IDENTIFCATION OF ROBUST BIOMARKERS USING MICROBIOME DNA SEQUENCING WITH A FOCUS ON THE ORAL MICROBIOME AND CANCER ASSOCIATIONS

By

Jacob T. Nearing

Submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy

at

Dalhousie University

Halifax, Nova Scotia

December 2022

© Copyright by Jacob T. Nearing, 2022

In the face of many choices sometimes knowing how they differ is just as important as knowing which is best.

Table of Contents

List of Tables

List of Figures

Abstract

The human microbiome can be defined as the community of microbes that live within and on the human body. Modern day microbiome research relies on the use of high throughput deoxyribonucleic acid (DNA) sequencing to characterize and identify microbes and community structures associated with human disease. Indeed, the human microbiome has been proposed as a useful source of biomarkers for numerous human health conditions including cancer. Yet often, microbial biomarkers identified through DNA sequencing efforts are not reproducible between studies. Herein, I present my work on examining computation choices during microbiome experiments and applying this knowledge to identify novel associations with the salivary microbiome and breast, prostate, and colon cancer.

Typical DNA sequenced based surveys of microbiomes, requires numerous choices to be made at each step of the experimental protocol. Often theses choices are unclear with no current gold standard within the field. Yet, these methods are used interchangeably within the literature without regard for how that choice might impact the biological conclusions that researchers find. To help address this issue I examined how choices made during two computational steps for processing DNA sequencing data impacted the biological conclusions drawn. In doing so I highlighted critical differences that can be attributed to bioinformatic tool choice and suggest potential solutions for these issues in the future.

Using information from the above chapters, the second half of this thesis represents analysis done on salivary samples from a large population cohort to both characterized salivary microbiome variation and how that variation is associated with breast, prostate, and colon cancer. Within these chapters we show that several daily life factors are significantly associated with salivary microbiome composition, yet they only explain a small amount of total community variance. We also show that the salivary microbiome contains little signal in cases of breast and prostate cancer. Contrastingly our work did show potential salivary microbiome associations within cases of colon cancer. These findings show the potential for future diagnostic research into the relation between the oral microbiome and colon cancer development.

List of Abbreviations Used

16S/18S/5S/5.8S/28S	16/18/5/5.8/28 Svedberg
ACR	Alberta Cancer Register
ALDEx2	Analysis of Variance (ANOVA)-like Differential Expression
ANCOM-II	Analysis of Compositions of Microbiomes Two
ASVs	Amplicon Sequence Variants
ATP	Alberta’s Tomorrow Project
AUC	Area under the curve
AUROC	Area under the receiver operator curve
BH	Benjamini-Hochberg
BMI	Body mass index
bp	Base-pair(s)
CanPath	Canadian Partnership for Tomorrow’s Health
CLR	Center-log-ratio
CoDa	Compositional data analysis
CRC	Colorectal cancer
DA	Differential abundance
DESeq2	Differential expression analysis for sequence count data 2
DNA	Deoxyribonucleic acid
edgeR	Empirical analysis of digital gene expression in R
ENA	European Nucleotide Archive
FDR	False discovery rate
HMP	Human Microbiome Project
IQR	Inter-quartile range
ITS	Internal transcribed spacer
LDA	Linear discriminant analysis
LEfSe	Linear discriminant analysis Effect Size

Limma	Linear models for microarray data
voom	Variance modeling at the observational level
MaAsLin2	Microbiome Multivariable Associations with Linear Models
MED	Minimum entropy decomposition
mRNA	Messenger ribonucleic acid
OSCC	Oral squamous cell carcinoma
OTU	Operational taxonomic unit
PATH	Partnership for Tomorrow’s Health
PCoA	Principal coordinates analysis
PCR	Polymerase chain reaction
PD	Phylogenetic Diversity
PERMANOVA	Permutational multivariate analysis of variance
pH	Potential of hydrogen
PICRUSt2	Phylogenetic Investigation of Communities by Reconstruction of Unobserved States Two
QIMME2	Quantitative Insighted Into Microbial Ecology Two
RDP	Ribosomal Database Project
REB	Regional ethics board
ROC	Receiver operator curve
rRNA	Ribosomal ribonucleic acid
SD	Standard deviation
T2DM	Type 2 Diabetes Mellitus
TMM	Trimmed mean of M-values
TMMwsp	Trimmed mean of M-values with singleton pairing
v	Version
V1/V3/V4/V5/V6/V8	Variable region 1/3/4/5/6/8
WHR	Waist-hip ratio

