation of host tissues by pathogens (Barr *et al.*, 2013; Buffie and Pamer, 2013; Rooks and Garrett, 2016; Oliphant and Allen-Vercoe, 2019). Alterations in the species composition of the microbiota can result from a myriad of host and environmental factors including but not limited to: age, sex, diet, pollution, host genetics, disease, antibiotics and drugs (Frank *et al.*, 2007; Turnbaugh *et al.*, 2008; Kau *et al.*, 2011; Yatsunenko *et al.*, 2012; Jašarević, Morrison and Bale, 2016; Langdon, Crook and Dantas, 2016; Bailey *et al.*, 2020; Weersma, Zhernakova and Fu, 2020). Microbiota alterations can result in dysbiosis, broadly defined as an imbalance between neutral/beneficial and harmful/pathogenic organisms (DeGruttola *et al.*, 2016).

The gut microbiota play a key role in host metabolism by providing additional metabolic pathways that are not encoded within the host genome. For example, the human gut microbiota to degrade otherwise indigestible plant polysaccharides, producing short-chain fatty acids (SCFAs) which can be metabolised by the host increased energy extraction from the diet (Gill *et al.*, 2006). The gut microbiota also contains the relevant metabolic pathways to synthesise vitamins and essential amino acids that humans are unable to produce themselves (Gill *et al.*, 2006). Studies in both mice and humans have shown that the diet of an individual influences the composition of the gut microbiota (Turnbaugh *et al.*, 2009; Ussar *et al.*, 2015). Differences in the microbiota of lean and obese individuals have also been reported (Turnbaugh *et al.*, 2006).

The microbiota of obese mice displays increased energy extraction from the diet, demonstrated by showing that germ-free mice displayed additional weight gain and adipose tissue content when receiving a microbiota transplant from obese mice compared lean mice (Turnbaugh *et al.*, 2006; Ridaura *et al.*, 2013). The microbiota also plays an essential role in normal immune development through providing molecular cues, in the form of metabolites and surface antigens, that are required for the fine tuning of the immune response (Rooks and Garrett, 2016). Germ free animals display severe immune defects and higher susceptibility to infection (Rooks and Garrett, 2016). A similar effect has also been observed in antibiotic-treated mice which display an impaired antiviral response to Influenza A infection without the activation of immune receptor TLR5 by commensal bacteria (Pfeiffer and Virgin, 2016). The relationship between the immune system and the microbiota is dynamic and multifactorial. Dietary, pathogen and commensal-derived antigens constantly interact with host receptors and complex molecular pathways integrate these signals to form a coordinated host response that tolerates commensals and excludes pathogens (Zhang *et al.*, 2020) (**Figure 1-1**). Regulating appropriate responses to this enormous array of antigens represents a significant challenge and failure to do so can lead to infection, allergy, metabolic syndromes or inflammation (Belkaid and Hand, 2014). An example of this can be seen in patients with Crohn's disease, whose gut microbiota show increased populations of inflammatory commensals, such as adherent-invasive *Escherichia coli*, and decreases in bacteria belonging to class *Clostridia*, which produce anti-inflammatory SCFAs (Belkaid and Hand, 2014). It has been proposed that dysregulated immune responses lead to the proliferation of pro-inflammatory commensals that are adapted to survive under immune stress, forming a positive feedback loop whereby these commensals invoke further inflammation and microbiome alteration (Belkaid and Hand, 2014).

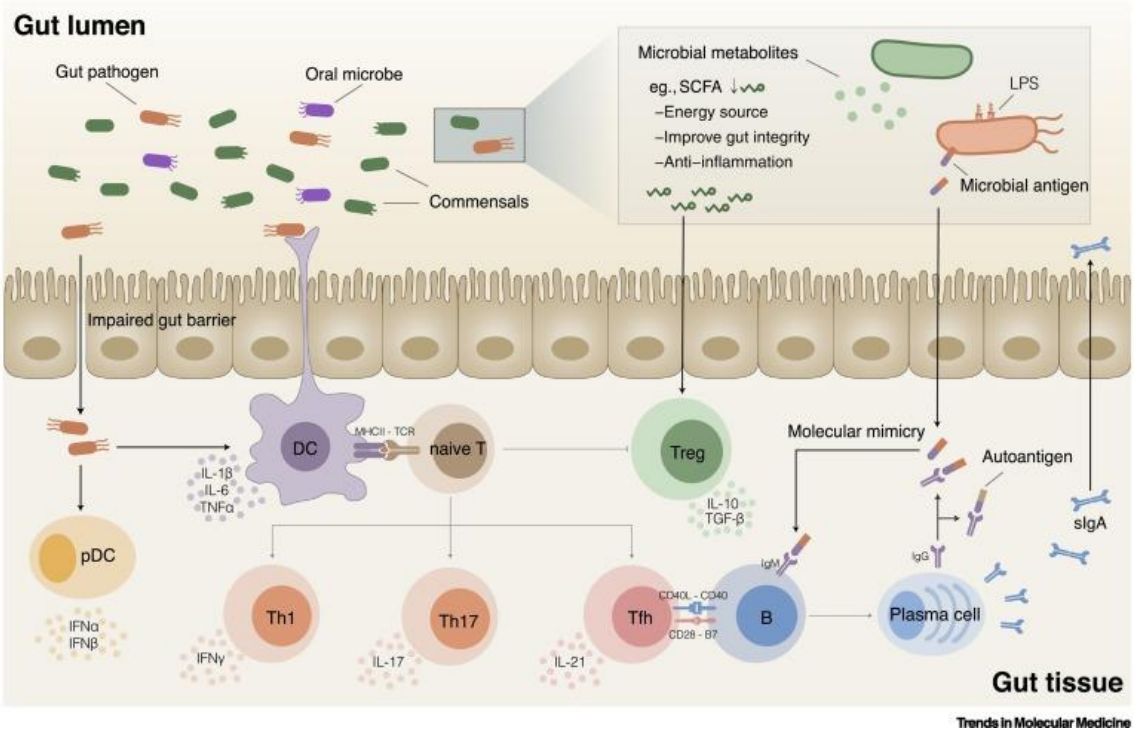

Figure 1-1 An overview of the interplay between the microbiome and host immunity. Reprinted (with permission) from Zhang X, Chen B di, Zhao L dan, Li H. The Gut Microbiota: Emerging Evidence in Autoimmune Diseases. Trends Mol Med. 2020 Sep 1;26(9):862–73.

The presence of the microbiota protects hosts from colonisation by pathogens, this activity has been termed colonisation resistance (Van der Waaij, Berghuis-de Vries and Lekkerkerk-van der Wees, 1971). Part of this effect occurs indirectly because of immunomodulation by the microbiota, termed immune-mediated colonisation resistance, that enhances the ability of the immune system to destroy pathogenic invaders. Colonisation resistance also results from direct interactions between members of the microbiota and invaders, including competition for nutrients and the production of inhibitory molecules that target invaders (Buffie and Pamer, 2013). Bacteriophage also contribute to colonisation resistance by adhering to host mucosal surfaces and protecting underlying host cells from bacterial infection (Barr *et al.*, 2013). Changes in the microbiome of an individual can alter the colonisation resistance of a host. For example; antibiotic treatment results in reduced diversity of the gut microbiota, leading to an increase in free metabolites such as sialic acid or lactate that can enable the invasion of the pathogens *Clostridium difficile* and *Salmonella Typhimurium* (Ng *et al.*, 2013).

1.2 The respiratory microbiome: structure, function and roles in health and disease

Most microbiome research to date has focussed on the gut. However, there is a growing appreciation that the respiratory tract microbiota are highly relevant to human health and disease, particularly in the context of respiratory infections and inflammatory disorders, such as chronic obstructive pulmonary disease (COPD) and asthma (Dickson *et al.*, 2016; Schenck, Surette and Bowdish, 2016; Cleary and Clarke, 2017; Krause *et al.*, 2017; Man, de Steenhuijsen Piters and Bogaert, 2017; Zhou *et al.*, 2019). The respiratory tract is a complex organ system that contains multiple anatomical structures which each have their own role in facilitating respiration through gas-exchange of oxygen and carbon dioxide. Broadly speaking, the respiratory tract can be broken down into two sections, upper respiratory tract (URT) and lower respiratory tract (LRT). The URT contains the nasal cavity, nasopharynx and oropharynx and is primarily responsible for the inhalation, humidification of air and the filtering of microbes. The LRT consists of the trachea and the lungs, the trachea transports air between the URT and the lungs where gas exchange takes place. Each of these sites has a distinct microbiota that varies in both density and composition due to their unique physiology and environmental conditions including temperature, pH and oxygen content (**Figure 1-2**).

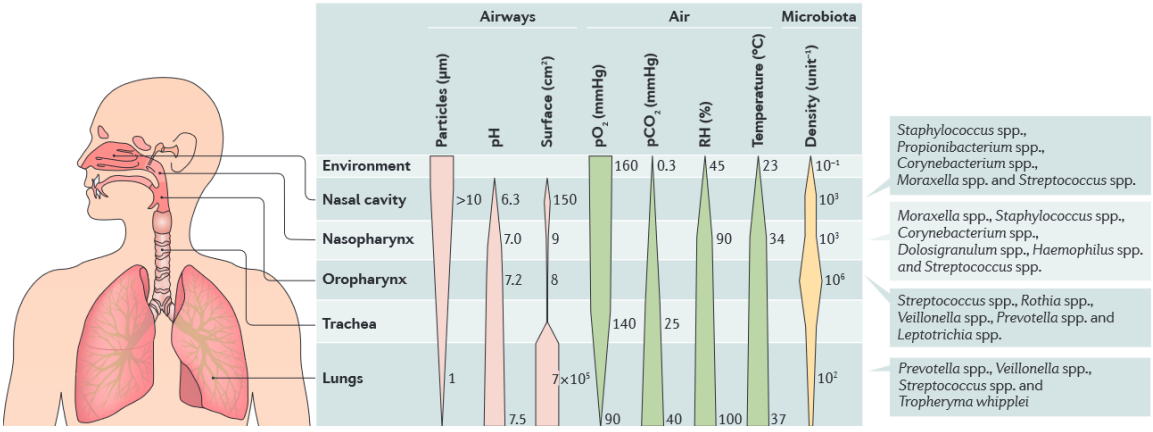

Figure 1-2 An overview of the anatomy of the respiratory tract, as well as the physiological and microbial differences that exist between separate sites. Reprinted (with permission) from Man WH, de Steenhuijsen Piters WAA, Bogaert D. The microbiota of the respiratory tract: gatekeeper to respiratory health. Nat Rev Microbiol. 2017 Mar 20;15(5):259–70.

Although these sites are physically remote from one another, a significant level of microbial migration occurs between them, microbes pass through the URT and can enter the lung via the inhalation of air, dispersal along mucous membranes in nasal secretions and saliva, or through the inhalation of oropharyngeal contents (microaspiration) (Dickson *et al.*, 2015; Schenck, Surette and Bowdish, 2016). Microbial clearance in the lungs transports bacteria back to the URT, clearance occurs through coughing and via the mucociliary escalator, where microbes are trapped in mucous are transported upwards by the beating of cilia (Dickson *et al.*, 2016). Migration of microbes is a key determinant of the respiratory microbiome, the microbial composition of the lung microbiome in adults has been found to be primarily determined by the migration and elimination of microbes from the oropharynx, rather than expansion of microbial populations inside the lung (Bassis *et al.*, 2015; Dickson *et al.*, 2015).

In contrast to the oropharynx, the nasopharyngeal microbiome (NPM) is not a major determinant of the lung microbiome in health (Bassis *et al.*, 2015). However it can act as an important reservoir for populations of potential pathogens to expand and spread to infect the lungs (Man, de Steenhuijsen Piters and Bogaert, 2017). The most common and abundant genera found in the adult NPM to date are *Staphylococcus*, *Haemophilus*, *Streptococcus*, *Sphingobacterium*, *Prevotella*, *Bifidobacterium*, *Rothia*, *Propionibacterium*, *Dolosigranulum*, *Corynebacterium* and *Moraxella* (Cremers *et al.*, 2014; Stearns *et al.*, 2015). Not all of these genera will necessary be present at once. There are numerous host and environmental factors that affect NPM composition and multiple distinct metagenomic profiles have already been described with varying proportions of these genera (Cremers *et al.*, 2014). Key pathogenic species that can be found in the NPM include *Staphylococcus aureus*, *Streptococcus pneumoniae*, *Haemophilus influenzae*, *Moraxella catarrhalis* and *Neisseria meningitidis* (Schenck, Surette and Bowdish, 2016; Cleary and Clarke, 2017; Man, de Steenhuijsen Piters and Bogaert, 2017; Hanada *et al.*, 2018). These species are known as pathobionts or potential pathogens, because they are carried asymptomatically in a proportion of the population but are capable of acutely infecting their host if the conditions are right, such as if host immunity wanes or a perturbation to the microbiome allowed them to overgrow (Schenck, Surette and Bowdish, 2016; Cleary and Clarke, 2017; Hanada *et al.*, 2018).

Ecological interactions between resident species of the NPM can be important factors in determining whether populations of potential pathogens can be prevented from establishing or maintained at safe levels (colonisation resistance), or whether they can overgrow and cause infections. Pathogenic species can engage in co-operative behaviour with one another, *S. pneumoniae*, *H. influenzae* and *M. catarrhalis* use shared quorum-sensing systems to create polyspecies biofilms that increase their resistance to antibiotics (Armbruster *et al.*, 2010; Perez *et al.*, 2014). Pathogens can also exclude one another, S. *pneumoniae* excludes *S. aureus* through the production of hydrogen-peroxide which trigger lysis of *S. aureus* through prophage-induction (Selva *et al.*, 2009). Resident commensals can protect the host from pathobionts, *Corynebacterium* and *Dolosigranulum* species have been found to engage in competition and exclusion of *S. aureus* in the nasal microbiome (Yan *et al.*, 2013; Liu *et al.*, 2015). Further study of the ecological interactions that are beneficial to commensals or restrictive to pathogens species is an exciting avenue for future research and could enable the use of exciting new microbiome interventions such as probiotics or phage therapy treatments to maintain microbiome health (Cleary and Clarke, 2017).

In health, the microbiota of the adult lung closely resembles that of the oropharynx, with the most abundant genera being *Prevotella*, *Veionella* and *Streptococcus*. Bacterial lower respiratory tract infections (LRTIs) occur when a bacterial pathogen is successful in colonising the URT and spreading to the lungs (Man, de Steenhuijsen Piters and Bogaert, 2017). Viral LRTI's may also originate as upper respiratory tract infections (URTIs) that spread to the lung, but some respiratory viruses can also directly bind to and infect the lung after being inhaled from the environment (van Riel *et al.*, 2006).

1.3 HIV infection and the respiratory tract

Approximately 38 million people worldwide are living with HIV and it is a leading cause of morbidity and mortality (Vos *et al.*, 2020; 'Global HIV & AIDS statistics — Fact sheet', 2022). Left untreated, HIV-infection causes the progressive loss of

$CD4^+$ T cells (CD4) over time and ultimately leads to acquired immunodeficiency syndrome (AIDS), defined when the CD4 count drops below 200 cells/μl blood (Deeks *et al.*, 2015). Antiretroviral therapy (ART) is a highly effective treatment that can suppress HIV replication and prevent CD4 depletion, however only 68% of current HIV cases worldwide are virally suppressed ('Global HIV & AIDS statistics — Fact sheet', 2022). Furthermore, it can take several years of ART for CD4 counts to recover and 33% of patients with low CD4 counts at the beginning of ART never fully returned to healthy CD4 levels after 7 years of treatment (Lok *et al.*, 2010). Therefore, even in the age of ART there are many people living with HIV (PLHIV) who are immunocompromised and at increased risk of infection.

Opportunistic infections are infections which rarely occur in immunocompetent individuals, a weakened immune system provides an opportunity for an infection that would otherwise be unsuccessful. HIV-infected individuals are at a higher risk of both common and opportunistic infections (Justiz Vaillant and Naik, 2022). Tuberculosis (TB) is an opportunistic respiratory infection, caused by the bacterium *Mycobacterium tuberculosis*, and is the leading cause of death among PLHIV (*WHO: Tuberculosis & HIV*, 2020). TB is extremely common in PLHIV in sub-Saharan Africa and can infect people even in the early stages of HIV; the risk of infection increases as CD4 count decreases below 500 cells/μl blood (Lawn *et al.*, 2009), although ART has been shown to significantly decrease TB risk (Lawn, Bekker and Wood, 2005). In Sub-Saharan Africa, bacterial pneumonia incidence is 10-20x higher in HIV-infected adults than the general population and influenza incidence is 3x higher (Feikin *et al.*, 2004; Ho *et al.*, 2018). A study of hospitalised adults in Malawi found 78.4% of hospitalised pneumonia cases occurred in HIV-infected patients (Aston *et al.*, 2019). HIV prevalence also drives the aetiology of pneumonia: pneumonia aetiology is dominated by influenza and rhinovirus in the US, whereas Mycobacterium tuberculosis and Streptococcus pneumoniae are the most common cause of pneumonia in sub-Saharan Africa (Scott *et al.*, 2000; Jain *et al.*, 2015; Aston *et al.*, 2019).

Few studies have looked at respiratory viral infections in PLHIV, however several relatively new studies have demonstrated a greater risk of influenza infection in HIV-infected individuals (Cohen *et al.*, 2013; Ho *et al.*, 2018). A common

complication of respiratory viral diseases, including influenza, are secondary bacterial infections that occur during or after the primary infection and cause pneumonia (Smith and McCullers, 2014). HIV-infected individuals are also at a greater risk of contracting secondary infections, HIV-positive patients hospitalised with influenza were found to be at a higher risk of pneumococcal secondary infection (Cohen *et al.*, 2013). Respiratory viral infections enable secondary infections either by strain/species specific interactions or more generally by damaging the respiratory epithelium which: breaches the physiological barrier to bacterial invasion, impairs bacterial clearance by cilia and exposes bacterial adhesion sites such as laminin and collagen (Smith and McCullers, 2014). Many species responsible for respiratory secondary infections, such as *Streptococcus pneumoniae* and *Staphylococcus aureus*, are present asymptomatically in the nasopharyngeal microbiome (NPM) but can infect the host given the right conditions (McCullers, 2014; Man, de Steenhuijsen Piters and Bogaert, 2017). If the NPM microbiome of HIV-infected individuals was different compared to HIV-uninfected individuals, for example by having higher carriage of pathobionts, this could contribute to the increased risk of pneumonia associated with HIV infection.

There are few studies comparing the respiratory microbiome between PLHIV and the general population and most have been carried out by the Lung HIV Microbiome Project (LHMP). The focus of the LHMP was on understanding how the HIV-infected lung microbiome might relate to non-infectious pulmonary conditions, such as chronic obstructive pulmonary disease (COPD), which occur more frequently in HIV-infected individuals, even those undergoing ART (Twigg, Weinstock and Knox, 2017). It was proposed that immunological defects in the lung associated with ART-treated or untreated HIV-infection, including CD4:CD8 T-cell imbalance and its associated inflammation, might lead microbiome changes (Twigg, Weinstock and Knox, 2017). In well HIV-infected adults with healthy CD4 counts $\geq$600 cells/µl, there were no major differences in the lung microbiome compared to healthy controls (Beck *et al.*, 2015; Cribbs *et al.*, 2016). A later study in US adults found that lung microbiome richness and evenness were lower in advanced stage HIV-infected individuals compared to healthy controls, meaning that there were fewer species total, and a small number of species tended to dominate the microbial population (Twigg *et al.*,

2016). Additionally, there was significantly greater dispersion of samples collected from HIV-infected individuals than those from healthy controls (ie. HIV-infected samples are more different to each other than HIV-uninfected samples) (Twigg *et al.*, 2016).

The increased dispersion seen in the HIV-infected lung microbiome is an example of the Anna Karenina Principle (AKP). The AKP applied to microbiomes states that when the host has a reduced ability to regulate its microbiome, changes in community composition become largely stochastic and can result in microbiomes from the same group being highly dispersed (Zaneveld, McMinds and Vega Thurber, 2017). Heterogeneity in the lung microbiome of HIV-infected individuals would complicate specific microbiome altering-treatments as there would be fewer targets (organisms) that are shared between individuals. Indeed, the researchers found that only *Streptococcus* was significantly more abundant in HIV-infected individuals compared to healthy controls (Twigg *et al.*, 2016). However, treatment with ART for 1 year reduced the observed differences in alpha richness, evenness and beta diversity dispersion between samples (Twigg *et al.*, 2016).

The researchers were able to find 16 taxa that were significantly more abundant in the 1-year ART group compared to healthy controls, including *Streptococcus*, *Veionella* and *Prevotella* species (Twigg *et al.*, 2016). *Prevotella* and *Veionella* commonly reside in the oral microbiome and enter the lung by microaspiration and might be important for the risk of COPD, as they have previously been linked to chronic lung inflammation via increased lung neutrophil and lymphocyte counts, as well as increased nitric oxide levels (Segal *et al.*, 2013). As previously discussed, *Streptococcus* species are clinically relevant to respiratory tract infections either on its own or as a secondary infection. A separate study into the lung mycobiome found that *Pneumocystis jirovecii and Ceriporia lacerata*, were overrepresented in HIV-infected adults undergoing ART and with relatively high CD4 counts (median 599 cells/µl) (Cui *et al.*, 2015). As these samples were collected from asymptomatic individuals, the carriage of these pathogens does not indicate an active infection, instead it represents a risk of opportunistic infection in the future.

1.4 Preparing and sequencing microbiome samples

Microbiome studies traditionally make use of molecular techniques and next-generation sequencing (NGS) technology to determine which microbes are present, and in what proportions, for a given sample.