Acknowledgements

I would first like to thank all my lab members and collaborators that I have worked with throughout my PhD. Much of the work done within this thesis would not have been possible without your critical feedback, support, work, and guidance. I'm grateful for everyone putting up with my late-night messages and constant "spam" in our slack channels.

Thank you to my supervisor, mentor, and friend Dr. Morgan Langille. Your guidance throughout my degree has been invaluable. Thank you for always encouraging me to search out my own ideas and your unwavering support throughout my degree.

I would also like to thank all my committee members for their feedback throughout my PhD: Dr. Zhenyu Cheng, Dr. Nikhil Thomas, Dr. Ketan Kulkarni, and Dr. John Rohde. A special thanks goes to Dr. Nikhil Thomas for your guidance as my co-supervisor. I would also like to thank Dr. Joesph Bielawski for his participation on my comprehensive exam.

Thank you to Dr. Michael Surrette from McMaster University for taking the time to be an external reviewer of this thesis.

I have been fortunate enough to be supported by several different funding agencies through my degree. These include researchNS, Beatrice Hunter Cancer Research Institute, Dalhousie University, and NSERC. Without your finical support none of this work would have been possible.

Thank you to my parents Joe and Lisa Nearing without your support I would have never gotten the opportunity to pursue my passion of science.

To Laura my wife: thank you for being there for me every step of the way. Your ability to put up with my scientific obsession and my excitement over “pretty” graphs continues to fuel my research curiosity.

Chapter 1 – Introduction

Since the advent of the microscope in the 17th century it has become clear that microbes play a critical role in numerous eco-systems including the human body (Dobell, 1932). Indeed, these microbial communities and their functions termed "the human microbiome" are not only important during active infection, but also play critical roles in other aspects of human health. For example, complex carbohydrates are fermented by various bacteria within the large intestine to produce products such as short chain fatty acids, an important immune regulator (Sanna et al., 2019). Due to the vast importance of these communities and their functions in our health, there has been significant interest to identify community characteristics associated with human diseases (Y. Fan & Pedersen, 2021). In fact, the human microbiome has been associated with a vast array of human health conditions ranging from autism (Morton, Donovan, & Taroncher-Oldenburg, 2022; Yap et al., 2021), to inflammatory bowel disease (Lloyd-Price et al., 2019), to multiple sclerosis (Jangi et al., 2016). For example, in the field of oncology both community wide composition measurements and specific microbial abundances have been associated with colon, prostate, breast, and pancreatic cancer (Cullin, Azevedo Antunes, Straussman, Stein-Thoeringer, & Elinav, 2021). Accordingly, it has been proposed that the human microbiome could harbor microbial biomarkers for a variety of different cancers. Highlighting the potential to identifying individuals at risk for disease progression and development through various microbial sequencing platforms (Marcos-Zambrano et al., 2021). However, results from microbiome studies are often not reliable study-study making the detection of robust biomarkers difficult (Schloss, 2018).