The first stage of sample processing involves isolating the microbial DNA/RNA of interest from the rest of the sample, so that it can be sequenced. At this stage it is important to enrich your samples for the organisms of interest by filtering out organisms that aren't to be studied, such as host cells, prior to sequencing (Thurber *et al.*, 2009). Human genomes are very large compared to most microbial/viral genomes and can therefore make up a disproportionately large amount of sequencing data (Thurber *et al.*, 2009). This is particularly relevant for samples that have low microbial density, such as those taken from the respiratory tract (Nelson *et al.*, 2019). Centrifugation can be an effective method for filtering out specific types of cells based on density. For instance, centrifugation at 5000g can be used to pellet eukaryotic cells (Nelson *et al.*, 2019). Eukaryotic/human cells can also be lysed with detergents or chaotropic agents (Nelson *et al.*, 2019). Once a sample is enriched, nucleic acid extractions can be carried out. Depending on the goal of the study, researchers may choose to DNAse/RNAse treat their samples prior to nucleic acid extractions; this degrades extracellular nucleic acids, such as those from extracellular virions, dead microbes or host cells that have been purposefully lysed (Thurber *et al.*, 2009). Nucleic acid extraction is then carried out by mechanical/chemical lysis of the remaining cells, followed by the separation of nucleic acids from cellular debris using spin columns or magnetic beads that bind nucleic acids (Ali *et al.*, 2017). At this point it can be beneficial to further deplete host DNA using antibody depletion kits, which specifically bind methylated epitopes in eukaryotic DNA (Nelson *et al.*, 2019).

Once nucleic acids have been extracted, they can be prepared for sequencing. Microbiome studies usually follow one of two approaches: amplicon-based sequencing or shotgun sequencing. Amplicon-based sequencing involves the amplification and sequencing of DNA from specific marker genes. Common amplicons include the 16s ribosomal RNA (rRNA) gene in bacteria/archaea and

18s rRNA gene in fungi, these genes have common regions meaning that it can be targeted for PCR amplification, but also contains hypervariable regions that can be used to determine its phylogeny (Jovel *et al.*, 2016; Banos *et al.*, 2018). After sequencing the amplicons, the sequences can be clustered into operational taxonomic units (OTUs) based on sequence similarity, taxonomy is then assigned by matching OTU sequences against reference databases (Knight *et al.*, 2018). Shotgun sequencing involves the untargeted sequencing of DNA or RNA in a sample, as opposed to specific amplicons, these sequences can then be classified to create taxonomic profiles like in 16s rRNA sequencing. Shotgun sequencing can also be used to assemble metagenomes, analyse community-encoded functions and investigate transcriptional activity of the community through transcriptomics (Knight *et al.*, 2018), but these are beyond the scope of our study.

There are advantages and drawbacks to both sequencing approaches. 16s rRNA sequencing is cheaper than shotgun sequencing, has better established pipelines and is better at dealing with low biomass/host contamination issues (Knight *et al.*, 2018). However it is restricted to bacteria and archaea, taxonomic resolution is restricted to genus-level, and the sequences provides no information about the gene-encoded functions of the community (Knight *et al.*, 2018). Conversely, shotgun sequencing allows identification of taxa at the species level and enables researchers to profile entire microbial communities, not just bacteria (Jovel *et al.*, 2016). Although shotgun sequencing is more expensive, has less established protocols for downstream analysis, and can suffer from host contamination issues, we opted for this approach in our study as species-level classification is important for our study. We also need to be able to differentiate between harmless commensals and potential pathogens from the same genera. Additionally, we wanted to analyse species from every kingdom as there is a growing appreciation that non-bacterial components of the microbiome have a key role in human health (Norman, Handley and Virgin, 2014; Pfeiffer and Virgin, 2016; Quince *et al.*, 2017)

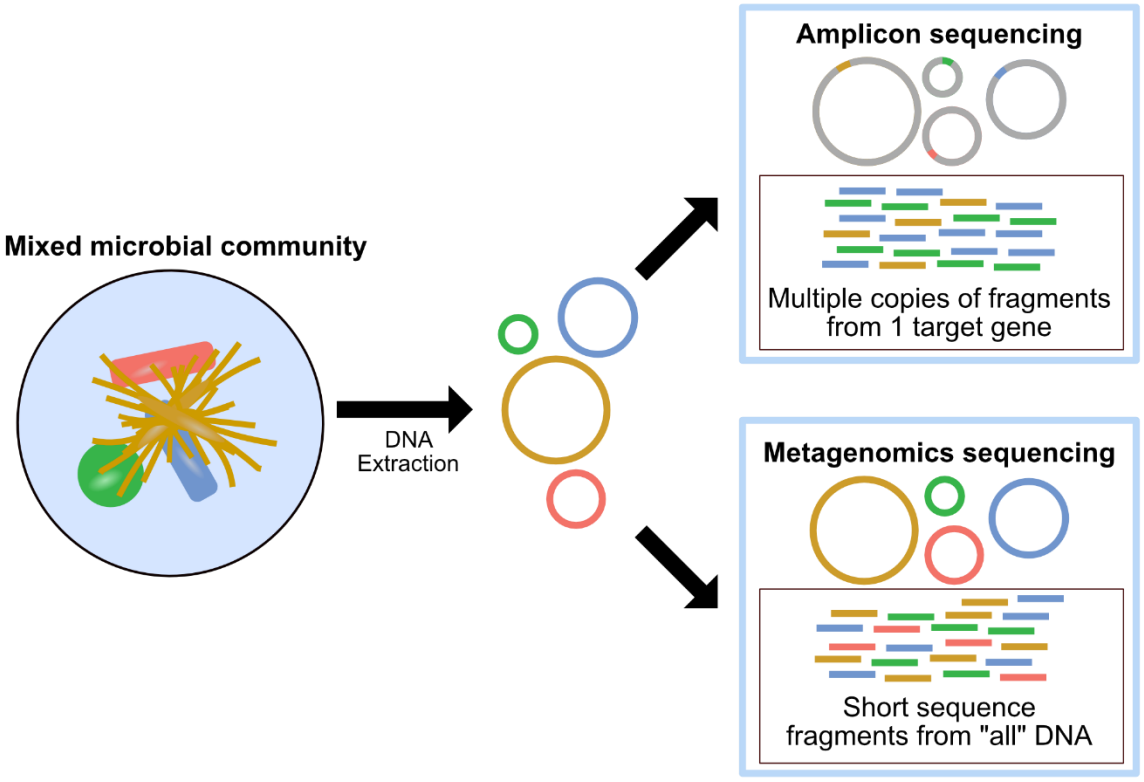

Figure 1-3 Diagrammatic overview of shotgun metagenomics vs amplicon sequencing. Reprinted (with permission) from the Happy Belly Bioinformatics github page (https://astrobiomike.github.io/misc/amplicon_and_metagen)

1.5 Processing metagenomic data from shotgun sequencing

1.5.1 Host read removal and sequence classification

Shotgun sequencing generates millions of short reads which need to be classified to determine microbial community composition. At this stage it is common to identify and remove any reads that originate from host DNA prior to classifying the remaining microbial reads, this can be done rapidly using tools such as bowtie2 to align all reads in the dataset against the host genome and remove those that successfully map (Knight *et al.*, 2018). After host read removal, the remaining microbial reads need to be classified. Many approaches are available, but they can be broadly grouped into two categories: assembly-based approaches and reference-based (assembly-free) approaches (Quince *et al.*,

2017). **Figure 1-4** gives an overview of the steps to produce, process and analyse a metagenomic dataset, however it doesn't show steps for human contaminant removal and relative abundance calculations that we added to our study.

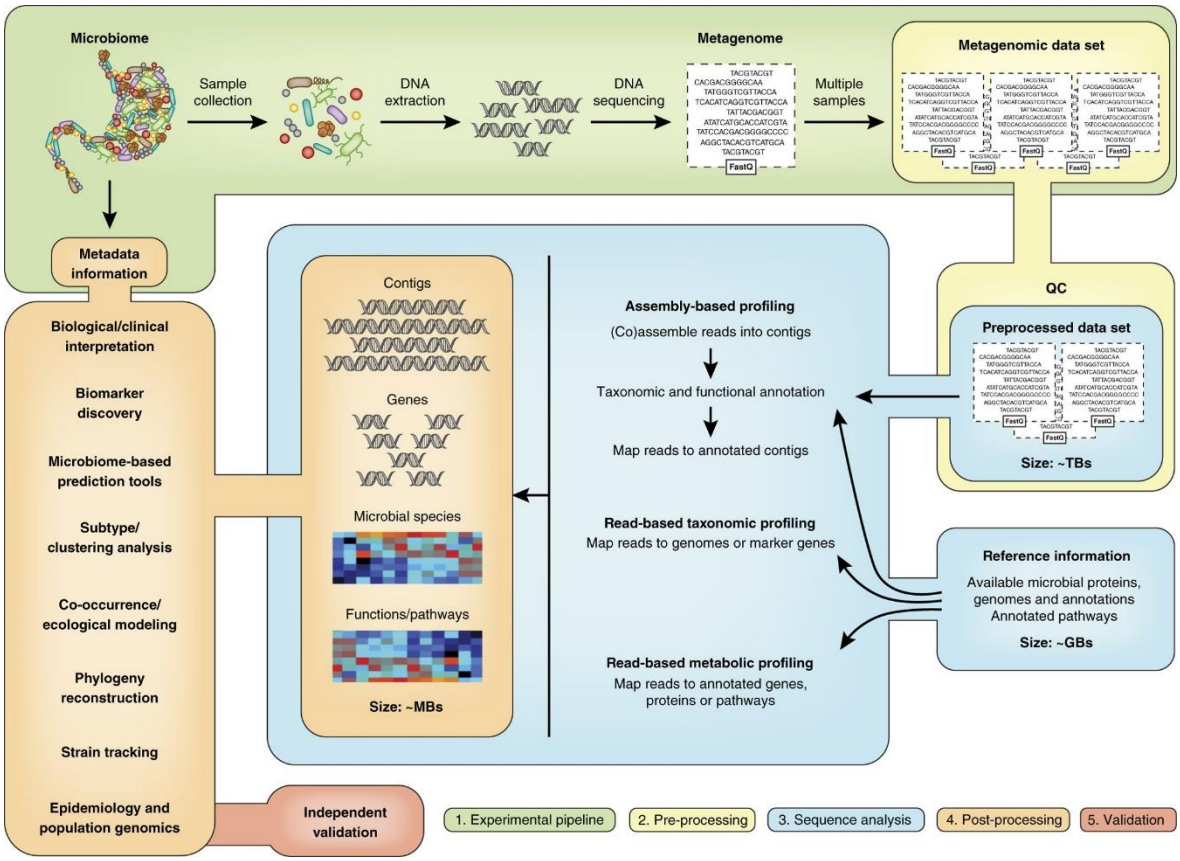

Figure 1-4 Procedural flowchart of data collection and analysis in shotgun metagenomics Reprinted (with permission) from Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol. 2017 Sep 12;35(9):833–44

Assembly-based approaches involve the *de novo* assembly of the reads into species-level contigs (Quince *et al.*, 2017). For all but the most abundant species in a sample, these contigs will usually be small genomic fragments, rather than full genomes, due to limited sequencing depth, repetitive sequences and strain-level variation (Alneberg *et al.*, 2014). Incomplete contigs from the same species can be grouped together by binning approaches; this is usually done in a reference-free manner because many microbial species are not present in reference databases (Quince *et al.*, 2017). For example, CONCOCT groups contigs into species bins by leveraging the fact that contigs belonging to the same species should have a) similar k-mer frequencies and b) correlated read-depth coverage across multiple samples (Alneberg *et al.*, 2014). Once binning is

complete, the reads can be classified by aligning them against the binned contigs.

Reference-based approaches, also referred to as taxonomic classifiers, involve the direct classification of reads using reference databases (Piro, Matschkowski and Renard, 2017). Taxonomic classifiers can be further broken into two categories, taxonomic binners and taxonomic profilers (Piro, Matschkowski and Renard, 2017; Sczyrba *et al.*, 2017). Binners attempt to classify every read to the most likely organism in a reference database, usually by sequence alignment or through compositional information, such as k-mer matching (Piro, Matschkowski and Renard, 2017; Quince *et al.*, 2017). Taxonomic profilers estimate the relative abundances of taxa rather than assigning individual reads (Piro, Matschkowski and Renard, 2017; Sczyrba *et al.*, 2017; Meyer *et al.*, 2019), they usually follow a marker-based approach, such as in MetaPhlan2 (Truong *et al.*, 2015) or mOTUs2 (Milanese *et al.*, 2019), where reads are aligned against a database of taxonomically informative 'marker genes'. Taxon relative abundances are then calculated from the hits to marker sequences, this contrasts with taxonomic binners which calculate the proportion of all reads mapping to each taxon (Piro, Matschkowski and Renard, 2017; Sczyrba *et al.*, 2017; Meyer *et al.*, 2019). Relative abundances can be estimated from the output of taxonomic binners, which is discussed later in the section.

In general, assembly-based methods are useful for high-confidence classifications and for identifying previously unclassified organisms (Quince *et al.*, 2017), but suffer from several drawbacks. Firstly, these methods are computationally expensive, which can be prohibitive for very large datasets (Quince *et al.*, 2017), whereas reference-based methods can be very efficient and fast (Menzel, Ng and Krogh, 2016). Secondly, assembly-based methods may end up grouping closely related species together into chimeric contigs (Ayling, Clark and Leggett, 2019). Finally, these methods struggle to identify low-abundance species as there are often insufficient number of reads present to assemble contigs for these species (Quince *et al.*, 2017). Marker-based taxonomic profilers also have problems detecting low-abundance taxa, as the reads are only aligned against reference databases of marker sequences, rather

than whole genome sequences, so rare taxa (that have low genome coverage) may not be detected (McIntyre *et al.*, 2017).

Taxonomic binners provide much better recovery of low-abundance taxa because they use more extensive reference databases, resulting in many more reads being classified. A downside to this approach is that many taxonomic binners suffer from high rates of false-positives due to a) a higher number of reads being classified and b) attempting to classify reads originating from genomic regions that are conserved across multiple species (Piro, Matschkowski and Renard, 2017). Some taxonomic binners employ least-common ancestor (LCA) methods that attempt to limit false-positives by assigning ambiguous reads that could match multiple taxa at a higher taxonomic level (Wood and Salzberg, 2014; Menzel, Ng and Krogh, 2016).

1.5.2 Accounting for uneven sequencing depth between samples

The number of reads obtained from a sequencing run can vary between samples, and this can affect downstream diversity estimates because as sequencing depth increases, more species are discovered (until a saturation point is reached) (Weiss *et al.*, 2017; Zaheer *et al.*, 2018). As a result, samples with lower sequencing depth can appear less diverse. It is therefore important to account for this effect before any meaningful comparison between samples can be drawn (Goodrich *et al.*, 2014; McMurdie and Holmes, 2014; Weiss *et al.*, 2017). The traditional normalisation approach is called rarefying, where for each sample an equal number of reads are sub-sampled without replacement (Goodrich *et al.*, 2014; Weiss *et al.*, 2017). This often leads to difficult trade-offs between the inclusion of more samples or retaining a higher sequencing depth, and whatever decision is made results in a loss of data (Goodrich *et al.*, 2014; Weiss *et al.*, 2017). Rarefying has been criticised for two major reasons: firstly, it discards potentially useful data. Second, sequencing data is inherently compositional, species/OTU read counts do not represent the absolute abundances, rather they represent the proportions of reads relative to each other. Rarefying does not treat the data as compositional and this leads to a theoretical increase in false-positive errors when testing for the differential abundance of taxa between groups (McMurdie and Holmes, 2014; Gloor *et al.*, 2017). Various other data

normalisation methods have been suggested, these involve transformations that account for compositionality of the data without the loss of data (Weiss *et al.*, 2017). More recently, tests on simulated data found that rarefying performed equally as well as other data transformations for differential abundance testing (Weiss *et al.*, 2017). Furthermore, rarefying has been identified as the only approach that was able to fully account for uneven sequencing depth, especially those with large library sizes (Weiss *et al.*, 2017; McKnight *et al.*, 2019). For this reason, rarefying remains the simplest and most robust method.

1.5.3 Calculating Relative abundances

As discussed in section 1.5.1, taxonomic binners classify each read individually, meaning the output is the total number (or proportion) of reads in each sample that were assigned to each taxon. The proportion of reads assigned to a taxon is distinct from its relative abundance because the likelihood of a read from a genome being sequenced increases with genome size (Nayfach and Pollard, 2016), creating a bias towards the sequencing of organisms with larger genomes. The relative abundance of each taxon (taxonomic profile) can thus be estimated as the total number of reads mapping to that taxon, divided by its genome size (McIntyre *et al.*, 2017; Piro, Matschkowski and Renard, 2017; LaPierre *et al.*, 2019). Unfortunately, calculating relative abundances in this way does come with some complications. In a metagenomic sample, the number of reads that map to a genome depend not only on its size, but also its size relative to the average genome size of all other genomes present (Nayfach and Pollard, 2016). Consequently, abundances can be correctly estimated when all genome sizes are known, but unclassified genomes with unknown length can skew results. To complicate matters further, many classifiers assign ambiguous reads at the taxonomic level of the LCA (e.g. genus) to avoid incorrect assignment (Menzel, Ng and Krogh, 2016; Wood, Lu and Langmead, 2019). Obviously a genus does not have a genome size, so an estimate must be made. Finally, deciding on the genome size to assign, even at species level, can be difficult because there may be multiple genome assemblies to choose from. While this process is by no means straightforward, it is essential to determine the true proportions of the species that are present in microbiome samples, rather than simply using the proportions of reads.

1.6 Microbiome analysis

A very common approach is differential abundance or "biomarker" testing, the identification of taxa whose abundance differs between groups. This can be challenging because microbiome datasets often have few replicates and hundreds or thousands of taxa, many of which are absent in most samples (data sparsity) (Knight *et al.*, 2018). Most importantly, microbiome data are compositional, reflecting the relative proportions of taxa rather than their absolute abundances (Gloor *et al.*, 2017; Weiss *et al.*, 2017; Knight *et al.*, 2018). Consequently, an increase in the relative abundance of one taxa must result in the decrease of others (such that their total sum is always 1) which does not necessarily reflect what is happening with their true abundances. Many tools have been developed to tackle some of these issues over the years, including DESeq2 (Love, Huber and Anders, 2014) and ANCOM (Mandal *et al.*, 2015) which are both very popular. More recently MaAasLin2 (Mallick *et al.*, 2021) was released, a highly customisable tool that is designed to overcome the difficulties associated with detecting differential abundance in sequencing data. Various log and log-like transformations can be chosen which accounts for compositionality in the data by linearising the associations between taxa. Minimum prevalence and abundance testing can be used to control i) sparsity of the dataset and ii) the number of tests being performed. There is also a choice of models including mixed models allow the use of cross-sectional and longitudinal samples to be used.

Alpha diversity measures are a simple way of characterising microbiome features based on the presence and abundance of taxa. Alpha diversity metrics measure properties of a single sample, such as richness: the number of species present, or evenness: how evenly distributed the abundance of species present are. Non-microbial ecosystems with a low richness and/or low evenness are more vulnerable to invasion (Levine and D'Antonio, 1999). Similar patterns have been observed in the gut microbiome, where reduced alpha-diversity results in more free metabolites, which facilitates pathogenic invasion (Gillis, Hughes, Spiga, Winter, Zhu, Carvalho, *et al.*, 2018; Herren and Baym, 2018).

Beta diversity metrics are a valuable technique for the pairwise comparison of samples, a commonly used metric is Bray-Curtis dissimilarity (BCD). BCD represents the differences in abundance of hundreds or thousands of species as a single metric, ranging from 0 (nothing in common) to 1 (identical species and abundances), that can be used to compare samples (Wagner *et al.*, 2018). In this way, each sample in a dataset can be compared with every other sample and their differences can be represented as a pairwise BCD matrix. Beta diversity matrices can be used alongside ordination methods such as principal coordinates analysis (PCoA), which condenses all of the compositional differences between many samples into a single ordination plot on a 2- or 3-dimensional axis that enables researchers to see if samples group by relevant characteristics (McKnight *et al.*, 2019). Beta diversity matrices can also be used with statistical tools such as Permutational Multivariate Analysis of Variance (PERMANOVA), which statistically test for differences in microbiome composition associated with sample metadata (Anderson, 2017). Longitudinal samples can be compared with beta diversity metrics to investigate how the microbiome changes over time. This is relevant because disease-associated microbiomes can be more unstable in response to external perturbations such as antibiotics, immunosuppression or invasion by pathogens (Zaneveld, McMinds and Vega Thurber, 2017).

1.7 Aims of the study

Our study aims to compare the nasopharyngeal microbiomes of 10 HIV-infected individuals vs. 6 HIV-uninfected controls over time. We will utilise shotgun sequencing of DNA and RNA samples to capture microbiome species from all kingdoms. We will include samples at 3 timepoints (0, 1 & 9 Months) from each participant to measure how the microbial community of the nasopharynx changes over time, in both HIV-infected and HIV-uninfected individuals. We hypothesise that the NPM differs between HIV-infected and HIV-uninfected individuals and that this might contribute to the increased susceptibility of HIV-infected individuals to secondary bacterial infections.

First, we will characterise the taxonomic composition and diversity of the nasopharynx in HIV-infected and HIV-uninfected individuals, identifying major taxa and also looking specifically at the abundances of common URT pathobionts

and HIV-associated pathogens. We will use biomarker identification tools to search for taxa that are differentially abundant between these groups. PERMANOVA will be used to determine whether HIV infection or other host or environmental factors are associated with significant changes to the overall composition of the nasopharyngeal microbiome. Finally, we will measure longitudinal changes in community composition using BCD and investigate whether HIV-infection leads to changes in the temporal stability of the nasopharyngeal microbiome.

To our knowledge, this study represents the first attempt to characterise the nasopharyngeal microbiome of HIV-infected individuals. The nasopharynx can either repel pathogens or act as a reservoir, so determining how the nasopharyngeal microbial community differs by HIV status could help us understand whether it is implicated in primary or secondary bacterial lung infections which are more common in PLHIV. This work is also significant as a proof of concept; to our knowledge, when we began the project there had been no published work which used shotgun sequencing to characterise the nasopharyngeal microbiome.

2 Materials and Methods

2.1 Study design and sample Collection

Nasopharyngeal swab samples had been previously collected as part of a previous cohort study that investigated the effect of HIV status on the incidence of influenza-like illness (ILI). The study was carried out between April 2013 and March 2015 at the Queen Elizabeth Central Hospital in Blantyre district, Malawi. It consisted of 608 Malawian adults and included HIV-infected and HIV-uninfected individuals who gave paired oropharyngeal and nasopharyngeal samples at routine bimonthly visits or when they experienced an ILI episode. The protocol for sample collection was as follows: expose the nostrils by applying gentle upwards pressure to the tip of the nose, insert a flocked nasopharyngeal swab along the floor of the nose until it reaches the posterior pharynx, leave the swab for 5 seconds and then withdraw it in a rotating motion, place the swab into a universal transport medium tube and frozen at -70°C. The chosen samples for this study were transported to the University of Glasgow on dry ice and stored at -70°C.

2.2 Sample Selection

The overall aim of this study was to investigate the effect of HIV status on the human nasopharyngeal microbiome. All samples that were taken within 30 days of the participant taking antibiotics or suffering an ILI episode were discarded as this would affect microbiome composition.

Using longitudinal samples was deemed to be important because i) more samples increases the probability of detecting real differences between individuals, as microbiome composition can change with time and each sample only represents a snapshot, and ii) longitudinal sampling enables the detection of how the microbiome changes with time and whether HIV status affect this. Most of the tools available for longitudinal microbiome analysis require samples that are taken at regular timepoints. However, not all of the samples from the study were available, and some patients weren't able to attend all of their bimonthly visits. We therefore searched the available data for potential subsets that i)

contained longitudinal samples taken at the same time intervals and ii) contained enough participants for downstream analysis. This dataset was chosen by analysing the number of cases and controls with sufficient longitudinal samples for all possible combinations of the following variables i) total number of timepoints(2-9), ii) timepoint range(0-24 months) and iii) permitted deviance of the sample date from the exact timepoint(0-14 days). A subset of the data was identified that consisted of 19 cases and 10 controls which all had 3 timepoints, baseline, 1 month and 9 months with a maximum deviance of each sample from the timepoint of 10 days.

We then looked into the participant metadata and identified the important factors that might affect NPM composition such as age, sex, HIV-status, CD4+ counts, asthma, chronic lung disease, smoking status, smoking history, previous tuberculosis, previous pneumonia, cooking fuel and number of children under 5 years in the household. We were able to select a subset of 10 cases and 6 controls who did not report a history of asthma or chronic lung disease, and were lifelong non-smokers, which allowed us to remove these variables from the analysis. We were unable to fully remove the factors of enrolment age, sex, previous tuberculosis, previous pneumonia, type of cooking fuel used and children under 5 in the household.

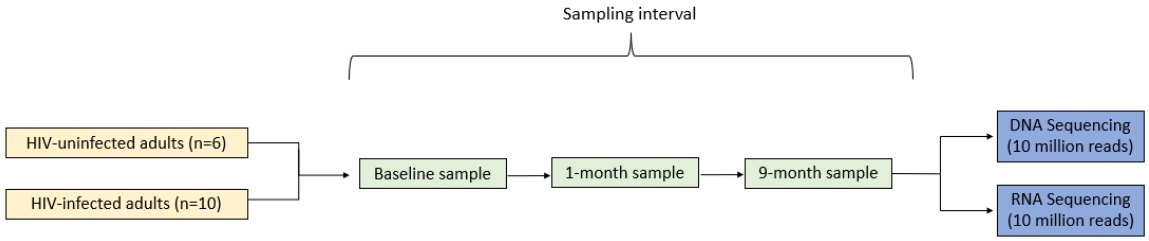

Figure 2-1 Diagrammatic flow chart showing the study cohort and sampling.

Table 1 - Participant metadata including Sex, Age, Children under 5 in the household, principal cooking fuel/energy source and medical history by HIV status

Characteristic	Participants, No./Total(%)	
	HIV infected (n=10)	HIV uninfected (n=6)
Female sex	7 (70%)	5 (83.3%)
Age, mean	36.8	34.9
Children under 5 in the household		
0	4 (40%	3 (50%)
1	5 (50%)	3 (50%)
2	1 (10%)	0 (0%)
Principal cooking fuel/energy source		
Firewood	3 (30%)	2 (33.3%)
Charcoal	7 (70%)	4 (66.7%)
Medical history		
Previous tuberculosis infection	2 (20%)	0 (0%)
Previous pneumonia	1 (10%)	0 (0%)

2.3 DNA Extraction, sample preparation and sequencing

Nucleic Acid extraction, library preparation and sequencing were carried out by Lily Tong and Chris Davis at the MRC-University of Glasgow Centre for Virus Research.

Briefly, up to 500 µl samples (nasopharyngeal swabs stored in viral transport medium) were slowly defrosted on ice and then centrifuged at 5000g for 1 minute to spin down human cells and debris (Thurber *et al.*, 2009). 50µl nucleic acid (both RNA and DNA) from each sample was extracted using the eMAG nucleic acid extraction platform and stored at -70°C. The nucleic acid of each sample was divided into 2 parts, which were processed for DNA and RNA sequencing respectively.

To reduce human genomic DNA contamination within the nucleic acid extraction, NEBNext Microbiome DNA Enrichment Kit (NEB, E2612) was used to remove CpG-

methylated host DNA prior to library preparation (Feehery *et al.*, 2013). The DNA was then sheared into proximal 350 base pair fragments by sonication (Covaris Sonicator LE220), and uniquely indexing tagged with NEBNext Multiplex Oligos for Illumina (New England Bio-Labs). KAPA LTP Library Preparation Kit (Roche7961880001) is used for this process. Libraries were then sequenced on Illumina NextSeq500 platform with paired ends for 2 x 150 base pair reads. A high output cartridge Kit v2.5 (300 Cycles) was used (Illumina 20024908).

2.4 Processing of the sequencing data

2.4.1 Adapter Trimming and Quality Filtering

Trimmomatic v0.39 (Bolger, Lohse and Usadel, 2014) was used to clip Illumina adaptor sequences and trim regions of low quality from the reads by performing sliding window trimming with a window size of 4 and a required average phred quality score of 28. Reads were discarded if their total length was less than 20 after trimming. The mate-pairs of reads that were discarded are retained as orphan reads for downstream analysis.

2.4.2 Identification and removal of host reads prior to taxonomic classification

Even after steps to reduce human contamination prior to sequencing, we still observed a high proportion of sequenced reads were of human origin. Human DNA contamination is a specific challenge to profiling the microbiota of the respiratory tract due to their low density (Man, de Steenhuijsen Piters and Bogaert, 2017), coupled with their relatively tiny genomes relative to human hosts. We therefore implemented a rigorous two-stage approach to human read removal using the alignment tools bowtie2 (Langmead and Salzberg, 2012) and SNAP (Zaharia *et al.*, 2011); in a benchmark comparison by Bush et al. this was the highest-performing method for the detection of human reads in microbial datasets (Bush *et al.*, 2020). After initial trimming/filtering, the remaining paired-end and orphan reads were aligned against the human genome GRCh38 (Schneider *et al.*, 2016) using bowtie2 (v2.3.1) on default settings; reads that successfully mapped onto GRCh38 were filtered from the dataset. The remaining

reads were then aligned against the newly completed CHM13 (Nurk *et al.*, 2022) human genome using SNAP (v1.0.18). For paired-end reads, if a single mate of a mate-pair could be aligned to the human genome, the entire read was considered human and discarded. Multiple human genome assemblies were chosen to increase the sensitivity of human read detection. Further filtering of human reads also occurred during (**section 2.4.3**) and after the taxonomic classification stage (**section 2.4.4**).

2.4.3 Taxonomic classification of reads

Sequencing reads that did not align to the human genome represent the DNA component of the nasopharyngeal microbiome. These reads were extracted using Samtools (v1.9) (Li *et al.*, 2009). The reads for each sample were then grouped into one file and formatted for classification by Kraken2 (v2.1.2) (Wood, Lu and Langmead, 2019), enabling Kraken2 to classify the paired-end and orphaned reads from each sample together. The read-merger.pl script (provided on Kraken2 github page) was used to concatenate each read pair into a single read sequence separated by an N character, the orphaned reads were then appended to the end of this file. Kraken2 classifies reads by computing a set of k-mers from each read and matching them to the lowest common ancestor of the genomes in the database that contain that k-mer (Wood and Salzberg, 2014). Kraken2 was run with default settings using the PlusPF database (Wood, Lu and Langmead, 2019) which contains k-mer profiles from all archaeal, viral, plasmid, human, protist and fungal sequences in the NCBI RefSeq database (O'Leary *et al.*, 2016).

Kraken2 was selected for taxonomic classification in our study for several reasons: Kraken2 is a taxonomic binner, meaning it will attempt to classify every read that has been sequenced, which is preferable to marker-based or assembly-based classification methods for samples with low read counts, as in our dataset. Kraken2 outperforms similar tools in both processing speed and memory usage (Wood, Lu and Langmead, 2019), these would be unfeasible to run with the limited computational resources available. Kraken2 allows the use of custom databases which is not possible with some comparable tools. Including human sequences in the classification database allows Kraken2 to correctly classify

human reads that have escaped the bowtie2/SNAP filtering stages, preventing these reads from being misclassified as microbial.

2.4.4 Post-classification removal of human reads

Taxonomic binners, including Kraken2 which was implemented in our pipeline, make use of reference databases to classify sequences (Quince *et al.*, 2017). Contamination in reference databases is a well-known issue, including human contamination of microbial reference sequences and microbial contamination of human reference sequences (Breitwieser *et al.*, 2019; Steinegger and Salzberg, 2020). The consequence of reference database contamination is the misclassification of reads when the classifier matches a read to a mislabeled sequence in the reference database, which can result in either i) a read of microbial origin being classified as human and is excluded, or ii) a read of human origin being classified as microbial, and is retained in the microbial dataset. Any human reads that leaked into the microbial dataset this way would be classified as the microbial taxa which contains the contaminated/mislabeled sequences in the database, which we have termed “contaminant taxa”.

We reasoned that any contaminant taxa would be highly conserved across all samples, because every sample contains human DNA in high abundance. Therefore, we compiled a list of 232 taxa (as classified by Kraken2) which were present in more than 80% of samples. We then used MegaBLAST (Morgulis *et al.*, 2008) to re-classify a randomised subset of 500 reads that had been assigned to each taxa by Kraken2, using the NCBI non-redundant nucleotide database (nr/nt) (NCBI Resource Coordinators, 2018) as a reference, and manually curated the results to identify any taxa that contained a large number of human reads. MegaBLAST is widely considered to be the gold-standard in metagenomic sequence classification (Bazinet *et al.*, 2018); it is the most sensitive and highest performing method. However, it is usually unfeasible to run MegaBLAST on large metagenomics datasets containing millions of reads as it is too computationally expensive (Ye *et al.*, 2019). Using MegaBLAST, we identified and removed 3 contaminant taxa (*Eukayota*, *Toxoplasma gondii ME49* and *Fusarium pseudograminearum CS3096*) from our Kraken2 classified dataset that totaled 10,045 reads, of which more than 80% were of human and non-microbial origin.

To measure our ability to detect and remove human contamination from our dataset prior to, during and post-taxonomic classification, we once again used MegaBLAST using nr/nt as a reference database to classify a random subset of 1000 reads from each sample prior to any filtering and after all filtering had taken place. Prior to any filtering, human reads comprised 43978/48000 (91.62%) of the total reads in the sample. After all filtering had taken place, there were 3247/48000 (6.76%) human reads that remained. We therefore estimate a total human contamination level of 6.76% in our dataset that we were unable to remove.

2.4.5 Rarefaction curves and rarefying samples

As previously discussed, the sequencing depth of a microbiome sample affects key diversity metrics and its ability to be compared to other samples, therefore microbiome studies need to implement a normalization approach to account for uneven sequencing depth across samples (McMurdie and Holmes, 2014; Weiss *et al.*, 2017). Rarefication was chosen as the normalization method for our study, where each sample is artificially subsampled to an even sequencing depth, and all samples with lower sequencing depth are removed. Rarefying involves a trade-off between the sequencing depth chosen and the number of samples that are retained.

Rarefaction curves are used to measure the effect of sequencing depth on sample characteristics. The alpha_rarefaction function from the diversity plugin for the QIIME2 microbiome bioinformatics platform (Bolyen *et al.*, 2018) was used to generate rarefaction curves. Curves were generated by simulating subsets of each sample at a sequencing read depth of 0-20,000 reads at 1000 read intervals and calculating the following alpha diversity metrics for each interval: species richness, Shannon index and Pielou's evenness. Once an appropriate sequencing depth cut-off had been chosen, the rarefy function in QIIME2 was used to rarefy the samples.

2.4.6 Generating taxonomic profiles

The total number (or proportion) of reads assigned to a taxon is distinct from its relative abundance because the likelihood of a read from a genome being

sequenced increases with genome size (Nayfach and Pollard, 2016), creating a bias towards the sequencing of larger genomes. The relative abundance of each taxon can thus be estimated as the total number of reads mapping to that taxon, divided by its genome size, followed by rescaling of the genome-size normalised read counts so that their sum is equal to 1 (McIntyre *et al.*, 2017; Piro, Matschkowski and Renard, 2017; LaPierre *et al.*, 2019). We calculated genome sizes for the taxa in our dataset using the NCBI Microbial Genomes database (NCBI Resource Coordinators, 2018). After calculating genome-size normalised read counts for each taxon, we then filtered reads that had been classified at the genus level or above (14.7% of total reads) and grouped the remaining taxa at the species level, the genome-normalised read counts were then converted to relative abundance by scaling them so that the total sum of taxa in each sample is equal to 1.

The output format of the taxonomic profile is a table where rows represent species, columns represent samples and cell values are the relative abundances of species in a sample. In addition to the taxonomic profile for the whole microbiome, we created a subset of taxonomic profiles for the different fractions of the microbiome split by superkingdom. These taxonomic profiles that were used for downstream analysis were: microbiome (all species), bacteriome (bacterial species only), virome (viral species only), eukaryome (eukaryote species only), archaeome (archaeal species only).

2.5 Microbiome analysis

2.5.1 Taxonomic composition visualisation

To give a broad overview of the species composition of the nasopharyngeal microbiome, an averaged taxonomic profile for cases and controls was created by taking the mean relative abundance of species of all the samples in these groups. These averaged taxonomic profiles were then loaded into Krona (Ondov, Bergman and Phillippy, 2011) to generate interactive reports, still images of these reports are included in the results section.

2.5.2 Alpha diversity

2.5.2.1 Alpha diversity metrics

For each sample, we calculated the richness as the number of observed species. We also calculated Pielou's evenness to quantify how equally distributed the taxa are based on their relative abundance. Species richness and Pielou's evenness metrics used were calculated using the diversity plugin for the QIIME2 (Bolyen *et al.*, 2018).

2.5.2.2 Modelling alpha diversity

To determine whether participant characteristics had any impact on alpha diversity metrics, ordinary least-square (OLS) linear regression models were produced to model the mean richness or evenness from all of a participants samples as a function of their HIV status, enrolment age and type of cooking fuel used. Models were created in python using the OLS function in the statsmodel package (Seabold and Perktold, 2010).

2.5.3 Beta diversity

2.5.3.1 Beta diversity metrics

The Bray-Curtis dissimilarity (BCD) between each pair of samples was calculated using the QIIME2 diversity plugin and used to generate pairwise dissimilarity matrices that were used for sample ordination, modelling longitudinal beta diversity, testing for differences in group dispersions/centroids and differential abundance analysis.

2.5.3.2 Sample ordination

To visualise the compositional differences among all of our samples simultaneously, a principal coordinate analysis (PCoA) was carried out using the QIIME2 diversity plugin. The subsequent PCoA matrix was visualised using the emperor plugin for QIIME2.

2.5.3.3 Modelling longitudinal beta diversity

To assess the stability of microbiome over time in our participants, and how host or environmental factors might affect this, we extracted the BCD between all longitudinal sample pairs for each participant. An OLS linear regression model was then fit to model these longitudinal BCD measurements as a function of participant age, HIV status and the type of cooking fuel used. Models were created in Python using the OLS function in the statsmodel package (Seabold and Perktold, 2010).

2.5.3.4 Testing group dispersions/centroids with PERMDISP/PERMANOVA

Permutational analysis of variance (PERMANOVA) (Anderson, 2017) was used to test for the effect of participant metadata on microbiome composition. Briefly, the BCD matrix is converted into principal coordinate space and OLS regression models are fit to determine whether the centroids of test groups are different. We generated a pairwise BCD matrix for each participant using the average species composition across all of their available samples. We then implemented PERMANOVA on the participant BCD matrix using the adonis2 function of the vegan (v2.6-2) package in R (Oksanen J. et al., 2022). We initially tested for the effects of all participant metadata (sex, enrolment age, cooking fuel, HIV status, number of children under 5 in the household, past history of pneumonia and tuberculosis) but later reduced this to HIV status, cooking fuel and enrolment age.

The betadisper function from the Vegan (v2.6-2) package in R was used to carry out Permutational analysis of multivariate dispersions (PERMDISP), which measures the dispersion of samples in different groups. For all the samples in each group, the BCD matrix is converted into principal coordinate space and a centroid is generated, the dispersion of the group is then measured as the average distance between each sample and the group centroid. A pseudo-F statistic is calculated that represents the difference in dispersion of the two groups, a significance value is then obtained by randomly permuting the group labels for a set number of iterations, re-calculating group dispersions and then observing the proportion of permutations where the pseudo-F statistic is greater than between the real groups.

2.5.3.5 Differential abundance analysis

The MaAsLin2 R package (Mallick *et al.*, 2021) was used to test for the differential abundance of species by participant age, HIV status and type of cooking fuel used. MaAsLin2 allows the use of multiple types of models, data transformations and normalisations to suit the requirements of the study. We ran MaAsLin2 as a linear mixed-effects model without any normalisations or transformations to the data, as these are designed to address uneven sampling depth, which has already been done for our data. We used all of the available samples with the participant ID being included as a random effect and participant age, HIV status and type of cooking fuel used as fixed effects.

3 Results

3.1 Sequencing depth normalisation

After filtering out species with non-RNA genomes from the RNA dataset, the remaining read counts from the identified species were very low (**Table 2**) so we opted to continue solely with the analysis of the DNA dataset. To account for highly uneven sequencing depth in our DNA samples (**Table 2**), we chose to rarefy our samples to a standard sequencing depth. Rarefaction curves were created to investigate the relationship between sequencing depth, key diversity metrics and the number of samples retained (**Figure 3-1**). Pielou's evenness index plateaued at the first sequencing depth interval of 1000 reads and very gradually declined as depth increased. Observed species richness increases rapidly at lower sequencing depth, with the rate of species discovery decreasing as sequencing depth increases. However there was still a steady increase in species discovery even at the maximum sequencing depth of 20,000 reads.

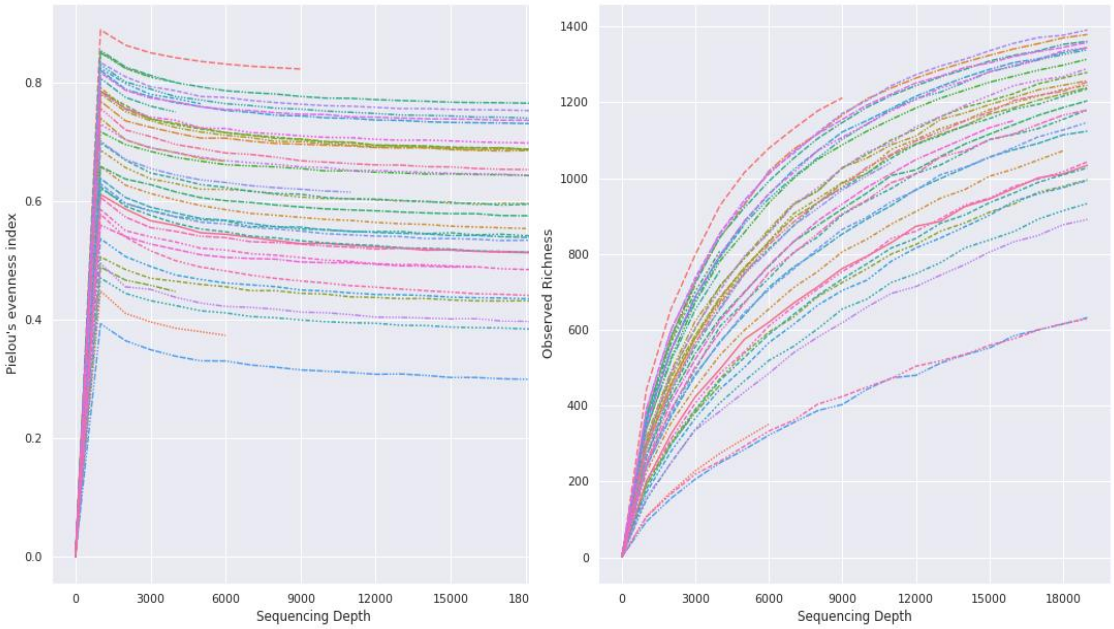

Figure 3-1. Rarefaction curves indicating the Pielou's evenness index (left) or observed species richness (right) of the samples in our dataset at varying sequencing depths.

Table 2 - Sample metadata including HIV status (HIV), CD4+ counts (CD4), enrolment age (enrage), sex, previous tuberculosis infection (pasttb), previous pneumonia infection (pastpn), type of fuel used for cooking (cook), children under 5 years old in the household (und5) and the number of reads classified at the species level for the DNA and RNA samples. Rows highlighted in yellow represent DNA samples that were filtered out because they had fewer species-level reads than the chosen rarefication depth of 4026 reads.