This thesis will explore various reasons why microbial biomarkers are not always robust across studies and present potential solutions to help solve this in future investigations. It is a publication-based thesis, which means all the main chapters other than the introduction (Chapter 1) and discussion (Chapter 6) have been published or have been submitted for publication. As a result, this introductory chapter will provide needed context and complementary information to the introductory sections of each results Chapter (Chapters 2-5). Within this chapter I will first give a brief overview of the general community make up and functional potential of the human microbiome with an emphasis on the oral cavity. I will also discuss how the oral microbiome has been associated with oral health conditions such as dental caries, and more distal diseases such as cancer. Following this I will introduce some of the most common ways to survey the human microbiome using DNA sequencing and the challenges and limitations that accompany these types of surveys. This includes a variety of biases associated with DNA sequencing and the various choices researchers must make throughout microbiome experiments. After which an in-depth discussion on dealing with sequencing error and the many choices that face researchers will be highlighted. Finally, we will discuss how microbial communities are examined using sequencing data and how biomarkers are generated from measures of diversity or specific taxonomic associations.

Following this introductory chapter, I will present four chapters that represent published or submitted (Chapter 5) manuscripts discussing the above topics. Within Chapters 2 and 3 we will explore how bioinformatic choices can impact biological results in microbiome studies. Chapter 2 will highlight the impact of methodology choices when dealing with sequencing errors and creating analytic units of analysis in marker gene sequencing. A step that is necessary during data processing but presents multiple choices with no clear best option. In Chapter 3 I will discuss current issues in microbiome differential abundance analysis, the goal of which is to find associations between specific microbes and their environment. This type of statistical analysis is found commonly within microbiome literature and is often used to indicate potential biomarkers.

Following these two Chapters we will slightly change focus to take a dive into the exploration of the oral microbiome and the various anthropometric, demographic, and dietary metrics that can impact its composition within healthy adults. Results from this chapter present key information required to understand the variation of the salivary microbiome in the absence of disease. In Chapter 5, we will use all this information to attempt to identify microbial biomarkers of disease within the oral cavity of retrospective and prospective cases of breast, prostate, and colon cancer. The goal of which is to assess the oral microbiome's applicability in risk detection of these diseases in a population setting.

1.1 - Human Microbiome Structure and Community Members

Despite what can be seen by the naked eye, there are trillions of microbes that live on and within the human body (Sender, Fuchs, & Milo, 2016). Indeed, studies dating back to the mid-1600s by Antonie van Leeuwenhoek, described a wide variety of different shapes and sizes of microbes living on our skin, in our stool, and within our saliva (Dobell, 1932). Today we now know these microbes as bacteria, archaea, fungi, viruses, and microeukaryotes. Following their discovery microbes were traditionally studied using various culturing techniques. These techniques uncovered their biochemical properties and the numerous contributions they make to various diverse ecosystems. For example, Alfred Nissle was the first to culture the probiotic Nissle 1917, a strain of *Escherichia coli*, after hypothesizing that microorganisms within the gut of a wounded soldier provide protective effects against infection (Sonnenborn, 2016). This isolate is still used today as a probiotic and led to the concept of colonization resistance, a key theory within modern day microbiota research (Buffie & Pamer, 2013).

While monumental discoveries on both microbes and microbial communities were made through culturing, it became apparent that many microbes were either difficult or not possible to cultivate within the laboratory (Staley & Konopka, 1985). This problem known as “The Great Plate Count Anomaly” would be addressed by Norman Pace in 1984 by sequencing environmental ribosomal ribonucleic acid (rRNA) genes to identify microbes found within hydrothermal vents (Stahl, Lane, Olsen, & Pace, 1984). Adapting from this work in 1996 Wilson and Blitchington compared cultured bacteria to those found by sequencing 16 Svedberg (16S) rRNA genes found within human stool (Wilson & Blitchington, 1996). This paper along with work by Kroes et al., Suau et al., and

Eckburg et al., showed that sequencing the 16S rRNA gene from human associated samples could be used to identify both culturable and nonculturable bacteria; vastly increasing the known diversity of the human microbiome (Eckburg et al., 2005; Kroes, Lepp, & Relman, 1999; Suau et al., 1999).