Patient ID	HIV	CD4	enrage	sex	pasttb	pastpn	cook	und5	Sample ID	Species reads (DNA)	Species reads (RNA)
Case1	Y	411	31.21	F	N	Y	Charcoal	0	Case1-1	6754	14
									Case1-2	33240	2
									Case1-3	477561	25
Case2	Y	357	37.63	M	N	N	Charcoal	0	Case2-1	4084	28
									Case2-2	40524	107
									Case2-3	37161	4
Case3	Y	407	49.08	M	N	N	Firewood	2	Case3-1	7085	26
									Case3-2	632414	33
									Case3-3	479013	27
Case4	Y	216	28.85	F	N	N	Charcoal	1	Case4-1	50273	79
									Case4-2	1877	496
									Case4-3	62764	42
Case5	Y	465	49.89	M	N	N	Firewood	0	Case5-1	28547	13
									Case5-2	128098	32
									Case5-3	17251	19
Case6	Y	508	47.3	F	N	N	Charcoal	1	Case6-1	4499	11
									Case6-2	21461	13
									Case6-3	235	478
Case7	Y	326	37.28	F	Y	Y	Charcoal	0	Case7-1	33460	9
									Case7-2	24530	95
									Case7-3	88728	30
Case8	Y	403	30.09	F	N	N	Charcoal	1	Case8-1	4724	34
									Case8-2	6150	29
									Case8-3	705	27
Case9	Y	340	31.1	F	N	N	Firewood	1	Case9-1	47482	64
									Case9-2	218520	1
									Case9-3	54975	20
Case10	Y	329	25.63	F	N	N	Charcoal	1	Case10-1	9889	18
									Case10-2	45354	23
									Case10-3	11842	48
Ctrl1	N		32.1	F	N	N	Charcoal	1	Ctrl1-1	4026	55
									Ctrl1-2	1065	20
									Ctrl1-3	1334	57
Ctrl2	N		22.73	M	N	N	Firewood	0	Ctrl2-1	26629	54
									Ctrl2-2	252	18
									Ctrl2-3	331495	35
Ctrl3	N		32.88	F	N	N	Charcoal	1	Ctrl3-1	19839	9
									Ctrl3-2	4842	4374
									Ctrl3-3	49296	43
Ctrl4	N		37.65	F	N	N	Firewood	0	Ctrl4-1	23120	27
									Ctrl4-2	48179	226
									Ctrl4-3	102339	150
Ctrl5	N		35.13	F	N	N	Charcoal	1	Ctrl5-1	109417	8
									Ctrl5-2	299618	1
									Ctrl5-3	599043	19
Ctrl6	N		48.8	F	N	N	Charcoal	0	Ctrl6-1	2500	93
									Ctrl6-2	368976	354
									Ctrl6-3	705314	31

We chose a cut-off value of 4026 reads for our rarefication depth as it represented a good trade-off between sample retention and dataset size for downstream analysis. At 4026 reads, Pielou's evenness index has largely stabilised and around 500-800 microbial species can already be detected in most samples (**Figure 3-1**). At the cut-off value of 4026 reads, 7 samples were removed from downstream analysis (**Table 2**). Ordination of the samples revealed that they grouped tightly together based on the number of reads they contained prior to rarefication, demonstrating that sequencing depth has a large effect on sample composition and as a result needs to be normalised prior to analysis (**Figure 3-2a**). After rarefication, the samples no longer cluster together based on original sequencing depth, indicating that rarefication removed this effect (**Figure 3-2b**).

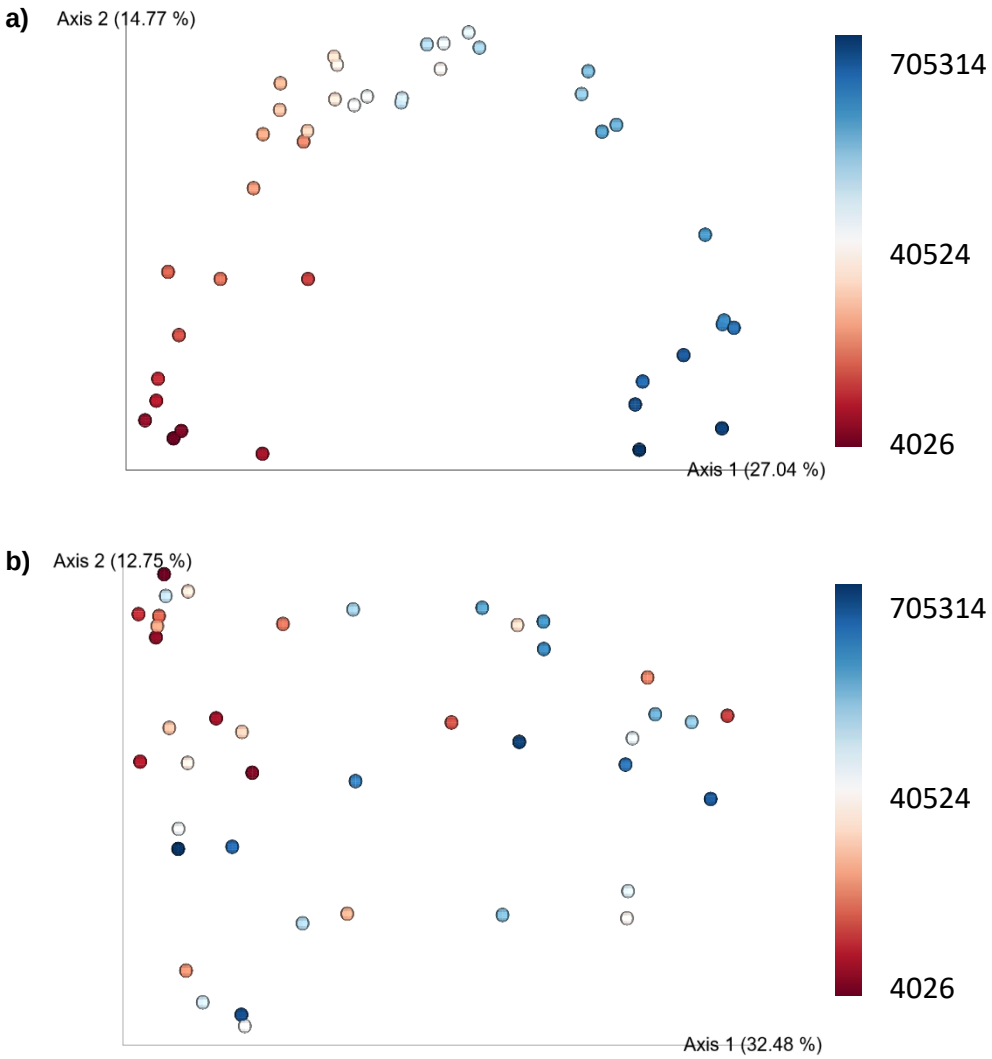

Figure 3-2. Principal coordinate analysis showing the ordination of samples by microbial composition before rarefication **(a)** or after rarefication **(b).** Samples are coloured according to the number of reads they contained prior to rarefication.

3.2 Overview of taxonomic composition

As part of the metagenomic pipeline, we calculated the relative abundance of each species by applying a transformation to the rarefied read counts. This transformation considers the genome size of the species in question, as the likelihood of sequencing a read from a given genome increases with its size. To determine the effect of the relative abundance calculations on our data, we visualised the composition of the four superkingdoms in our dataset using relative read proportions and the newly calculated relative abundances (**Figure 3-3**). Prior to any relative abundance calculations, bacteria dominated the dataset, with over 94% of the reads being classified as bacterial in 40/41 samples. The relative abundance calculations had a dramatic effect on the relative proportion of the viral superkingdom. Viral relative abundances were much higher than their relative read proportions, which is to be expected as viruses have tiny genomes relative to bacteria and eukaryotes in our dataset. Interestingly, there was a large amount of variance in viral relative abundance between samples, and viral relative abundance appeared to be much higher in controls compared to HIV-infected cases.

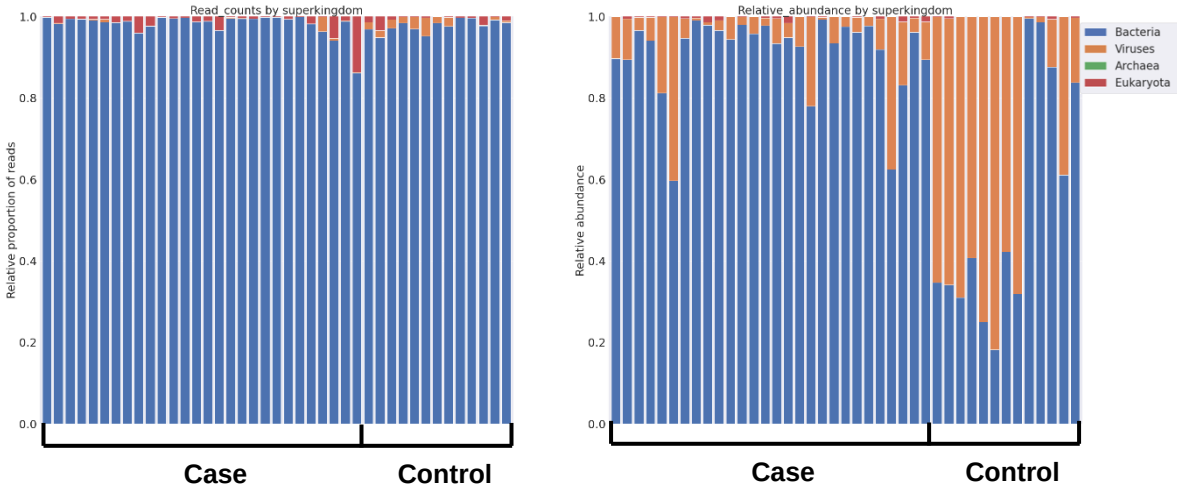

Figure 3-3. Stacked bar plots showing the relative read proportion (left) and relative abundance (right) of each superkingdom per sample.

Next, we investigated the composition of the NPM at multiple taxonomic levels for cases and controls. Krona reports were used to visualise the mean relative abundance of taxa across all case samples or all control samples. Archaea and eukaryotes were underrepresented in both cases and controls, with viruses and

bacteria accounting for over 99% of the microbial population (**Figure 3-4)**. As noted previously, mean viral relative abundance was higher in controls than cases (43% vs 9%). The overwhelming majority of the viruses detected in the cases and controls belong to the order Caudoviricetes, which consists of tailed bacteriophage (**Figure 3-5**). The genus Pahexavirus made up 96% of the bacteriophage in controls and 56% in cases, within this genus we identified 40 unique species that all infect the bacterium *Cutibacterium acnes* (**Supplementary Table 1**). We also detected several other phage species that were shared in cases and controls. While bacterial relative abundance was lower in controls than cases (57% vs 90%), the taxonomic composition of the bacteriome was similar for both groups (**Figure 3-6**). Major bacterial genera included *Corynebacterium*, *Cutibacterium*, *Staphylococcus*, *Kocuria*, *Micrococcus*, *Janibacter*, *Nocardioides* and *Paracoccus*. Most of these genera were present in very similar proportions when averaged across cases and controls. The largest differences in mean relative abundance were in the genera *Cutibacterium* (18% of bacteria in controls, 7% in cases) and Staphylococcus (3% of bacteria in controls, 8% in cases). The mean relative abundance of *C. acnes* was much higher in controls than cases (14% of bacteria vs 5%), whereas *Staphylococcus warneri* and *Staphylococcus aureus* were more abundant in cases than controls (**Supplementary Table 2**).

In most samples there were 9 major genera which accounted for 60% or more of the total microbial population in the sample. However, the relative proportions of these genera varied considerably between individuals and within-individuals over time (**Figure 3-7**). Of these major genera, *Corynebacterium*, *Cutibacterium*, *Kocuria*, *Micrococcus*, *Janibacter*, *Nocardioides* and *Paracoccus* were detected in all samples, *Staphylococcus* was detected in 40/41 samples and *Pahexavirus* was detected in 25/41 samples (**Table 3**). Some samples were dominated by one or a few genera, while others had much more evenly distributed populations. *Corynebacterium*, *Pahexavirus* and *Staphylococcus* were among the most variable genera which dominated (>20% relative abundance) some samples but were much rarer or not present in others. *Corynebacterium* and *Staphylococcus* were dominant in a higher proportion of case samples than control samples. *Pahexavirus* was found in 5/6 controls and was present in all samples analysed for 4/6 controls, it tended to dominate the microbial

population in controls but not cases, and its abundance was closely associated with its bacterial host *Cutibacterium*.

Table 3 - Frequency of detection for 9 major genera in cases and controls

Genus	**Case samples (n=27)**	**Control Samples (n=14)**
Corynebacterium	27 (100%)	14 (100%)
Cutibacterium	27 (100%)	14 (100%)
Janibacter	27 (100%)	14 (100%)
Kocuria	27 (100%)	14 (100%)
Micrococcus	27 (100%)	14 (100%)
Nocardioides	27 (100%)	14 (100%)
Pahexavirus	14 (52%)	11 (79%)
Paracoccus	27 (100%)	14 (100%)
Staphylococcus	26 (96%)	14 (100%)

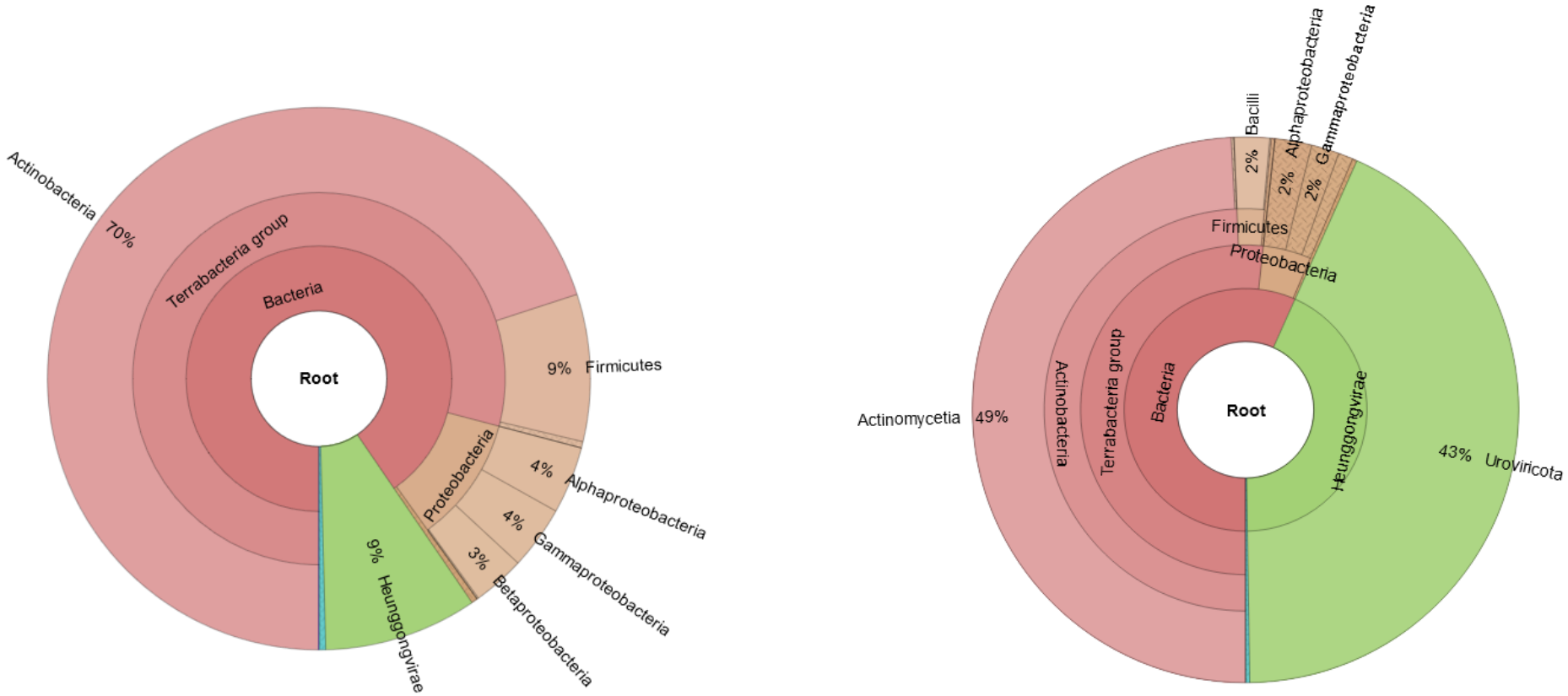

Figure 3-4. Krona reports showing the mean taxonomic composition of the microbiome of cases (left) and controls (right).

Figure 3-5. Krona reports showing the mean taxonomic composition of the virome for cases (left) and controls (right)

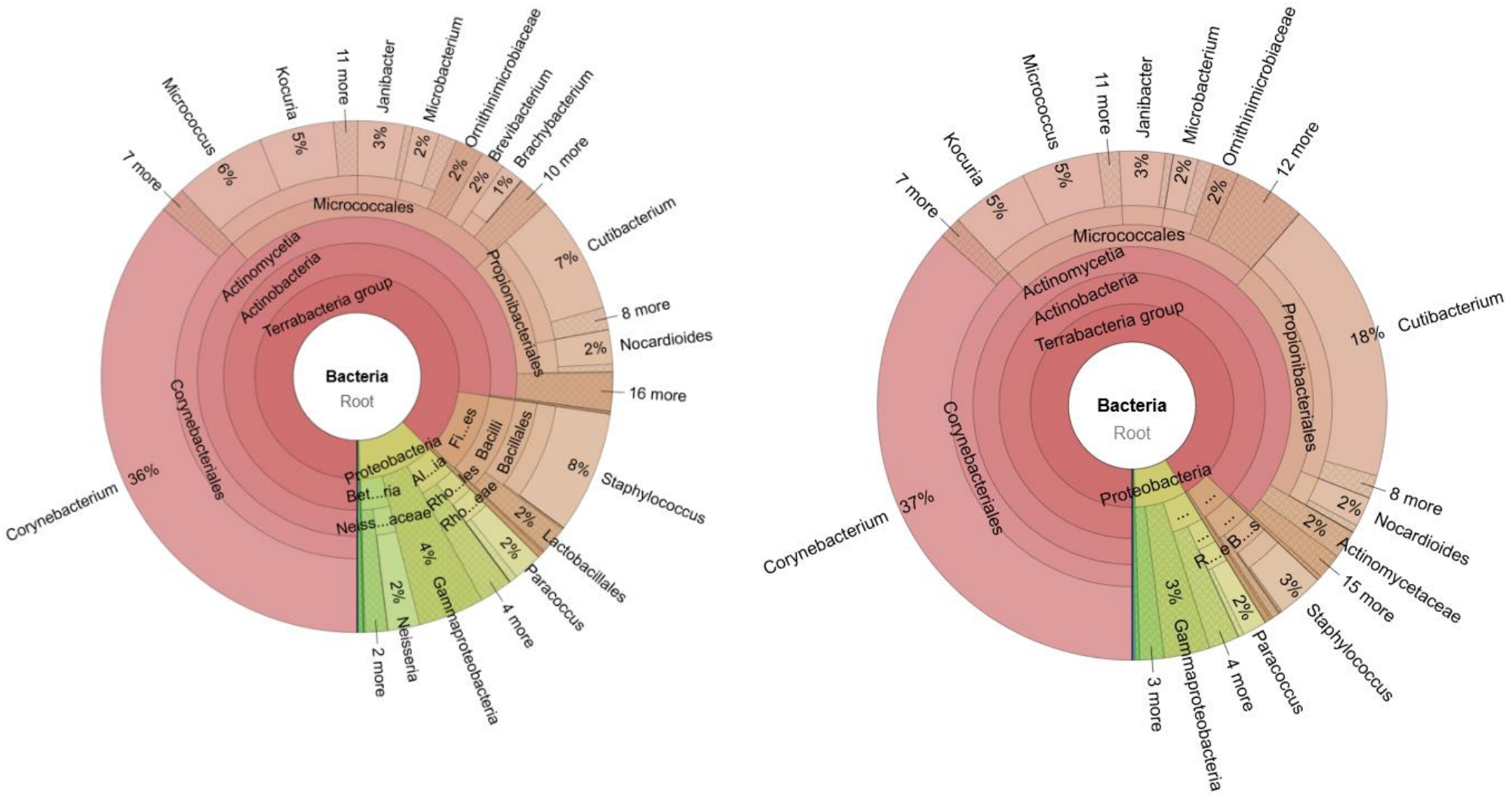

Figure 3-6. Krona reports showing the mean taxonomic composition of the bacteriome for cases (left) and controls (right)

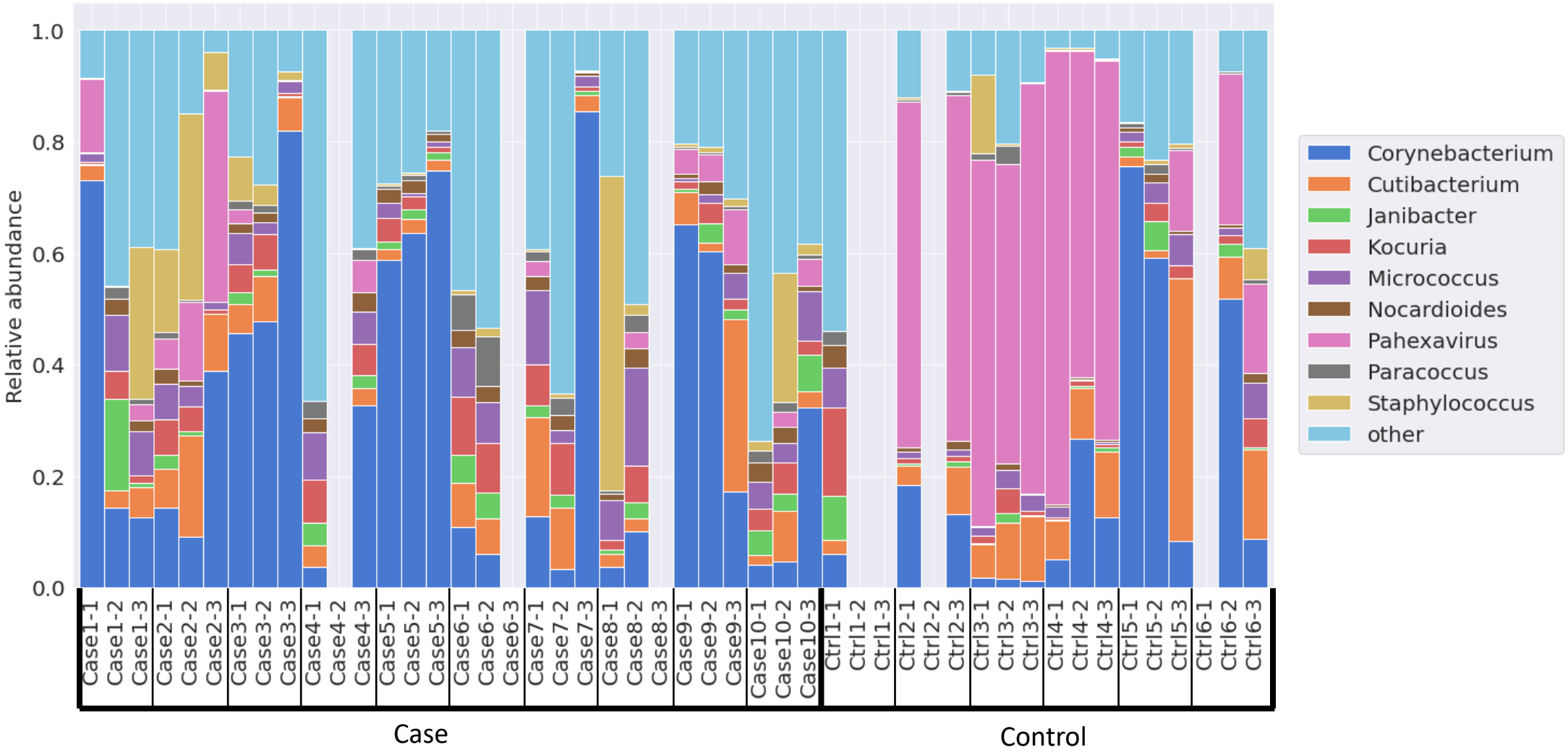

Figure 3-7. Stacked bar plots showing the relative abundance of 9 major genera in each sample. Samples are ordered longitudinally by patient and grouped by HIV status.

We searched for the presence of pathogens commonly associated with HIV-infection that could potentially be detected in the nasopharynx including: *Microbacterium tuberculosis*, *Cryptococcus neoformans*, *Cryptosporidium spp.*, *Histoplasma capsulatum*, *Pneumocystis jirovecii*, Herpes simplex virus and Human cytomegalovirus (CMV) ('CDC: AIDS and Opportunistic Infections', 2021). The only HIV-related pathogen we identified CMV, which was detected in at least 1 sample from 5/10 cases and not in any of the 6 controls (**Figure 3-8**). We also measured the abundance of common URT pathobionts whose overgrowth can lead to URT and LRT infections including: *Staphylococcus aureus*, *Haemophilus influenzae*, *Streptococcus pneumoniae*, *Moraxella catarrhalis* and *Neisseria meningitidis*. Relative abundance of pathobionts was low across all samples apart from Case8-1 where *S. aureus* constituted 60% of total bacterial abundance in the sample.

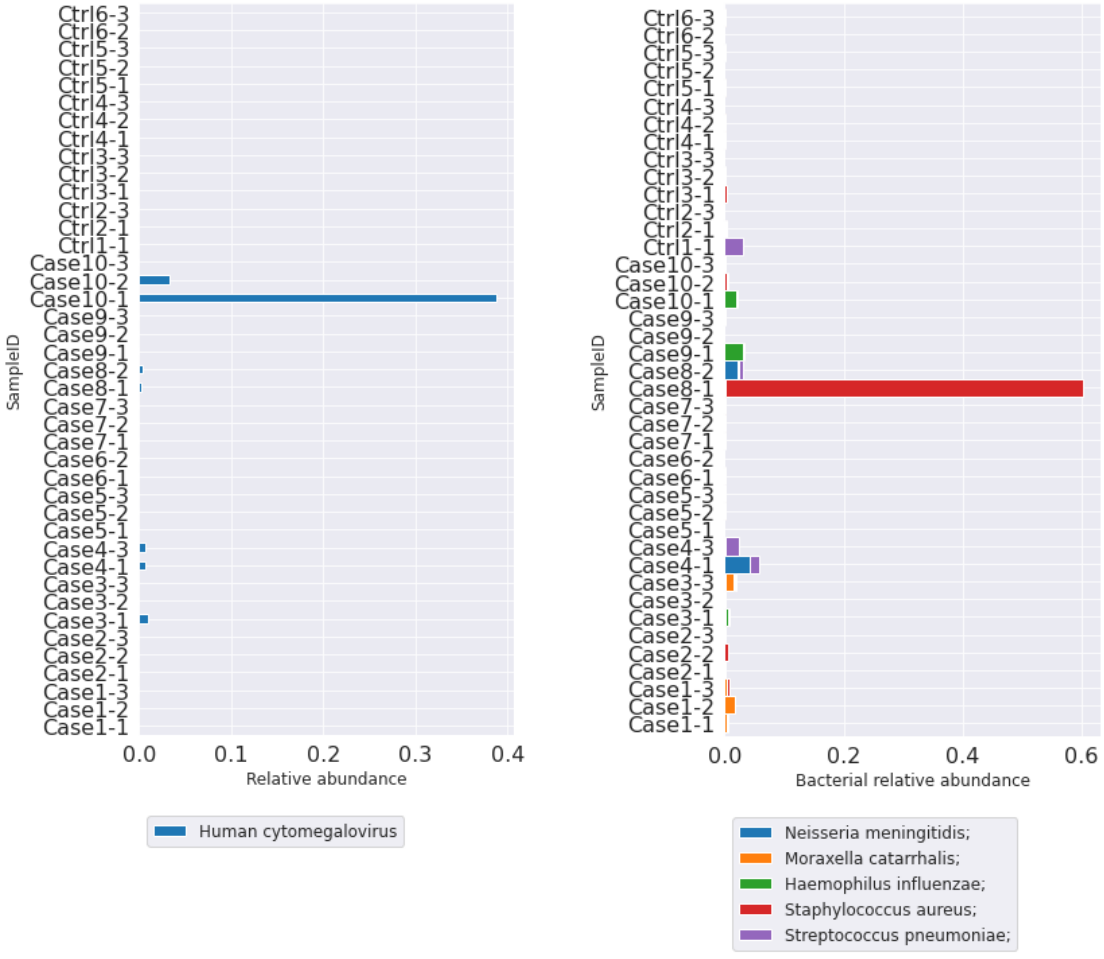

Figure 3-8. Horizontal bar plots showing (left) The relative abundance of HIV-associated pathogens in our samples and (right) Relative bacterial abundance of common URT pathogens in our samples.

3.3 Factors affecting microbiome composition

We used PERMANOVA to test the hypothesis that HIV-infected adults would significantly differ in NPM compared to healthy controls (**Table 4**). Initially we ran PERMANOVA with all the available participant metadata as explanatory variables, but this led to too much partitioning of the data to identify any significant associations. There was also collinearity between HIV status and past history of pneumonia/TB, as well as sex and cooking fuel (**Table 2**). We therefore re-ran PERMANOVA using HIV-status, type of cooking fuel used and enrolment age as explanatory variables as these were the predictors with the highest F-scores in the first test. With these parameters, HIV-status was found to be a significant predictor of microbiome composition. The type of cooking fuel used was borderline significant as a predictor of microbiome composition ($p<0.1$),but was a significant predictor of bacteriome composition($p<0.05$). We also tested the dispersion of sample composition by HIV-status and type of cooking fuel using PERMDISP, neither factor had a significant effect on group dispersion (**Figure 3-9**).

Table 4 – PERMANOVA results showing factors that affect microbiome composition with degrees of freedom (Df), effect size(F) and statistical significance (Pr>F) using different models. Statistical significance: (.) designates a Pr>F of less than 0.1 and (*) less than 0.05

Model	Factor	Df	F	Pr>F
All explanatory variables	sex	1	0.570	0.833
	cook	1	1.130	0.328
	enrage	1	0.873	0.516
	und5	1	0.696	0.696
	pasttb	1	0.458	0.885
	pastpn	1	0.499	0.916
	hiv	1	1.419	0.189
HIV, cooking fuel and enrolment age	hiv	1	2.378	0.027(*)
	cook	1	1.8461	0.072(.)
	enrage	1	1.0411	0.3687
HIV, cooking fuel and enrolment age (bacteriome)	hiv	1	1.047	0.354
	cook	1	2.299	0.035(*)
	enrage	1	1.191	0.2685

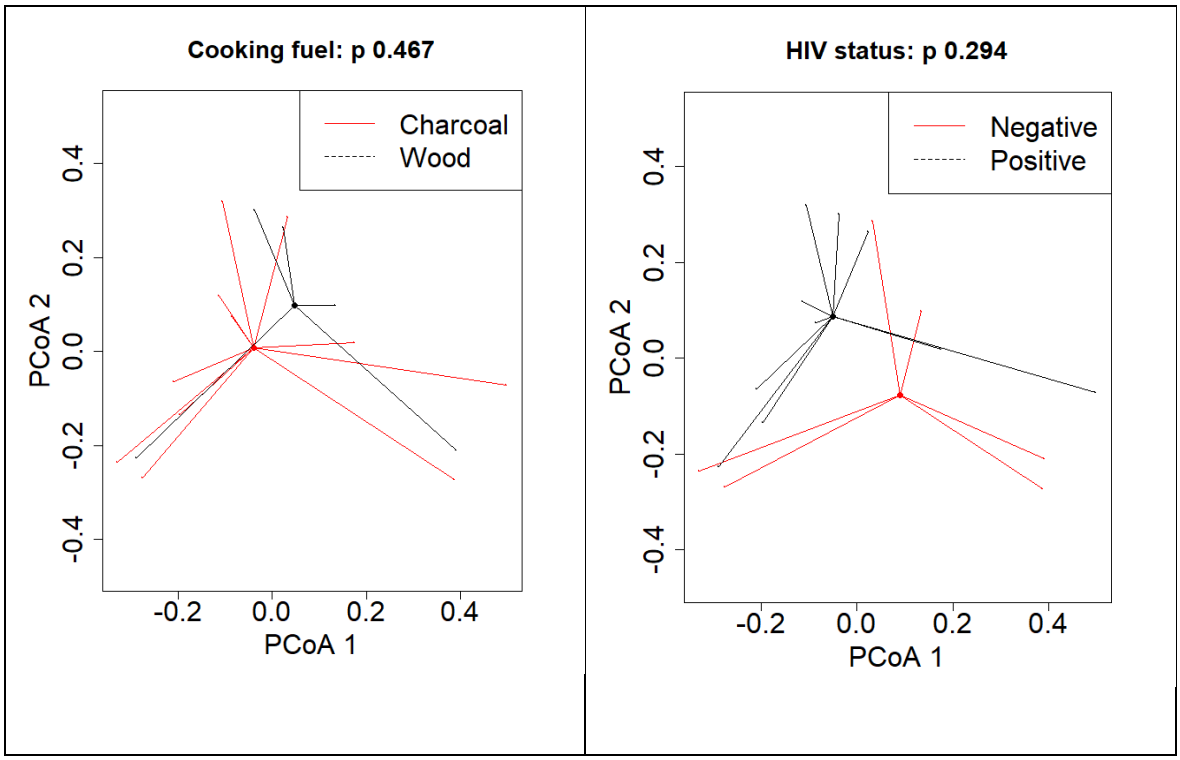

Figure 3-9. Principal Coordinate Analysis plots showing the dispersion of samples by type of cooking fuel used (left) and HIV-status (right). Statistical significance is labelled above the plot.

3.4 Alpha diversity analysis

We used multivariate regression to study the relationship between participant characteristics and the richness or evenness of their microbiome and its compartments (bacteriome, virome, eukaryome and archaeome). We were unable to detect an effect of cooking fuel, HIV status or enrolment age on overall species richness or evenness (**Table 5**). In the viral species richness model we observed a highly significant association between viral species richness and HIV status, with controls having 15 more viral species on average after accounting for other explanatory factors ($p<0.01$). For the bacterial evenness model, HIV status was a borderline significant predictor of bacterial evenness with controls having 10.9% lower bacterial evenness on average after accounting for other explanatory factors ($p<0.1$). These differences can be seen in the richness and evenness plots below (**Figure 3-10**).

Table 5 – Alpha diversity multivariate regression model results showing the coefficients, standard errors, t-statistic and statistical significance (P>|t|) of test factors. Statistical significance: (.) designates a P>|t| of less than 0.1, (*) less than 0.05, (**) less than 0.01

Model	Factor	Coef.	Std.Err.	t	P>|t|
All species richness	Intercept	775.612	141.059	5.498	
	cook[T.Firewood]	-18.349	68.201	-0.269	0.792
	Hiv[T.control]	-0.273	64.835	-0.004	0.997
	enrage	-3.618	3.826	-0.946	0.363
Viral species richness	Intercept	26.445	10.083	2.623	
	cook[T.Firewood]	7.428	4.875	1.524	0.154
	Hiv[T.control]	15.319	4.635	3.305	0.006(**)
	enrage	-0.28	0.274	-1.025	0.325
All species evenness	Intercept	0.552	0.128	4.319	
	hiv[T.control]	-0.068	0.057	-1.185	0.259
	cook[T.Firewood]	0.021	0.06	0.353	0.73
	enrage	-0.001	0.003	-0.26	0.8
Bacterial species evenness	Intercept	0.589	0.116	5.076	
	hiv[T.control]	-0.109	0.052	-2.094	0.058(.)
	cook[T.Firewood]	-0.006	0.055	-0.116	0.909
	enrage	-0.002	0.003	-0.556	0.588

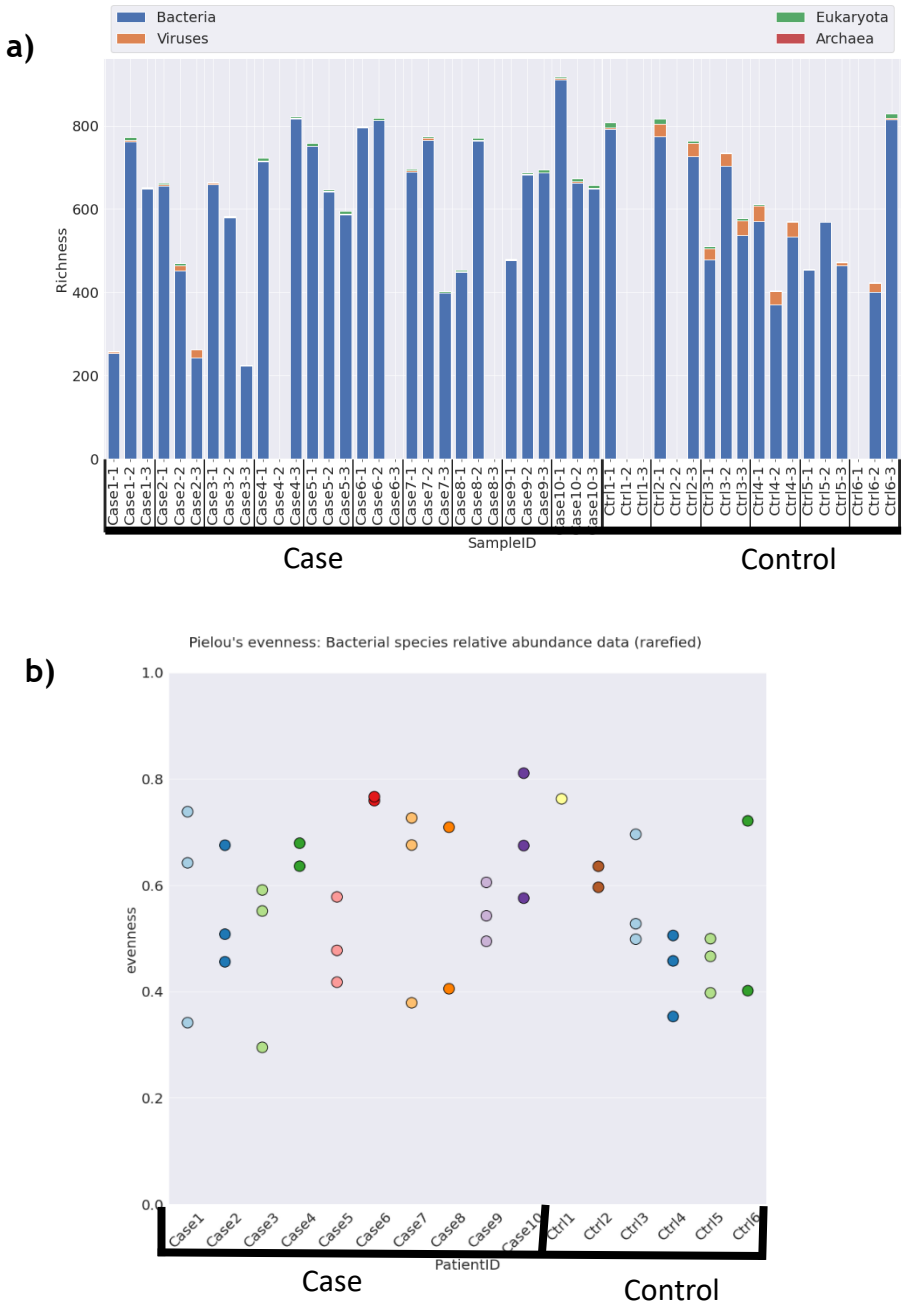

Figure 3-10. (a) Stacked bar plot of overall species richness per sample coloured by superkingdom **(b)** Pielou's evenness index for the bacteriome of each sample, grouped by individual.

3.5 Longitudinal beta diversity analysis

Multivariate linear regression models were used to determine whether participant characteristics had any effect on the stability of their microbiome over time. We modelled the Bray-Curtis dissimilarity (BCD) between all possible longitudinal sample pairings. We used the time between samples and participant HIV status, type of cooking fuel used and enrolment age as explanatory variables (**Table 6**). For the all-species model, cooking fuel and time between samples

were significant predictors of BCD ($p<0.05$), whereas HIV status was borderline significant($p<0.1$). BCD between samples was lower in controls than cases and those who used firewood instead of charcoal, and increased time between sampling led to increased BCD. For the virome model, HIV status and cooking fuel were both significant predictors ($p<0.05$) of BCD between sample pairs, but longer time between sampling was not associated with larger BCD between sample pairs.

Table 6 - Multivariate regression models of longitudinal beta diversity results showing the coefficients, standard errors, t-statistic and statistical significance (P>|t|) of each factors. Statistical significance: (.) designates a P>|t| of less than 0.1, (*) less than 0.05, (**) less than 0.01

Model	**Factor**	**Coef.**	**Std.Err.**	**t**	**P>|t|**
Longitudinal Bray-Curtis: All species	Intercept	0.726	0.118	6.134	0
	hiv[T.control]	-0.09	0.048	-1.869	0.072(.)
	cook[T.Firewood]	-0.17	0.05	-3.404	0.002(**)
	Time_difference[T.8.0]	0.16	0.055	2.917	0.007(**)
	Time_difference[T.9.0]	0.129	0.054	2.376	0.024(*)
	enrage	-0.004	0.003	-1.322	0.196
Longitudinal Bray-Curtis: Viral species	Intercept	0.825	0.206	4.008	0.001
	hiv[T.control]	-0.208	0.081	-2.567	0.017(*)
	cook[T.Firewood]	-0.243	0.086	-2.837	0.009(**)
	Time_difference[T.8.0]	-0.128	0.096	-1.328	0.197
	Time_difference[T.9.0]	0.033	0.095	0.345	0.733
	enrage	0.004	0.006	0.661	0.515

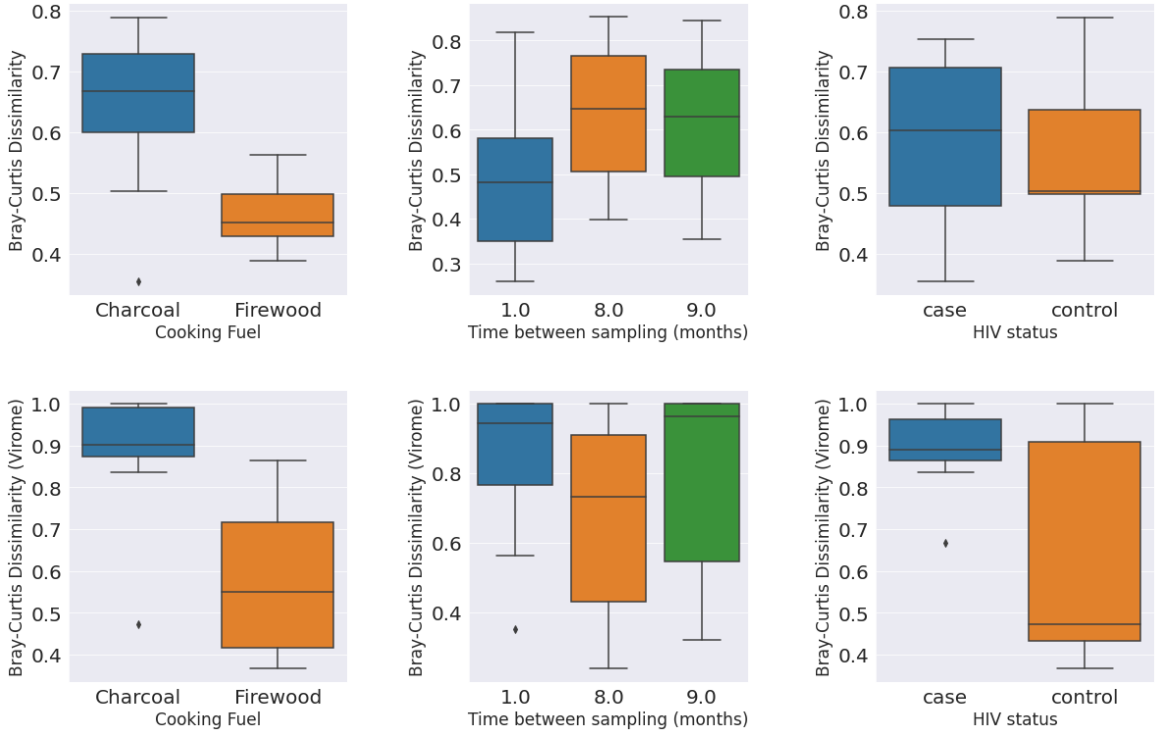

Figure 3-11. Box plots showing the distribution of BCD values by sample grouping for (top) the all-species model, and (bottom) the virome model.