Following the above early work came the revolution of high throughput sequencing machines that led the way to modern day microbiota research. Indeed, in 2007 the National Institutes of Health launched the Human Microbiome Project (Turnbaugh et al., 2007), which was the first to publish reference data on the microbiomes found in 256 healthy United States volunteers (Huttenhower et al., 2012). The Human Microbiome Project explored several human sites including the gut, mouth, nose, skin, and vagina, showing the unique characteristics of each community. These early studies that cataloged microbial composition represented an important stride in understanding the ecological importance of microbial communities within the human body (Gilbert et al., 2018).

From these works it was clear that among all sites within the human body the gut, specifically the lower intestine, harbors the richest microbial community (Huttenhower et al., 2012). In fact over 1000 different bacterial species have been cultured from the gut microbiome of humans (Rajilić-Stojanović & de Vos, 2014). Typically, this community is dominated by bacteria, bacteriophages, and lower numbers of archaea, fungi, and single cell eukaryotes (Laforest-Lapointe & Arrieta, 2018; Yatsunenko et al., 2012). Indeed, the number of bacterial cells within our lower intestine is thought to be similar to the total number of cells that make up the human body (Sender et al., 2016). Of these cells over half of them belong to two phyla, Bacteroidetes and Firmicutes. Bacteroidetes are made up of a diverse group of Gram-negative anaerobes while Firmicutes are typically Gram-positive and can be both anaerobes or aerobes depending on species (Paster, Dewhirst, Olsen, & Fraser, 1994; Vos et al., 2011). Interestingly, the ratio between these two phyla within the gut has been associated with several aspects of human health such as obesity and aging. However, this result has not been generalizable across all cohort studies (Ley, Turnbaugh, Klein, & Gordon, 2006; Magne et al., 2020). These differing results are thought to be due to a number of issues, including differing experimental protocols and methods of analysis (Bahl, Bergström, & Licht, 2012; Vebø, Karlsson, Avershina, Finnby, & Rudi, 2016).

Along with Firmicutes and Bacteroidetes, Actinobacteria make up a large proportion of the bacterial members within the gut microbiota. Within the gut this phylum is mainly represented by *Bifidobacterium,* which have been linked to several beneficial traits and is among the earliest colonizers of the human gastrointestinal tract (Martino et al., 2022). Indeed, *Bifidobacterium* species such as *Bifidobacterium longum*, *Bifidobacterium breve*, and *Bifidobacterium infantis* have been suggested as probiotics due to their beneficial health effects. For example, higher abundances of *B. longum* has been associated with improved remission rates in ulcerative colitis (Furrie et al., 2005; Tamaki et al., 2016). Furthermore, a mixture of *Bifidobacterium* has also been shown to lead to improved responses to immune checkpoint inhibitors; a recent approached used in treating a variety of cancers (S. Sun et al., 2020).

Another prominent group within the gut microbiome is the phylum Proteobacteria (Huttenhower et al., 2012). Proteobacteria first described as “purple bacteria and their relatives” by Woese in 1987 are a group of Gram-negative bacteria that can be found within the skin, oral cavity, vagina, and gut of humans (Carl Richard Woese, 1987). A number of members within this phylum are well known human pathogens including *Escherichia coli*, *Salmonella bongori*, *Salmonella enterica*, *Helicobacter pylori*, *Yersinia pestis*, and various *Shigella* species (Rizzatti, Lopetuso, Gibiino, Binda, & Gasbarrini, 2017). Interestingly, the increased relative abundance of this phyla has been associated with a number of conditions including infectious complications during chemotherapy (Nearing et al., 2019), or individuals with inflammatory bowel disease (Lloyd-Price et al., 2019; Shin, Whon, & Bae, 2015). While work using mouse models of inflammatory bowel disease have reflected these results, the role of this phyla during inflammation remains unclear. Nevertheless, it is speculated that the expansion of this phyla may be due to differing oxygen levels within the gut during inflammation (Rigottier-Gois, 2013).