3.6 Differential abundance analysis

We carried out differential abundance testing using MaAsLin2 to identify taxa that were differentially abundant between groups (**Table 7**). Due to our low number of participants (6 cases vs 10 controls), any effect size would need to be large to achieve statistical significance. As the number of statistical tests increases, the minimum detectable effect size decreases because a multiple testing correction is applied for each test to control the false discovery rate. We therefore opted to limit the number of statistical tests being carried out by restricting the groups analysed to HIV status and cooking fuel, and limiting the taxa analysed to the 9 major genera identified in Section 3.2. *Micrococcus* abundance was associated with cooking fuel, average relative abundance of *Micrococcus* was 3.8% lower in those who used firewood($q<0.05$). For HIV status, *Pahexavirus* relative abundance was 33.1% lower in cases on average ($q<0.05$), and *Nocardioides* abundance was 0.9% higher on average($q<0.1$). We also did differential abundance testing of the major bacterial genera using their relative bacterial abundance. Using relative bacterial abundances negates the effect of Pahexavirus relative abundance, which is much higher in controls, on the

relative abundance of bacterial genera and therefore improves sensitivity to detect differential abundance of bacterial genera. When testing in this way, we found that relative bacterial abundance of Cutibacterium was significantly higher in controls than cases and Corynebacterium was more abundant in those who used firewood as fuel (**Table 8**).

Table 7 – MaAasLin2 results showing the differential abundance of 9 major genera according to HIV status and type of cooking fuel used. Results show the coefficients (coef), standard error (stderr), statistical significance (pval) and statistical significance after Benjamini-Hochberg procedure to correct for multiple testing (qval). Statistical significance: (.) designates a qval of less than 0.1 and (*) less than 0.05

Genus	metadata	value	coef	stderr	pval	qval
Micrococcus	cook	Firewood	-0.038	0.01	0.001	0.014(*)
Corynebacterium	cook	Firewood	0.208	0.095	0.048	0.173
Janibacter	cook	Firewood	-0.018	0.009	0.067	0.201
Kocuria	cook	Firewood	-0.03	0.016	0.11	0.247
Nocardioides	cook	Firewood	-0.006	0.003	0.104	0.247
Pahexavirus	cook	Firewood	0.149	0.1	0.158	0.259
Paracoccus	cook	Firewood	-0.014	0.009	0.139	0.259
Staphylococcus	cook	Firewood	-0.058	0.036	0.146	0.259
Cutibacterium	cook	Firewood	-0.009	0.028	0.763	0.763
Pahexavirus	HIV	Case	-0.331	0.096	0.004	0.039(*)
Nocardioides	HIV	Case	0.009	0.003	0.012	0.073(.)
Micrococcus	HIV	Case	0.024	0.01	0.028	0.126
Staphylococcus	HIV	Case	0.052	0.036	0.185	0.277
Corynebacterium	HIV	Case	0.124	0.094	0.21	0.283
Cutibacterium	HIV	Case	-0.035	0.028	0.22	0.283
Janibacter	HIV	Case	0.01	0.009	0.276	0.331
Paracoccus	HIV	Case	0.008	0.009	0.338	0.381
Kocuria	HIV	Case	0.008	0.016	0.642	0.679

Table 8 - MaAasLin2 results (bacterial abundance data) showing the bacterial differential abundance of 9 major bacterial genera according to HIV status and type of cooking fuel used. Results show the coefficients (coef), standard error (stderr), statistical significance (pval) and statistical significance after Benjamini-Hochberg procedure to correct for multiple testing (qval). Statistical significance: (.) designates a qval of less than 0.1 and (*) less than 0.05

Genus	metadata	value	coef	stderr	pval	qval
Corynebacterium	cook	Firewood	0.293	0.084	0.001	0.011(*)
Micrococcus	cook	Firewood	-0.04	0.012	0.002	0.012(*)
Janibacter	cook	Firewood	-0.017	0.01	0.087	0.252
Microbacterium	cook	Firewood	-0.008	0.005	0.097	0.252
Kocuria	cook	Firewood	-0.03	0.014	0.078	0.252
Staphylococcus	cook	Firewood	-0.075	0.044	0.098	0.252
Paracoccus	cook	Firewood	-0.016	0.01	0.129	0.291
Nocardioides	cook	Firewood	-0.003	0.004	0.46	0.755
Cutibacterium	cook	Firewood	0.02	0.038	0.594	0.755
Cutibacterium	HIV	Case	-0.141	0.038	0.001	0.011(*)
Paracoccus	HIV	Case	0.004	0.01	0.705	0.755
Corynebacterium	HIV	Case	0.031	0.084	0.713	0.755
Janibacter	HIV	Case	0.005	0.01	0.594	0.755
Microbacterium	HIV	Case	0.003	0.005	0.581	0.755
Micrococcus	HIV	Case	0.005	0.012	0.688	0.755
Nocardioides	HIV	Case	0.002	0.004	0.587	0.755
Staphylococcus	HIV	Case	0.035	0.044	0.437	0.755
Kocuria	HIV	Case	0	0.014	0.976	0.976

4 Discussion

The study captured and compared the nasopharyngeal microbiomes of 10 HIV-infected individuals against 6 HIV-uninfected controls using shotgun sequencing of DNA and RNA samples. We characterised the nasopharyngeal microbiome by processing the sequencing data to create taxonomic profiles representing the relative abundance of each species in our samples. Nasopharyngeal microbiome composition was compared between cases and controls, and we also attempted to measure the impact of other variables including participant age, sex, cooking fuel, children under 5 in the household and medical history of respiratory infections.

4.1 Taxonomic composition of the nasopharyngeal microbiome

This work was a pilot study of using shotgun metagenomics to characterise the nasopharyngeal microbiome. In the absence of a known ground truth to validate our results against, comparing what we have found against published 16s rRNA studies can help us validate whether our results make sense. The major bacterial genera that we detected were *Corynebacterium*, *Cutibacterium*, *Staphylococcus*, *Kocuria*, *Micrococcus*, *Janibacter*, *Nocardioides* and *Paracoccus* (**Figure 3-6**). Our results differ from a previous characterisation of the NPM in Canadian adults using 16s rRNA sequencing (Stearns *et al.*, 2015), which were primarily dominated by *Staphylococcus*, *Rothia*, *Streptococcus* and *Veionella*. Although *Staphylococcus* were identified in both studies, the average abundance was much higher in the Canadian study (~40% vs ~5%). 16s rRNA sequencing has also been used to characterise the NPM of healthy adults in the UK and the Netherlands (Cremers *et al.*, 2014; Haak *et al.*, 2022). These studies had more similar results to each other with major taxa, including *Corynebacterium*, *Dolosigranulum*, and *Staphylococcus*. Few of the major genera that we found in our study were also detected in the others: the UK study identified *Kocuria*, *Micrococcus* and *Cutibacterium*, while the Netherlands study found *Paracoccus*, *Moraxella* and *Cutibacterium*. It is clear from these studies that NPM composition varies dramatically between individuals and between studies, indicating that there are many factors that influence the NPM, including the

country of testing. It is therefore not surprising to find that the NPM of Malawian adults differs from those tested in the UK, Netherlands, and Canada. Identifying taxa that are shared between our study and other studies does lend some credibility to our results, however.

Using PERMANOVA we demonstrated that on average, the NPM composition of individuals was significantly different based on the type of cooking fuel they used and by their HIV status (**Table 4**). We found that on average, the viral portion of the microbiome was much higher in controls than cases, and this appeared to be driven by an expansion of many bacteriophage species in the *Pahexavirus* genus (**Figure 3-5**). All *Pahexavirus spp.* that were detected infect the bacterium *Cutibacterium acnes* (**Supplementary Table 1**). Our differential abundance testing showed statistically significant association between HIV status and *Pahexavirus* relative abundance, but not *Cutibacterium* relative abundance (**Table 7**). However, when we performed relative abundance testing using only bacterial species, we were able to detect a statistically significant association between HIV status and *Cutibacterium* relative abundance (**Table 8**); the reasons underlying this are discussed in depth in section 4.4. The differences observed in *Cutibacterium* abundance was primarily in the species *Cutibacterium acnes*, which made up on average 14% of mean bacterial relative abundance in controls and only 5% in cases (**Supplementary Table 2**).

C. acnes (previously *Propionibacterium acnes*) is best known as a resident of the human skin microbiome. While it is mostly non-pathogenic, it can illicit strong immune responses and has been implicated in the inflammatory skin condition Acne Vulgaris (Taylor, Gonzalez and Porter, 2011). Very little is known about the clinical relevance of *C. acnes* in the upper respiratory tract (URT), however it has been associated with pro-inflammatory responses and the formation of granuloma in the lungs (Werner *et al.*, 2017). Inactivated *C. acnes* is currently used in equine medicine as an immune modulator to activate respiratory macrophages. Intravenous injection of inactivated *C. acnes* is used either as a prophylactic treatment or to treat respiratory diseases alongside antibiotics (Vail *et al.*, 1990; Paillot, 2013). While little is known about the clinical relevance of *C. acnes* in the URT, the role of the microbiome in priming the immune system against pathogens has previously been established (Abt *et al.*, 2012; Belkaid and Hand, 2014; Pfeiffer and Virgin, 2016; Man, de Steenhuijsen Piters and Bogaert,

2017). *C. acnes* has been shown to be immunogenic in the skin and lungs, if this were also the case in the URT it's possible that *C. acnes* might play an important role in microbiota-mediated immune modulation. In this scenario, the decrease in *C. acnes* population that we observed in the nasopharynx of HIV-infected individuals might make them further susceptible to respiratory infections, although this hypothesis would require further research to test.

4.2 Alpha diversity analysis

We found that mean viral species richness was significantly higher in controls vs. cases (**Figure 3-10a**) and that this was driven by the presence of more *Pahexavirus* species, as control samples with high viral richness also have expanded *Pahexavirus* populations (**Figure 3-7**). Bacterial species evenness was 11% lower on average in HIV-uninfected individuals (p=0.053) (**Figure 3-10b**) and this may partially result from the increased dominance of *Cutibacterium acnes* that we observed in these samples. Contrasting with our findings in the nasopharynx, Twigg et al. found a reduction of bacterial evenness in the lung microbiome of HIV-infected individuals, however this finding was in individuals with more advanced HIV-infection than in our study (Twigg *et al.*, 2016). Low evenness is observed when one or a few taxa tend to dominate the microbial population, which potentially poses a risk of infection to the host depending on the pathogenicity of the taxa that dominates. There was increased domination of the genus *Streptococcus* in the lung microbiome of HIV-infected individuals, which might lead to increased risk of infection by *S. pneumoniae* (Twigg *et al.*, 2016). Conversely, in the nasopharynx of controls we observed increased dominance of *C. acnes*, which hasn't been implicated as a pathogen of the URT and therefore unlikely to pose an infection risk.

Measures of alpha diversity in the microbiome have been proposed to affect resilience to invasion, stability and have also been correlated with disease status (Levine and D'Antonio, 1999; Gillis, Hughes, Spiga, Winter, Zhu, de Carvalho, *et al.*, 2018; Herren and Baym, 2018). Our findings highlight that alpha diversity measures can be a useful tool for identifying differences in microbial populations between test groups. However, it is important to place these findings within the

appropriate biological context by considering which specific microbes drive alpha diversity changes and how they are relevant to the research in question.

4.3 Longitudinal changes and beta diversity analysis

Our results clearly show that an individual's NPM microbiome composition varies significantly over time. Few other studies have characterised longitudinal NPM samples. Allen *et al*. used 16s rRNA to characterise the nasopharyngeal bacteriome over time in health and in response to rhinovirus challenge (Allen *et al.*, 2014). We are unable to directly compare our measurements of NPM stability with theirs as we have used different beta diversity metrics to measure the differences between samples. However our findings are broadly similar in that the taxonomic composition of longitudinal samples varies widely with time.

As expected, we found that samples taken 8 or 9 months apart were significantly more different than samples taken 1 month apart (**Figure 3-11**), showing that the NPM composition is likely to drift further with time but that this relationship is not linear. We also found that the Bray-Curtis dissimilarity (BCD) between longitudinal sample pairs was higher in cases than controls (9% coefficient, p=0.072) (**Table 6**). This represents a relative instability of the microbiome in HIV-infected individuals compared to healthy controls. In the virome we identified an even stronger association between HIV status and stability (20.8% coefficient, p=0.017). This observation is likely explained by the differential abundance of *Pahexavirus* in cases and controls. In most control samples, the virome is dominated by large, relatively stable populations of *Pahexavirus* which reduces the BCD between samples, however these are not there in most case samples (**Figure 3-7**).

4.4 Relative abundance data and the 'curse of compositionality'

We identified many bacteriophages, belonging to the Pahexavirus genus, which infect *Cutibacterium acnes* in our study. We were able to detect a statistically significant association between HIV status and *Pahexavirus* relative abundance, but no association was found between HIV status and *Cutibacterium* relative

abundance. *Cutibacterium* makes up a larger proportion of the bacterial population in controls compared to cases (18% vs 7%), but this difference is less pronounced when considering the whole microbial population (10% vs 7%) (**Figure 3-6**). This is because the bacterial portion of the microbiome is much lower in controls than cases (57% vs 90%), which is primarily driven by differences in *Pahexavirus* abundance between these groups (**Figure 3-4**). Differential abundance analysis using only bacteria finds a highly significant association between HIV status and *Cutibacterium* abundance (**Table 8**). We therefore propose that the difference in relative abundance of *Pahexavirus* between cases and controls was so large that it actually masked other differences between the samples.

The proposed masking of compositional differences between cases and controls by *Pahexavirus* is an example of what is referred to as 'the curse of compositionality', a technical issue that can make biomarker testing very difficult (Weiss *et al.*, 2016; Knight *et al.*, 2018). The 'curse' is that for compositional data: the measured abundance of each taxon is related to the measured abundance of all other taxa, therefore an increase in the relative abundance of one taxon must result in the decrease in the relative abundance of others. To place this in the context of our data, the simplest scenario is that the absolute abundance of *Pahexavirus* has increased by more than other species, pushing down their relative abundance. In this case it would be safe to conclude that the absolute abundance of *Cutibacterium* had also increased. However, we can't know this for sure because another possible, albeit more unlikely, scenario is that the absolute abundance of *Pahexavirus* is equal in cases and controls but the absolute abundance of the other taxa are lower in cases. To summarise, a key limitation of relative abundance is that it cannot be used to determine whether a particular taxon is more or less abundant, or quantify the amount of change, between two samples (Barlow, Bogatyrev and Ismagilov, 2020).

Quantifying microbiomes by absolute abundance would make differential abundance testing more sensitive and robust, as well as improving our understanding of which taxa are positively or negatively associated with test variables. It would also enable the identification of changes that affect total microbial abundance but don't affect the relative proportions of taxa, an effect that has already been observed in the gut microbiome of mice following a

ketogenic diet (Barlow, Bogatyrev and Ismagilov, 2020). Spike-ins are one way of estimating absolute abundances, where a known concentration of specific DNA/RNA is added to a sample that provides a known reference point around which to convert relative abundances to absolutes (Barlow, Bogatyrev and Ismagilov, 2020). We initially tried to use DNA spike-ins with our data, however we couldn't detect them after sequencing, highlighting a technical difficulty in choosing an appropriate concentration of DNA for the spike-in.

4.5 Types of cooking fuel used and their effect on the microbiome

The combustion of biomass fuels for cooking and heating, including firewood and charcoal, produces particulate matter that is referred to as Household Air Pollution (HAP) (Torres-Duque *et al.*, 2008). HAP is associated with the risk of chronic obstructive pulmonary disease (COPD) in women and pneumonia in children (Smith, Mehta and Maeusezahl-feuz, 2004). 16s rRNA profiling of the lung microbiome in Malawian adults found that exposure to HAP was associated with increased abundance of *Neisseria* and *Streptococcus* (Rylance *et al.*, 2016). In the nasopharynx of Ghanaian infants, HAP from cooking fires (compared to gas stoves) was associated with higher prevalence of potential pathogens *Haemophilus influenzae*, *Streptococcus pneumoniae* and *Moraxella catarrhalis* (Carrión *et al.*, 2019). No studies to date have compared the effects of firewood vs charcoal use on the respiratory microbiome. Charcoal burning is cleaner than firewood, particulate matter concentrations are 90% lower in houses using charcoal than those using open fires (Bailis, Ezzati and Kammen, 2003).

Our results show that longitudinal NPM composition is more stable in individuals using firewood as fuel rather than charcoal (**Figure 3-11**). This result seems counterintuitive as we expected that high HAP associated with firewood might lead to dysbiosis and instability of the NPM. However there is a high level of collinearity between sex and cooking fuel usage in our data (3/4 men use firewood and 10/12 women use charcoal) (**Table 2**). In Ethiopia and Uganda, women are exposed to 5-10x the amount of HAP than men because they spend more time indoors and more time cooking than men (Okello, Devereux and Semple, 2018). It is therefore difficult to attribute any of the changes in

composition or stability that we observed to the usage of firewood or charcoal as these changes could also result from differences in sex and the degree of exposure to cooking fuel.

4.6 Issues with RNA samples

The number of reads in the RNA dataset associated with species that have RNA genomes (RNA viruses) was very low, so we opted to focus our analysis on the DNA samples. Despite using a sequencing depth of 10 million reads, we found fewer than 100 reads from RNA viruses in most of our samples. 94.25% of the total sequenced reads were human, 3.09% were bacterial, 2.54% were unclassified, 0.10% were non-human eukaryotes and 0.02% were viral (**Supplementary Table 3**). Of the viral reads, less than 8% came from RNA viruses. As the vast majority of the RNA reads sequenced were of human origin, the crucial improvement for future studies would be more efficient removal of human cells and RNAs prior to sequencing, which is discussed in section 4.7.

Nevertheless, even with effective removal of human contamination, viruses only make up 0.35% of microbial RNA reads. Many microbial RNA reads will be from bacterial/eukaryote ribosomal RNA (rRNA) and messenger RNA (mRNA). rRNA depletion kits have previously been used in metagenomics to improve RNA virus recovery (Manso, Bibby and Mbisa, 2017; Fitzpatrick *et al.*, 2021), but this is an expensive additional step and did not improve RNA virus recovery in our limited tests (data not shown), so we opted not to include this step for our samples.

Our poor recovery of RNA viruses demonstrates that a successful characterisation of the RNA microbiome of the respiratory tract would require additional steps to enrich for viruses or better deplete non-viral RNAs. Samples can be enriched for viruses by Tangential-flow filtering, where a fine mesh is used to physically filter out cells (Thurber *et al.*, 2009). Ultracentrifugation can also be used to separate cells from viruses (Vibin *et al.*, 2018). It may also be necessary to amplify the remaining viral RNAs after extraction, as the yield may not actually be high enough for a sequencing run, however this can also introduce bias (Thurber *et al.*, 2009). Taken together, these steps could drastically improve RNA virus recovery, but they will also introduce biases. Some viruses can be lost during

filtering steps due to their size/shape (Thurber *et al.*, 2009). Furthermore, any intracellular viruses will be lost when cells are filtered.

It's also worth considering whether the benefit of sequencing RNA viruses outweighs the cost for the study in question. The price of preparing and sequencing DNA and RNA samples is similar, but the DNA samples contain information about populations of bacteria, archaea, eukaryotes, and DNA viruses whereas the RNA samples only measure RNA viral populations and their abundances can't be directly compared to microbial populations from DNA samples. Previous 16s rRNA sequencing studies in the respiratory tract microbiome have also used multiplex PCR to analyse the presence/absence of respiratory viruses (Teo *et al.*, 2015; Haak *et al.*, 2022). Using multiplex PCR to confirm the presence/absence of clinically relevant RNA viruses might be a much more cost-effective solution than shotgun RNA sequencing.

4.7 The feasibility of shotgun metagenomics in the respiratory tract

The nasopharynx has a relatively low microbial density (Man, de Steenhuijsen Piters and Bogaert, 2017), so we knew that host contamination would be a significant issue if we didn't take steps to deplete human DNA from our samples prior to sequencing. We attempted to deplete human cells and DNA from the sample in 2 stages. Prior to DNA extraction, we centrifuged samples at 5000g for 1 minutes to pellet eukaryotic cells and separated the supernatant (Nelson *et al.*, 2019). After DNA extraction, we used the NEBNext Microbiome DNA Enrichment Kit to remove CpG-methylated host DNA prior to library preparation (Feehery *et al.*, 2013).

Despite the steps we took to reduce human contamination, 98.2% of the reads we sequenced were of human origin (**Supplementary Table 4**). 1.17% of all reads were microbial and classified at the species level with an average of 114595 reads per sample. However, the number of species-level microbial reads varied drastically by sample, ranging from 235 to 705314 reads (**Table 2**). Consequently, when normalising for sequencing depth in our samples, we had to

rarefy samples to a very low read count of 4026; while this did make our samples comparable it also resulted in a large loss of data for downstream analysis.

The rarefaction curves that we generated show that by rarefying our samples, we fail to detect hundreds of microbial species (**Figure 3-1**). Furthermore, our detection is biased against viruses. Viral genomes can be hundreds or thousands of times smaller than microbes, meaning there is less viral DNA per single replicating unit so they must be present at higher densities than other microbes to meet our detection threshold. For example: viruses made up 0.39% of nonhuman reads in our dataset but they accounted for 43% of the total relative abundance in controls and 9% in cases (**Figure 3-4**). Overall we detected very few viral species in our samples, viral detection could be improved by recovering more microbial reads through more effectively removing human DNA prior to sequencing. This is supported by the observation that average viral richness was doubled in our unrarefied samples compared to the rarefied samples we used for analysis (**Supplementary Table 5**).

While human reads dominated all the samples, there were drastic differences in the number of species-level microbial reads per sample. The source of this variation is unclear, one possibility is that the proportions of human and microbial reads reflect true differences in the absolute abundance of the nasopharyngeal microbiome. In reality this is unlikely to be the only factor, we observed differences in the human:microbial read ratio of up to 100-fold in samples taken 1 month apart in the same individuals. This extreme level of variation would not be expected in such a short time and in the absence of a major microbiome perturbation, such as antibiotic usage or URT infection, which were actively tracked as part of the study. Another possible factor is variation in technique during collection of the nasopharyngeal swab, this hasn't been tested for shotgun sequencing, but swabbing technique did not significantly affect nasopharyngeal samples used for 16s rRNA sequencing or PCR (Akmatov *et al.*, 2017; Kinloch *et al.*, 2020). Sample storage is another factor that could contribute to a higher ratio of human:microbial reads, DNA/RNA degradation can be observed in frozen samples that thawed for as little as 1 hour (Cardona *et al.*, 2012). Repeated freeze-thaw cycles can also affect the DNA of microbes differently, eukaryotic cells were robust to freeze-thaw cycles whereas DNA

degradation was observed in gram-negative bacteria (Poulsen *et al.*, 2021). While we took great care to keep our samples frozen, the samples used in our study were originally stored in Malawi from a previous study and the number of freeze-thaw cycles is unknown. Samples also had to be transported from Malawi to the UK over 2 days on dry ice, therefore a degree of sample thawing and DNA degradation is a possibility.

While we can't pinpoint exactly why some samples have a higher proportion of human reads than others, there was an extremely high level of human contamination across all samples. Taking further steps to reduce the amount of human DNA/RNA in our samples would i) prevent wasted sequencing effort on human reads, ii) improve our ability to detect species with low abundance and/or small genomes and iii) make our overall microbiome characterisations more accurate. A significant source of human DNA/RNA in microbiome samples is extracellular (Nelson *et al.*, 2019). In our protocol we opted not to DNAse/RNAse treat our samples prior to nucleic acid extraction as this can degrade some viruses (Thurber *et al.*, 2009). In hindsight we would recommend degrading extracellular nucleic acids prior to nucleic acid extraction. There are several commercial kits available for human DNA depletion including MolYsis and benzonase1 that involve the targeted lysis of eukaryotic cells and endonuclease degradation of extracellular DNA (Nelson *et al.*, 2019). Further testing would be required to see how well these kits improve host depletion when used in combination with the NEBNext Microbiome DNA Enrichment Kit that we deployed in our study.

4.8 Future Directions

Our small pilot study used shotgun metagenomics to compare the NPM of HIV-infected and HIV-uninfected adults. We successfully deployed DNA shotgun sequencing in the nasopharynx and identified compositional differences in the NPM associated with HIV status, however our study had several limitations. First, an overwhelming majority of sequencing reads were of human origin and the number of microbial reads recovered varied widely between samples. Our analysis was therefore conducted at a much shallower sequencing depth than planned, reducing our ability to detect species present at lower abundances

and/or those with smaller genomes, particularly viruses. Second, the compositional nature of sequencing data makes it difficult to detect relationships between microbes and sample metadata. It also restricted any conclusions to be about the relative abundances of taxa between groups, not their true abundances. Finally, due to limited sample size and the presence of multiple variables that could affect NPM composition, we could only perform limited testing and detect relationships with very large effect sizes.

We believe our finding that the NPM differs by HIV status shows that this is a promising area for further research. Future studies should be able to expand upon our findings if they are able to i) further deplete human contamination, ii) quantify absolute microbial abundances and iii) increase sample size.

Appendix 1

Supplementary Table 1 – *Pahexavirus* bacteriophage species identified in our samples and their bacterial hosts as annotated in the Virus-Host database (Mihara *et al.*, 2016).

Virus name	Virus tax id	Other names	Host name
Pahexavirus ATCC29399BC	1229794	Propionibacterium phage ATCC29399B_C	Cutibacterium acnes
Pahexavirus P100A	1229790	Propionibacterium phage P100_A	Cutibacterium acnes
Pahexavirus P100D	1229789	Propionibacterium phage P100D	Cutibacterium acnes
Pahexavirus P144	1229784	Propionibacterium phage P14	Cutibacterium acnes
Pahexavirus P91	1229782	Propionibacterium phage P9.1	Cutibacterium acnes
Pahexavirus PA6	376758	Propionibacterium phage PA6	Cutibacterium acnes
Pahexavirus PAD20	504501	Propionibacterium phage PAD20	Cutibacterium acnes
Pahexavirus PAS50	504553	Propionibacterium phage PAS50	Cutibacterium acnes
Pahexavirus PHL025M00	1500799	Propionibacterium phage PHL025M00	Cutibacterium acnes
Pahexavirus PHL041M10	1500801	Propionibacterium phage PHL041M10	Cutibacterium acnes
Pahexavirus PHL060L00	1235647	Propionibacterium phage PHL060L00	Cutibacterium acnes
Pahexavirus PHL070N00	1500807	Propionibacterium phage PHL070N00	Cutibacterium acnes
Pahexavirus PHL071N05	1235650	Propionibacterium phage PHL071N05	Cutibacterium acnes
Pahexavirus PHL092M00	1500813	Propionibacterium phage PHL092M00	Cutibacterium acnes
Pahexavirus PHL095N00	1500814	Propionibacterium phage PHL095N00	Cutibacterium acnes
Pahexavirus PHL114L00	1235656	Propionibacterium phage PHL114L00	Cutibacterium acnes
Pahexavirus PHL114L00	1500815	Propionibacterium phage PHL114N00	Cutibacterium acnes
Pahexavirus PHL116M00	1500816	Propionibacterium phage PHL116M00	Cutibacterium acnes
Pahexavirus PHL116M00	1500817	Propionibacterium phage PHL116M10	Cutibacterium acnes
Pahexavirus PHL132N00	1500820	Propionibacterium phage PHL132N00	Cutibacterium acnes
Pahexavirus PHL141N00	1500821	Propionibacterium phage PHL141N00	Cutibacterium acnes
Pahexavirus PHL152M00	1500825	Propionibacterium phage PHL152M00	Cutibacterium acnes
Pahexavirus PHL171M01	1500827	Propionibacterium phage PHL171M01	Cutibacterium acnes
Pahexavirus PHL179M00	1500828	Propionibacterium phage PHL179M00	Cutibacterium acnes
Pahexavirus PHL199M00	1500830	Propionibacterium phage PHL199M00	Cutibacterium acnes
Pahexavirus PHL301M00	1500831	Propionibacterium phage PHL301M00	Cutibacterium acnes
Pahexavirus SKKY	1655020	Propionibacterium phage SKKY	Cutibacterium acnes
Pahexavirus attacne	1655012		Cutibacterium acnes
Pahexavirus lauchelly	1655015		Cutibacterium acnes
Pahexavirus ouroboros	1655017		Cutibacterium acnes
Pahexavirus P11	1229792		Cutibacterium acnes
Pahexavirus procrass1	1655019		Cutibacterium acnes
Pahexavirus solid	1655021		Cutibacterium acnes
Pahexavirus stormborn	1655022		Cutibacterium acnes
Pahexavirus wizzo	1655023		Cutibacterium acnes
unclassified Pahexavirus	1654740		Cutibacterium acnes
unclassified Pahexavirus	1654780		Cutibacterium acnes
unclassified Pahexavirus	1747271		Cutibacterium acnes
unclassified Pahexavirus	1690805		Cutibacterium acnes

Appendix 2

Supplementary Table 2 – The bacterial relative abundance and total relative abundance of the top 10 most abundant bacterial species in cases and controls.

	Relative abundance (bacterial species)		Relative abundance (all species)	
Bacterial species	**Case**	**Control**	**Case**	**Control**
Corynebacterium segmentosum	0.198	0.206	0.183	0.135
Cutibacterium acnes	0.052	0.173	0.045	0.081
Micrococcus luteus	0.055	0.05	0.051	0.027
Corynebacterium propinquum	0.059	0.012	0.054	0.009
Staphylococcus warneri	0.046	0.032	0.040	0.012
Cutibacterium granulosum	0.021	0.036	0.019	0.017
Corynebacterium macginleyi	0.026	0.024	0.024	0.017
Kocuria palustris	0.022	0.025	0.019	0.017
Corynebacterium tuberculostearicum	0.014	0.039	0.012	0.017
Staphylococcus aureus	0.023	0.000	0.021	0.000

Appendix 3

Supplementary Table 3 – Summary of read classifications for the RNA sequencing dataset

Classification		Reads	Mean reads per sample	Percentage of nonhuman reads	Percentage of total reads
Total		480000000	10000000		100
Human		452387892	9424748		94.25
Non-human		27612108	575252	100	5.75
Unclassified		12178458	253718	44.11	2.54
Microbial		15433650	321534	55.89	3.22
	Archaea	8227	171	0.03	0
	Bacteria	14847414	309321	53.77	3.09
	Eukaryotes	474605	9888	1.72	0.1
	Viruses	95983	2000	0.35	0.02
	RNA Viruses	7421	155	0.03	0

Appendix 4

Supplementary Table 4 – Summary of read classifications for the DNA sequencing dataset

Classification		Reads	Mean reads per sample	Percentage of nonhuman reads	Percentage of total reads
Total		480000000	10000000		100
Human		471493781	9822787		98.23
Non-human		8506219	177213	100	1.77
Unclassified		2617897	54540	30.78	0.55
Microbial		5888322	122673	69.22	1.23
	Archaea	5353	112	0.06	0
	Bacteria	5768679	120181	67.82	1.2
	Eukaryotes	81042	1688	0.95	0.02
	Viruses	33248	693	0.39	0.01

Appendix 5

Supplementary Table 5 – Average species richness of each microbial superkingdom for cases and controls in the rarefied data and unrarefied data.

	Richness (Rarefied data)			Richness (Raw data)		
Superkingdom	**Case**	**Control**	**All**	**Case**	**Control**	**All**
Archaea	0	0	0	0	0	0
Bacteria	620	585	608	1211	1330	1252
Eukaryote	4	5	4	11	13	11
Virus	3	21	9	11	33	19

Bibliography

Abt, M.C. *et al.* (2012) 'Commensal Bacteria Calibrate the Activation Threshold of Innate Antiviral Immunity', *Immunity*, 37(1), pp. 158–170. Available at: https://doi.org/10.1016/j.immuni.2012.04.011.

Akmatov, M.K. *et al.* (2017) 'Determination of nasal and oropharyngeal microbiomes in a multicenter population-based study – findings from Pretest 1 of the German National Cohort', *Scientific Reports*, 7(1), p. 1855. Available at: https://doi.org/10.1038/s41598-017-01212-6.

Ali, N. *et al.* (2017) 'Current Nucleic Acid Extraction Methods and Their Implications to Point-of-Care Diagnostics', *BioMed Research International*, 2017, p. 9306564. Available at: https://doi.org/10.1155/2017/9306564.

Allen, E.K. *et al.* (2014) 'Characterization of the nasopharyngeal microbiota in health and during rhinovirus challenge', *Microbiome*, 2, p. 22. Available at: https://doi.org/10.1186/2049-2618-2-22.

Alneberg, J. *et al.* (2014) 'Binning metagenomic contigs by coverage and composition', *Nature Methods*, 11(11), pp. 1144–1146. Available at: https://doi.org/10.1038/nmeth.3103.

Anderson, M.J. (2017) 'Permutational Multivariate Analysis of Variance (PERMANOVA)', in *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, Ltd, pp. 1–15. Available at: https://doi.org/10.1002/9781118445112.stat07841.

Armbruster, C.E. *et al.* (2010) 'Indirect pathogenicity of Haemophilus influenzae and Moraxella catarrhalis in polymicrobial otitis media occurs via interspecies quorum signaling', *mBio*, 1(3), pp. e00102-10. Available at: https://doi.org/10.1128/mBio.00102-10.

Aston, S.J. *et al.* (2019) 'Etiology and Risk Factors for Mortality in an Adult Community-acquired Pneumonia Cohort in Malawi', *American Journal of Respiratory and Critical Care Medicine*, 200(3), pp. 359–369. Available at: https://doi.org/10.1164/rccm.201807-1333OC.

Ayling, M., Clark, M.D. and Leggett, R.M. (2019) 'New approaches for metagenome assembly with short reads', *Briefings in Bioinformatics* [Preprint]. Available at: https://doi.org/10.1093/bib/bbz020.

Bailey, M.J. *et al.* (2020) 'Exposure to air pollutants and the gut microbiota: a potential link between exposure, obesity, and type 2 diabetes', *Gut Microbes*, 11(5), pp. 1188–1202. Available at: https://doi.org/10.1080/19490976.2020.1749754.

Bailis, R., Ezzati, M. and Kammen, D.M. (2003) 'Greenhouse Gas Implications of Household Energy Technology in Kenya', *Environmental Science & Technology*, 37(10), pp. 2051–2059. Available at: https://doi.org/10.1021/es026058q.

Banos, S. *et al.* (2018) 'A comprehensive fungi-specific 18S rRNA gene sequence primer toolkit suited for diverse research issues and sequencing platforms', *BMC*

Microbiology, 18(1), p. 190. Available at: https://doi.org/10.1186/s12866-018-1331-4.

Barlow, J.T., Bogatyrev, S.R. and Ismagilov, R.F. (2020) 'A quantitative sequencing framework for absolute abundance measurements of mucosal and lumenal microbial communities', *Nature Communications*, 11(1), p. 2590. Available at: https://doi.org/10.1038/s41467-020-16224-6.

Barr, J.J. *et al.* (2013) 'Bacteriophage adhering to mucus provide a non-host-derived immunity', *Proceedings of the National Academy of Sciences*, 110(26), pp. 10771–10776. Available at: https://doi.org/10.1073/pnas.1305923110.

Bassis, C.M. *et al.* (2015) 'Analysis of the Upper Respiratory Tract Microbiotas as the Source of the Lung and Gastric Microbiotas in Healthy Individuals', *mBio*, 6(2), pp. e00037-15. Available at: https://doi.org/10.1128/mBio.00037-15.

Bazinet, A.L. *et al.* (2018) 'BLAST-based validation of metagenomic sequence assignments', *PeerJ*, 6, p. e4892. Available at: https://doi.org/10.7717/peerj.4892.

Beck, J.M. *et al.* (2015) 'Multicenter Comparison of Lung and Oral Microbiomes of HIV-infected and HIV-uninfected Individuals', *American Journal of Respiratory and Critical Care Medicine*, 192(11), pp. 1335–1344. Available at: https://doi.org/10.1164/rccm.201501-0128OC.

Belkaid, Y. and Hand, T. (2014) 'Role of the Microbiota in Immunity and inflammation', *Cell*, 157(1), pp. 121–141. Available at: https://doi.org/10.1016/j.cell.2014.03.011.

Bolger, A.M., Lohse, M. and Usadel, B. (2014) 'Trimmomatic: a flexible trimmer for Illumina sequence data', *Bioinformatics*, 30(15), pp. 2114–2120. Available at: https://doi.org/10.1093/bioinformatics/btu170.

Bolyen, E. *et al.* (2018) *QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science*. e27295v2. PeerJ Inc. Available at: https://doi.org/10.7287/peerj.preprints.27295v2.

Breitwieser, F.P. *et al.* (2019) 'Human contamination in bacterial genomes has created thousands of spurious proteins', *Genome Research*, 29(6), pp. 954–960. Available at: https://doi.org/10.1101/gr.245373.118.

Buffie, C.G. and Pamer, E.G. (2013) 'Microbiota-mediated colonization resistance against intestinal pathogens', *Nature Reviews Immunology*, 13(11), pp. 790–801. Available at: https://doi.org/10.1038/nri3535.

Bush, S.J. *et al.* (2020) 'Evaluation of methods for detecting human reads in microbial sequencing datasets', *Microbial Genomics*, 6(7), p. mgen000393. Available at: https://doi.org/10.1099/mgen.0.000393.

Cardona, S. *et al.* (2012) 'Storage conditions of intestinal microbiota matter in metagenomic analysis', *BMC Microbiology*, 12, p. 158. Available at: https://doi.org/10.1186/1471-2180-12-158.

Carrión, D. *et al.* (2019) 'Examining the relationship between household air pollution and infant microbial nasal carriage in a Ghanaian cohort', *Environment International*, 133, p. 105150. Available at: https://doi.org/10.1016/j.envint.2019.105150.

'CDC: AIDS and Opportunistic Infections' (2021). Centers for Disease Control and Prevention. Available at: https://www.cdc.gov/hiv/basics/livingwithhiv/opportunisticinfections.html (Accessed: 22 September 2022).

Cleary, D.W. and Clarke, S.C. (2017) 'The nasopharyngeal microbiome', *Emerging Topics in Life Sciences*, 1(4), pp. 297–312. Available at: https://doi.org/10.1042/ETLS20170041.

Cohen, C. *et al.* (2013) 'Severe Influenza-associated Respiratory Infection in High HIV Prevalence Setting, South Africa, 2009–2011', *Emerging Infectious Diseases*, 19(11), pp. 1766–1774. Available at: https://doi.org/10.3201/eid1911.130546.

Cremers, A.J. *et al.* (2014) 'The adult nasopharyngeal microbiome as a determinant of pneumococcal acquisition', *Microbiome*, 2, p. 44. Available at: https://doi.org/10.1186/2049-2618-2-44.

Cribbs, S.K. *et al.* (2016) 'Correlation of the lung microbiota with metabolic profiles in bronchoalveolar lavage fluid in HIV infection', *Microbiome*, 4. Available at: https://doi.org/10.1186/s40168-016-0147-4.

Cui, L. *et al.* (2015) 'Topographic Diversity of the Respiratory Tract Mycobiome and Alteration in HIV and Lung Disease', *American Journal of Respiratory and Critical Care Medicine*, 191(8), pp. 932–942. Available at: https://doi.org/10.1164/rccm.201409-1583OC.

Deeks, S.G. *et al.* (2015) 'HIV infection', *Nature Reviews Disease Primers*, 1(1), pp. 1–22. Available at: https://doi.org/10.1038/nrdp.2015.35.

DeGruttola, A.K. *et al.* (2016) 'Current understanding of dysbiosis in disease in human and animal models', *Inflammatory bowel diseases*, 22(5), pp. 1137–1150. Available at: https://doi.org/10.1097/MIB.0000000000000750.

Dickson, R.P. *et al.* (2015) 'Spatial Variation in the Healthy Human Lung Microbiome and the Adapted Island Model of Lung Biogeography', *Annals of the American Thoracic Society*, 12(6), pp. 821–830. Available at: https://doi.org/10.1513/AnnalsATS.201501-029OC.

Dickson, R.P. *et al.* (2016) 'The Microbiome and the Respiratory Tract', *Annual Review of Physiology*, 78(1), pp. 481–504. Available at: https://doi.org/10.1146/annurev-physiol-021115-105238.

Feehery, G.R. *et al.* (2013) 'A method for selectively enriching microbial DNA from contaminating vertebrate host DNA', *PloS One*, 8(10), p. e76096. Available at: https://doi.org/10.1371/journal.pone.0076096.

Feikin, D.R. *et al.* (2004) 'Global strategies to prevent bacterial pneumonia in adults with HIV disease', *The Lancet. Infectious Diseases*, 4(7), pp. 445–455. Available at: https://doi.org/10.1016/S1473-3099(04)01060-6.

Fitzpatrick, A.H. *et al.* (2021) 'High Throughput Sequencing for the Detection and Characterization of RNA Viruses', *Frontiers in Microbiology*, 12, p. 621719. Available at: https://doi.org/10.3389/fmicb.2021.621719.

Frank, D.N. *et al.* (2007) 'Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases', *Proceedings of the National Academy of Sciences of the United States of America*, 104(34), pp. 13780–13785. Available at: https://doi.org/10.1073/pnas.0706625104.

Gilbert, J.A. *et al.* (2018) 'Current understanding of the human microbiome', *Nature Medicine*, 24(4), pp. 392–400. Available at: https://doi.org/10.1038/nm.4517.

Gill, S.R. *et al.* (2006) 'Metagenomic Analysis of the Human Distal Gut Microbiome', *Science*, 312(5778), pp. 1355–1359. Available at: https://doi.org/10.1126/science.1124234.

Gillis, C.C., Hughes, E.R., Spiga, L., Winter, M.G., Zhu, W., Carvalho, T.F. de, *et al.* (2018) 'Dysbiosis-Associated Change in Host Metabolism Generates Lactate to Support Salmonella Growth', *Cell Host & Microbe*, 23(1), pp. 54-64.e6. Available at: https://doi.org/10.1016/j.chom.2017.11.006.

Gillis, C.C., Hughes, E.R., Spiga, L., Winter, M.G., Zhu, W., de Carvalho, T.F., *et al.* (2018) 'Dysbiosis-associated change in host metabolism generates lactate to support Salmonella growth', *Cell host & microbe*, 23(1), pp. 54-64.e6. Available at: https://doi.org/10.1016/j.chom.2017.11.006.

'Global HIV & AIDS statistics — Fact sheet' (2022). UNAIDS. Available at: https://www.unaids.org/sites/default/files/media_asset/UNAIDS_FactSheet_en.pdf.

Gloor, G.B. *et al.* (2017) 'Microbiome Datasets Are Compositional: And This Is Not Optional', *Frontiers in Microbiology*, 8. Available at: https://doi.org/10.3389/fmicb.2017.02224.

Goodrich, J.K. *et al.* (2014) 'Conducting a Microbiome Study', *Cell*, 158(2), pp. 250–262. Available at: https://doi.org/10.1016/j.cell.2014.06.037.

Haak, B.W. *et al.* (2022) 'Bacterial and Viral Respiratory Tract Microbiota and Host Characteristics in Adults With Lower Respiratory Tract Infections: A Case-Control Study', *Clinical Infectious Diseases*, 74(5), pp. 776–784. Available at: https://doi.org/10.1093/cid/ciab568.

Hanada, S. *et al.* (2018) 'Respiratory Viral Infection-Induced Microbiome Alterations and Secondary Bacterial Pneumonia', *Frontiers in Immunology*, 9. Available at: https://doi.org/10.3389/fimmu.2018.02640.

Herren, C. and Baym, M. (2018) 'Stronger connectivity of the resident gut microbiome lends resistance to invading bacteria'. Available at: https://doi.org/10.1101/261750.

Ho, A. *et al.* (2018) 'Impact of Human Immunodeficiency Virus on the Burden and Severity of Influenza Illness in Malawian Adults: A Prospective Cohort and

Parallel Case-Control Study', *Clinical Infectious Diseases*, 66(6), pp. 865–876. Available at: https://doi.org/10.1093/cid/cix903.

Jain, S. *et al.* (2015) 'Community-Acquired Pneumonia Requiring Hospitalization among U.S. Adults', *New England Journal of Medicine*, 373(5), pp. 415–427. Available at: https://doi.org/10.1056/NEJMoa1500245.

Jašarević, E., Morrison, K.E. and Bale, T.L. (2016) 'Sex differences in the gut microbiome–brain axis across the lifespan', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1688), p. 20150122. Available at: https://doi.org/10.1098/rstb.2015.0122.

Jovel, J. *et al.* (2016) 'Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics', *Frontiers in Microbiology*, 7. Available at: https://doi.org/10.3389/fmicb.2016.00459.