Several other phyla including Fusobacteria, Verrucomicrobia, Tenericutes, Cyanobacteria, and TM7 are also found within the gut and have been associated with various health conditions (Huttenhower et al., 2012). Overall, bacteria within the gut play diverse roles within the human body including the production of various metabolites, providing colonization resistance, regulating immune function, and aiding gut physiology (Ducarmon et al., 2019; Lloyd-Price, Abu-Ali, & Huttenhower, 2016), clearly highlighting their importance to human health. In fact some suggest that human physiology evolved closely with these microbial inhabitants leading to evolutionary models of human holobionts (Madhusoodanan, 2019). Despite this research, the origin of many of these gut microbes remains elusive, although it has been shown that at least a third are transferred from the oral cavity (Schmidt et al., 2019).

The oral cavity is the second most diverse human microbiome after the colon (Huttenhower et al., 2012). It houses unique communities throughout, with specific microbial community compositions found on the tongue, gingival sulcus, cheek, teeth, soft palate, hard palate, and saliva (Hall et al., 2017; J. He, Li, Cao, Xue, & Zhou, 2015). In fact over 1,000 different species have been identified within the oral cavity yet each individual surface is only typically colonized by ~50 species (Dewhirst et al., 2010).

Hard surfaces such as the supragingival dental plaque are characterized by large proportions of Proteobacteria, Firmicutes, Fusobacteria, and Bacteroidetes, although other phyla such as Actinobacteria, and TM7 are also present (Xu et al., 2015). In contrast within the oral mucosa, microbial communities are mainly dominated by only two phyla, Firmicutes and Proteobacteria, which can make up to 90% of the total community (Xu et al., 2015).

This has led to the idea of a "site-specialist hypothesis" indicating that despite these communities being in close proximity and often in contact with the same fluids they maintain distinct populations (Mark Welch, Dewhirst, & Borisy, 2019). This is because the competitiveness of bacteria within different regions of the oral cavity is dependent on their ability to adhere to the particular surface of interest, compete for nutrients within the region, and the availability of oxygen (Lamont, Koo, & Hajishengallis, 2018). Indeed, while teeth are smooth hard surfaces coated in enamel, the tongue check and other soft surfaces are covered by mucosal epithelium leading to distinctive landscapes within these sites. For example, *Corynebacterium*, a predominantly aerobic genera, is consistently found within plaque communities and gingival crevices of individuals. However, this genus is only found in trace abundances in other areas of the oral cavity (Mark Welch, Rossetti, Rieken, Dewhirst, & Borisy, 2016). Contrastingly, areas of reduced oxygen availability such as the subgingival cavity or near the tooth base, are predominantly anaerobic taxa such as *Fusobacterium*, *Porphyromonas*, and *Capnocyotphaga* a group of bacteria dependent on high levels of CO_2 (Lamont et al., 2018).

While broad taxonomic differences are apparent between oral sites, even species within the same genera can show significant differences in relative abundance between them. For instance, while *Actinomyces graevenitzii* is found in significantly higher proportions on the tongue than teeth, *Actinomyces massiliensis* a close relative shows the opposite trend (Mark Welch et al., 2019). Even different subtypes within several species such as *Haemophilus parainfluenza*, and *Granulicatella adiacens* have been suggested to differ in abundance between oral sites (Costello et al., 2009).

Yet, despite these distinct environments some microbes can still be found across multiple sites. For instance, work by Hall et al., showed that a number of bacterial species across five different phyla could be found within the same individuals tongue plaque, supragingival plaque, and saliva (Hall et al., 2017). This is because microbial communities within the oral cavity are constantly being washed and exposed to saliva, a free flowing fluid that not only provides aid during digestion and maintaining pH (Mandel, 1987), but also acts to mix and move microbes across different sites (Lamont et al., 2018). Thus, saliva samples from humans are typically made up of a wide variety of microbes both anaerobic and aerobic that originated from areas such as the teeth, throat, tongue, and tonsils. Although, communities inferred from saliva samples often correspond better with communities found on the tongue than within dental plaques

(Segata et al., 2012). Due to saliva being easy to collect and its ability to contact several oral sites, it has become a popular sample choice for the analysis of the oral microbiome.