Justiz Vaillant, A.A. and Naik, R. (2022) 'HIV-1 Associated Opportunistic Infections', in *StatPearls*. Treasure Island (FL): StatPearls Publishing. Available at: http://www.ncbi.nlm.nih.gov/books/NBK539787/ (Accessed: 30 August 2022).

Kau, A.L. *et al.* (2011) 'Human nutrition, the gut microbiome, and immune system: envisioning the future', *Nature*, 474(7351), pp. 327–336. Available at: https://doi.org/10.1038/nature10213.

Kinloch, N.N. *et al.* (2020) 'Evaluation of Nasopharyngeal Swab Collection Techniques for Nucleic Acid Recovery and Participant Experience: Recommendations for COVID-19 Diagnostics', *Open Forum Infectious Diseases*, 7(11), p. ofaa488. Available at: https://doi.org/10.1093/ofid/ofaa488.

Knight, R. *et al.* (2018) 'Best practices for analysing microbiomes', *Nature Reviews Microbiology*, 16(7), pp. 410–422. Available at: https://doi.org/10.1038/s41579-018-0029-9.

Krause, R. *et al.* (2017) 'Mycobiome in the Lower Respiratory Tract – A Clinical Perspective', *Frontiers in Microbiology*, 7. Available at: https://doi.org/10.3389/fmicb.2016.02169.

Langdon, A., Crook, N. and Dantas, G. (2016) 'The effects of antibiotics on the microbiome throughout development and alternative approaches for therapeutic modulation', *Genome Medicine*, 8, p. 39. Available at: https://doi.org/10.1186/s13073-016-0294-z.

Langmead, B. and Salzberg, S.L. (2012) 'Fast gapped-read alignment with Bowtie 2', *Nature Methods*, 9(4), pp. 357–359. Available at: https://doi.org/10.1038/nmeth.1923.

LaPierre, N. *et al.* (2019) 'MiCoP: microbial community profiling method for detecting viral and fungal organisms in metagenomic samples', *BMC Genomics*, 20(5), p. 423. Available at: https://doi.org/10.1186/s12864-019-5699-9.

Lawn, S.D. *et al.* (2009) 'Short-term and long-term risk of tuberculosis associated with CD4 cell recovery during antiretroviral therapy in South Africa',

AIDS, 23(13), pp. 1717–1725. Available at: https://doi.org/10.1097/QAD.0b013e32832d3b6d.

Lawn, S.D., Bekker, L.-G. and Wood, R. (2005) 'How effectively does HAART restore immune responses to Mycobacterium tuberculosis? Implications for tuberculosis control', *AIDS*, 19(11), pp. 1113–1124. Available at: https://doi.org/10.1097/01.aids.0000176211.08581.5a.

Levine, J.M. and D'Antonio, C.M. (1999) 'Elton Revisited: A Review of Evidence Linking Diversity and Invasibility', *Oikos*, 87(1), pp. 15–26. Available at: https://doi.org/10.2307/3546992.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), pp. 2078–2079. Available at: https://doi.org/10.1093/bioinformatics/btp352.

Liu, C.M. *et al.* (2015) 'Staphylococcus aureus and the ecology of the nasal microbiome', *Science Advances*, 1(5), p. e1400216. Available at: https://doi.org/10.1126/sciadv.1400216.

Lok, J.J. *et al.* (2010) 'Long-term increase in CD4+ T-cell counts during combination antiretroviral therapy for HIV-1 infection', *AIDS (London, England)*, 24(12), pp. 1867–1876. Available at: https://doi.org/10.1097/QAD.0b013e32833adbcf.

Love, M.I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*, 15(12), p. 550. Available at: https://doi.org/10.1186/s13059-014-0550-8.

Mallick, H. *et al.* (2021) 'Multivariable association discovery in population-scale meta-omics studies', *PLOS Computational Biology*, 17(11), p. e1009442. Available at: https://doi.org/10.1371/journal.pcbi.1009442.

Man, W.H., de Steenhuijsen Piters, W.A.A. and Bogaert, D. (2017) 'The microbiota of the respiratory tract: gatekeeper to respiratory health', *Nature Reviews Microbiology*, 15(5), pp. 259–270. Available at: https://doi.org/10.1038/nrmicro.2017.14.

Mandal, S. *et al.* (2015) 'Analysis of composition of microbiomes: a novel method for studying microbial composition', *Microbial Ecology in Health and Disease*, 26. Available at: https://doi.org/10.3402/mehd.v26.27663.

Manso, C.F., Bibby, D.F. and Mbisa, J.L. (2017) 'Efficient and unbiased metagenomic recovery of RNA virus genomes from human plasma samples', *Scientific Reports*, 7, p. 4173. Available at: https://doi.org/10.1038/s41598-017-02239-5.

McCullers, J.A. (2014) 'The co-pathogenesis of influenza viruses with bacteria in the lung', *Nature Reviews Microbiology*, 12(4), pp. 252–262. Available at: https://doi.org/10.1038/nrmicro3231.

McIntyre, A.B.R. *et al.* (2017) 'Comprehensive benchmarking and ensemble approaches for metagenomic classifiers', *Genome Biology*, 18(1), p. 182. Available at: https://doi.org/10.1186/s13059-017-1299-7.

McKnight, D.T. *et al.* (2019) 'Methods for normalizing microbiome data: An ecological perspective', *Methods in Ecology and Evolution*, 10(3), pp. 389–400. Available at: https://doi.org/10.1111/2041-210X.13115.

McMurdie, P.J. and Holmes, S. (2014) 'Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible', *PLOS Computational Biology*, 10(4), p. e1003531. Available at: https://doi.org/10.1371/journal.pcbi.1003531.

Menzel, P., Ng, K.L. and Krogh, A. (2016) 'Fast and sensitive taxonomic classification for metagenomics with Kaiju', *Nature Communications*, 7, p. 11257. Available at: https://doi.org/10.1038/ncomms11257.

Meyer, F. *et al.* (2019) 'Assessing taxonomic metagenome profilers with OPAL', *Genome Biology*, 20(1), p. 51. Available at: https://doi.org/10.1186/s13059-019-1646-y.

Mihara, T. *et al.* (2016) 'Linking Virus Genomes with Host Taxonomy', *Viruses*, 8(3), p. 66. Available at: https://doi.org/10.3390/v8030066.

Milanese, A. *et al.* (2019) 'Microbial abundance, activity and population genomic profiling with mOTUs2', *Nature Communications*, 10(1), p. 1014. Available at: https://doi.org/10.1038/s41467-019-08844-4.

Morgulis, A. *et al.* (2008) 'Database indexing for production MegaBLAST searches', *Bioinformatics (Oxford, England)*, 24(16), pp. 1757–1764. Available at: https://doi.org/10.1093/bioinformatics/btn322.

Nayfach, S. and Pollard, K.S. (2016) 'Toward Accurate and Quantitative Comparative Metagenomics', *Cell*, 166(5), pp. 1103–1116. Available at: https://doi.org/10.1016/j.cell.2016.08.007.

NCBI Resource Coordinators (2018) 'Database resources of the National Center for Biotechnology Information', *Nucleic Acids Research*, 46(D1), pp. D8–D13. Available at: https://doi.org/10.1093/nar/gkx1095.

Nelson, M.T. *et al.* (2019) 'Human and Extracellular DNA Depletion for Metagenomic Analysis of Complex Clinical Infection Samples Yields Optimized Viable Microbiome Profiles', *Cell reports*, 26(8), pp. 2227-2240.e5. Available at: https://doi.org/10.1016/j.celrep.2019.01.091.

Ng, K.M. *et al.* (2013) 'Microbiota-liberated host sugars facilitate post-antibiotic expansion of enteric pathogens', *Nature*, 502(7469), pp. 96–99. Available at: https://doi.org/10.1038/nature12503.

Norman, J.M., Handley, S.A. and Virgin, H.W. (2014) 'Kingdom-Agnostic Metagenomics and the Importance of Complete Characterization of Enteric Microbial Communities', *Gastroenterology*, 146(6), pp. 1459–1469. Available at: https://doi.org/10.1053/j.gastro.2014.02.001.

Nurk, S. *et al.* (2022) 'The complete sequence of a human genome', *Science*, 376(6588), pp. 44–53. Available at: https://doi.org/10.1126/science.abj6987.

Okello, G., Devereux, G. and Semple, S. (2018) 'Women and girls in resource poor countries experience much greater exposure to household air pollutants

than men: Results from Uganda and Ethiopia', *Environment International*, 119, pp. 429–437. Available at: https://doi.org/10.1016/j.envint.2018.07.002.

Oksanen J. et al. (2022) '_vegan: Community Ecology Package_. R package version 2.6-2'.

O'Leary, N.A. *et al.* (2016) 'Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation', *Nucleic Acids Research*, 44(D1), pp. D733-745. Available at: https://doi.org/10.1093/nar/gkv1189.

Oliphant, K. and Allen-Vercoe, E. (2019) 'Macronutrient metabolism by the human gut microbiome: major fermentation by-products and their impact on host health', *Microbiome*, 7(1), p. 91. Available at: https://doi.org/10.1186/s40168-019-0704-8.

Ondov, B.D., Bergman, N.H. and Phillippy, A.M. (2011) 'Interactive metagenomic visualization in a Web browser', *BMC Bioinformatics*, 12(1), p. 385. Available at: https://doi.org/10.1186/1471-2105-12-385.

Paillot, R. (2013) 'A systematic review of the immune-modulators Parapoxvirus ovis and Propionibacterium acnes for the prevention of respiratory disease and other infections in the horse', *Veterinary Immunology and Immunopathology*, 153(1–2), pp. 1–9. Available at: https://doi.org/10.1016/j.vetimm.2013.01.010.

Perez, A.C. *et al.* (2014) 'Residence of Streptococcus pneumoniae and Moraxella catarrhalis within polymicrobial biofilm promotes antibiotic resistance and bacterial persistence in vivo', *Pathogens and Disease*, 70(3), pp. 280–288. Available at: https://doi.org/10.1111/2049-632X.12129.

Pfeiffer, J.K. and Virgin, H.W. (2016) 'Transkingdom control of viral infection and immunity in the mammalian intestine', *Science (New York, N.Y.)*, 351(6270). Available at: https://doi.org/10.1126/science.aad5872.

Piro, V.C., Matschkowski, M. and Renard, B.Y. (2017) 'MetaMeta: integrating metagenome analysis tools to improve taxonomic profiling', *Microbiome*, 5(1), p. 101. Available at: https://doi.org/10.1186/s40168-017-0318-y.

Poulsen, C.S. *et al.* (2021) 'Standard Sample Storage Conditions Have an Impact on Inferred Microbiome Composition and Antimicrobial Resistance Patterns', *Microbiology Spectrum*, 9(2), pp. e01387-21. Available at: https://doi.org/10.1128/Spectrum.01387-21.

Quince, C. *et al.* (2017) 'Shotgun metagenomics, from sampling to analysis', *Nature Biotechnology*, 35(9), pp. 833–844. Available at: https://doi.org/10.1038/nbt.3935.

Ridaura, V.K. *et al.* (2013) 'Gut Microbiota from Twins Discordant for Obesity Modulate Metabolism in Mice', *Science*, 341(6150), pp. 1241214–1241214. Available at: https://doi.org/10.1126/science.1241214.

van Riel, D. *et al.* (2006) 'H5N1 Virus Attachment to Lower Respiratory Tract', *Science*, 312(5772), pp. 399–399. Available at: https://doi.org/10.1126/science.1125548.

Rooks, M.G. and Garrett, W.S. (2016) 'Gut microbiota, metabolites and host immunity', *Nature reviews. Immunology*, 16(6), pp. 341–352. Available at: https://doi.org/10.1038/nri.2016.42.

Rylance, J. *et al.* (2016) 'Household air pollution and the lung microbiome of healthy adults in Malawi: a cross-sectional study', *BMC Microbiology*, 16(1), p. 182. Available at: https://doi.org/10.1186/s12866-016-0803-7.

Schenck, L.P., Surette, M.G. and Bowdish, D.M.E. (2016) 'Composition and immunological significance of the upper respiratory tract microbiota', *Febs Letters*, 590(21), pp. 3705–3720. Available at: https://doi.org/10.1002/1873-3468.12455.

Schneider, V.A. *et al.* (2016) 'Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly'. bioRxiv, p. 072116. Available at: https://doi.org/10.1101/072116.

Scott, J. *et al.* (2000) 'Aetiology, outcome, and risk factors for mortality among adults with acute pneumonia in Kenya', *The Lancet*, 355(9211), pp. 1225–1230. Available at: https://doi.org/10.1016/S0140-6736(00)02089-4.

Sczyrba, A. *et al.* (2017) 'Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software', *Nature Methods*, 14(11), pp. 1063–1071. Available at: https://doi.org/10.1038/nmeth.4458.

Seabold, S. and Perktold, J. (2010) 'Statsmodels: Econometric and Statistical Modeling with Python', in. *Python in Science Conference*, Austin, Texas, pp. 92–96. Available at: https://doi.org/10.25080/Majora-92bf1922-011.

Segal, L.N. *et al.* (2013) 'Enrichment of lung microbiome with supraglottic taxa is associated with increased pulmonary inflammation', *Microbiome*, 1(1). Available at: https://doi.org/10.1186/2049-2618-1-19.

Selva, L. *et al.* (2009) 'Killing niche competitors by remote-control bacteriophage induction', *Proceedings of the National Academy of Sciences of the United States of America*, 106(4), pp. 1234–1238. Available at: https://doi.org/10.1073/pnas.0809600106.

Smith, A.M. and McCullers, J.A. (2014) 'Secondary Bacterial Infections in Influenza Virus Infection Pathogenesis', *Influenza Pathogenesis and Control - Volume I*, 385, pp. 327–356. Available at: https://doi.org/10.1007/82_2014_394.

Smith, K.R., Mehta, S. and Maeusezahl-feuz, M. (2004) 'Indoor Air Pollution from Household Use of Solid Fuels." Comparative Quantification of Health Risks 18:1435–1492'.

Stearns, J.C. *et al.* (2015) 'Culture and molecular-based profiles show shifts in bacterial communities of the upper respiratory tract that occur with age', *The ISME Journal*, 9(5), pp. 1246–1259. Available at: https://doi.org/10.1038/ismej.2014.250.

Steinegger, M. and Salzberg, S.L. (2020) 'Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank',

Genome Biology, 21, p. 115. Available at: https://doi.org/10.1186/s13059-020-02023-1.

Taylor, M., Gonzalez, M. and Porter, R. (2011) 'Pathways to inflammation: acne pathophysiology', *European journal of dermatology: EJD*, 21(3), pp. 323–333. Available at: https://doi.org/10.1684/ejd.2011.1357.

Teo, S.M. *et al.* (2015) 'The Infant Nasopharyngeal Microbiome Impacts Severity of Lower Respiratory Infection and Risk of Asthma Development', *Cell Host & Microbe*, 17(5), pp. 704–715. Available at: https://doi.org/10.1016/j.chom.2015.03.008.

Thurber, R.V. *et al.* (2009) 'Laboratory procedures to generate viral metagenomes', *Nature Protocols*, 4(4), pp. 470–483. Available at: https://doi.org/10.1038/nprot.2009.10.

Torres-Duque, C. *et al.* (2008) 'Biomass Fuels and Respiratory Diseases', *Proceedings of the American Thoracic Society*, 5(5), pp. 577–590. Available at: https://doi.org/10.1513/pats.200707-100RP.

Truong, D.T. *et al.* (2015) 'MetaPhlAn2 for enhanced metagenomic taxonomic profiling', *Nature Methods*, 12(10), pp. 902–903. Available at: https://doi.org/10.1038/nmeth.3589.

Turnbaugh, P.J. *et al.* (2006) 'An obesity-associated gut microbiome with increased capacity for energy harvest', *Nature*, 444(7122), pp. 1027–1031. Available at: https://doi.org/10.1038/nature05414.

Turnbaugh, P.J. *et al.* (2008) 'Marked alterations in the distal gut microbiome linked to diet-induced obesity', *Cell host & microbe*, 3(4), pp. 213–223. Available at: https://doi.org/10.1016/j.chom.2008.02.015.

Turnbaugh, P.J. *et al.* (2009) 'A core gut microbiome in obese and lean twins', *Nature*, 457(7228), pp. 480–484. Available at: https://doi.org/10.1038/nature07540.

Twigg, H.L. *et al.* (2016) 'Effect of Advanced HIV Infection on the Respiratory Microbiome', *American Journal of Respiratory and Critical Care Medicine*, 194(2), pp. 226–235. Available at: https://doi.org/10.1164/rccm.201509-1875OC.

Twigg, H.L., Weinstock, G.M. and Knox, K.S. (2017) 'Lung microbiome in human immunodeficiency virus infection', *Translational Research*, 179, pp. 97–107. Available at: https://doi.org/10.1016/j.trsl.2016.07.008.

Ussar, S. *et al.* (2015) 'Interactions between Gut Microbiota, Host Genetics and Diet Modulate the Predisposition to Obesity and Metabolic Syndrome', *Cell Metabolism*, 22(3), pp. 516–530. Available at: https://doi.org/10.1016/j.cmet.2015.07.007.

Vail, C.D. *et al.* (1990) 'Adjunct treatment of equine respiratory disease complex (ERDC) with the propionibacterium acnes, immunostimulant, EqStim®', *Journal of Equine Veterinary Science*, 10(6), pp. 399–403. Available at: https://doi.org/10.1016/S0737-0806(06)80132-2.

Van der Waaij, D., Berghuis-de Vries, J.M. and Lekkerkerk-van der Wees, J.E.C. (1971) 'Colonization resistance of the digestive tract in conventional and antibiotic-treated mice', *Journal of Hygiene*, 69(03), pp. 405–411. Available at: https://doi.org/10.1017/S0022172400021653.

Vibin, J. *et al.* (2018) 'Metagenomics detection and characterisation of viruses in faecal samples from Australian wild birds', *Scientific Reports*, 8(1), p. 8686. Available at: https://doi.org/10.1038/s41598-018-26851-1.

Vos, T. *et al.* (2020) 'Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019', *The Lancet*, 396(10258), pp. 1204–1222. Available at: https://doi.org/10.1016/S0140-6736(20)30925-9.

Wagner, B.D. *et al.* (2018) 'On the Use of Diversity Measures in Longitudinal Sequencing Studies of Microbial Communities', *Frontiers in Microbiology*, 9. Available at: https://doi.org/10.3389/fmicb.2018.01037.

Wang, B. *et al.* (2017) 'The Human Microbiota in Health and Disease', *Engineering*, 3(1), pp. 71–82. Available at: https://doi.org/10.1016/J.ENG.2017.01.008.

Weersma, R.K., Zhernakova, A. and Fu, J. (2020) 'Interaction between drugs and the gut microbiome', *Gut*, 69(8), pp. 1510–1519. Available at: https://doi.org/10.1136/gutjnl-2019-320204.

Weiss, S. *et al.* (2016) 'Correlation detection strategies in microbial data sets vary widely in sensitivity and precision', *The ISME Journal*, 10(7), pp. 1669–1681. Available at: https://doi.org/10.1038/ismej.2015.235.

Weiss, S. *et al.* (2017) 'Normalization and microbial differential abundance strategies depend upon data characteristics', *Microbiome*, 5. Available at: https://doi.org/10.1186/s40168-017-0237-y.

Werner, J.L. *et al.* (2017) 'Induction of Pulmonary Granuloma Formation by Propionibacterium acnes Is Regulated by MyD88 and Nox2', *American Journal of Respiratory Cell and Molecular Biology*, 56(1), pp. 121–130. Available at: https://doi.org/10.1165/rcmb.2016-0035OC.

WHO: Tuberculosis & HIV (2020). Available at: https://www.who.int/teams/global-hiv-hepatitis-and-stis-programmes/hiv/treatment/tuberculosis-hiv (Accessed: 30 August 2022).

Wood, D.E., Lu, J. and Langmead, B. (2019) 'Improved metagenomic analysis with Kraken 2', *Genome Biology*, 20(1), p. 257. Available at: https://doi.org/10.1186/s13059-019-1891-0.

Wood, D.E. and Salzberg, S.L. (2014) 'Kraken: ultrafast metagenomic sequence classification using exact alignments', *Genome Biology*, 15(3), p. R46. Available at: https://doi.org/10.1186/gb-2014-15-3-r46.

Yan, M. *et al.* (2013) 'Nasal microenvironments and interspecific interactions influence nasal microbiota complexity and S. aureus carriage', *Cell host &*

microbe, 14(6), pp. 631–640. Available at: https://doi.org/10.1016/j.chom.2013.11.005.

Yatsunenko, T. *et al.* (2012) 'Human gut microbiome viewed across age and geography', *Nature*, 486(7402), pp. 222–227. Available at: https://doi.org/10.1038/nature11053.

Ye, S.H. *et al.* (2019) 'Benchmarking Metagenomics Tools for Taxonomic Classification', *Cell*, 178(4), pp. 779–794. Available at: https://doi.org/10.1016/j.cell.2019.07.010.

Zaharia, M. *et al.* (2011) 'Faster and More Accurate Sequence Alignment with SNAP'. arXiv. Available at: https://doi.org/10.48550/arXiv.1111.5572.

Zaheer, R. *et al.* (2018) 'Impact of sequencing depth on the characterization of the microbiome and resistome', *Scientific Reports*, 8(1), p. 5890. Available at: https://doi.org/10.1038/s41598-018-24280-8.

Zaneveld, J.R., McMinds, R. and Vega Thurber, R. (2017) 'Stress and stability: applying the Anna Karenina principle to animal microbiomes', *Nature Microbiology*, 2(9), p. 17121. Available at: https://doi.org/10.1038/nmicrobiol.2017.121.

Zhang, X. *et al.* (2020) 'The Gut Microbiota: Emerging Evidence in Autoimmune Diseases', *Trends in Molecular Medicine*, 26(9), pp. 862–873. Available at: https://doi.org/10.1016/j.molmed.2020.04.001.

Zhou, Y. *et al.* (2019) 'The upper-airway microbiota and loss of asthma control among asthmatic children', *Nature Communications*, 10(1), p. 5714. Available at: https://doi.org/10.1038/s41467-019-13698-x.