Human microbiome saliva samples are typically dominated by four to five phyla: Firmicutes, Bacteroidetes, Proteobacteria, Fusobacteria and Actinobacteria. Other prominent members include TM7, Spirochaetes and Synergistes (Bik et al., 2010; Segata et al., 2012). In general the abundance of various taxa within the salivary microbiome is much more evenly distributed than other oral communities or even the gut microbiome (Segata et al., 2012). Although it should be noted that some genera are still represented in higher abundances than others including *Steptococcus*, *Veillonella*, *Prevotella*, *Neisseria*, *Fusobacterium*, *Rothia* and *Haemophlius* (Bik et al., 2010; Segata et al., 2012).

Saliva itself is sterile when it enters the oral cavity and as previously mentioned only develops a microbiome signature through interaction with various oral surfaces (Schrøder, Bardow, Eickhardt-Dalbøge, Johansen, & Homøe, 2017). Interestingly, despite significant external exposure through eating and activities such as tooth brushing and kissing, the overall composition of the salivary microbiome has been shown to be temporarily stable within individuals (Caporaso et al., 2011; Leake, Pagni, Falquet, Taroni, & Greub, 2016). Although recent work has shown that over long periods of time significant changes to environmental factors, such as an Antarctic expedition, can led to compositional changes (Bhushan, Yadav, Singh, & Ganju, 2019). However, despite this stability there remains a significant amount of variation between individuals.

Indeed, despite broad patterns of community composition at the phylum there remains significant deviation between humans at lower taxonomic ranks such as species and strain (Caporaso et al., 2010; Huttenhower et al., 2012). Indeed, work on the oral microbiome has shown that specific taxon can be associated with both host and environmental factors such as demographics, physical features, and lifestyle choices (X. Fan, Peters, et al., 2018; Mason, Nagaraja, Camerlengo, Joshi, & Kumar, 2013; J. Wu et al., 2016; Y. Yang, Cai, Zheng, et al., 2019). However, only a small number of studies have examined how each of these factors contribute to overall variance within the oral microbiome.

With the goal of identifying microbial biomarkers of disease in mind, it is important to understand what contributes to this variation in healthy individuals first. Answering this key question is the focus of Chapter 4 of this thesis. I, along with colleagues attempt to address this question using a large population cohort of healthy Atlantic Canadians. Within Chapter 4 you will also find a brief introduction on previous works within this area along with a discussion which puts our results into context with previous literature. Results from this Chapter will be critical in the assessment of salivary biomarkers identified in the following results Chapter.

Unsurprisingly, variation within the salivary microbiome has also been associated with numerous different oral health conditions such as dental caries (Baker et al., 2021), periodontitis (Lundmark et al., 2019; Yong Zhang et al., 2021), and oral cancesrs (Hsiao et al., 2018). For example, in the case of dental caries, microbes play a critical role through the formation of biofilms. Biofilms on teeth if left unmanaged can propagate, especially during consumption of simple carbohydrates such as those found in common soda pops (Lamont et al., 2018). In fact, it is thought that the consumption of simple carbohydrates such as those found in sugary drinks is one of the leading causes of dental caries worldwide (Peres et al., 2019). This is because biofilm formation and expansion can led to increased levels of acidogenic and aciduric bacteria (Lamont et al., 2018). Common microbial culprits include *Streptococcus mutans*, *Streptococcus sobrinus*, and various lactobacilli and bifidobacteria (Simón-Soro & Mira, 2015). These acidogenic communities then begin generating various acids which can demineralize tooth enamel, eventually leading to the formation of caries (Simón-Soro & Mira, 2015).

Changes in the composition of the salivary microbiome have also been associated with periodontitis a condition highlighted by inflammation within the gingiva. Studies on saliva comparing healthy controls to those with periodontitis have uncovered several potential microbial biomarkers with increases in the abundance of *Porphyromonas gingivalis* being one of the most common features (Belstrøm, 2020). Moreover, a recent study using 16S rRNA gene sequencing showed that the relative abundance of *P. gingivalis* could be used to discriminate healthy individuals from those with periodontitis, demonstrating an area under the receiver operator curve (AUROC) of 0.80 (Damgaard et al., 2019). These findings highlight the need to explore *P. gingivalis* in larger population-based settings.

A number of cancers of the mouth, throat and neck have been associated with changes in salivary microbiome composition (Irfan, Delgado, & Frias-Lopez, 2020). The most well studied of which has been oral squamous cell carcinoma (OSCC). Various works have identified over twenty-two different taxa associated with OSCC diagnosis within the oral microbiome (Irfan et al., 2020). While no clinically useful biomarkers have yet to be validated due to high variability between study results, exploration into the oral microbiome, biomarker discovery, and OSCC remains a fruitful endeavor as we learn more about reducing study to study bias.

The composition of the oral microbiome has also been associated with several health conditions distal to the oral cavity. These include type two diabetes mellitus (T2DM), obesity, and various cancers (Belstrøm, 2020). In the case of T2DM various reports have identified specific taxon associations, although agreement between studies is sparse. For example, work by Sabharwal et al., has reported decreased oral microbial diversity, where as Chen et al., reported the opposite (B. Chen et al., 2020; Sabharwal et al., 2019). Furthermore, work by Almeida-Santos et al., reported no differences in diversity between non-diabetics and those with T2DM (Almeida-Santos, Martins-Mendes, Gayà-Vidal, Pérez-Pardal, & Beja-Pereira, 2021). Yet despite these differences high disease classification rates from oral microbiome data have been achieved in study specific cohorts (Omori et al., 2022). Again, these results highlight the need to understand how variation within the oral microbiome can impact biomarker detection and to what degree biological and technical variation explain these differences.

In the case of non-oral cancers several types have been associated with differences in oral microbiome composition with some of the most well studied being pancreatic, and colon cancer (Belstrøm, 2020; X. Fan, Alekseyenko, et al., 2018; Vogtmann et al., 2022a). In the case of pancreatic cancer several studies have identified shifts in oral microbiome composition in affected patients. Studies on saliva have shown variable results with the most commonly associated taxa being *Haemophilus* and *Porphyromonas* (Irfan et al., 2020). Interestingly, signal from these taxa have not only been associated in saliva (X. Fan, Alekseyenko, et al., 2018; Torres et al., 2015; Vogtmann et al., 2020) but also tongue plaque (H. Lu et al., 2019). Furthermore, work within mice on pancreatic cancer has further shown that lipopolysaccharide a component of the outer membrane of Gram-negative bacteria, like *Porphyromonas* and *Haemophilus* is sufficient to promote carcinogenesis within the pancreas (Ochi et al., 2012). These studies highlight a potential mechanism of association between these bacterial taxa and pancreatic cancer develop. However, whether these mechanisms are at play or have biological significance remains a question within the field. Highlighting the need for further mechanistic work in animal models.

Colon cancer is the most well studied non-oral cancer in relation to the oral microbiome (Belstrøm, 2020; Irfan et al., 2020). Over the past decade several different oral taxa have been identified as potential biomarkers of disease and several studies have shown high classification accuracy from microbial oral samplings (Flemer et al., 2017; S. Zhang et al., 2020). However, the effectiveness of the oral microbiome as a biomarker for colon cancer development still lacks comprehensive evaluations within population-based settings. Furthermore, the association between disease progression and shifts within the oral microbiome is still not well understood. To help address these questions Chapter 5 of this thesis will focus on the potential relationship between the oral microbiome and both retrospective and prospective cases of colon cancer within a large population cohort. Further information on the current literature of colon cancer and the oral microbiome can be found within this Chapter's introduction and discussion.

Other non-oral cancers have also been associated with the oral microbiome, although research within this field is still emerging. For example, a recent study on lung cancer and the oral microbiome highlighted several taxa associated with disease development (Vogtmann et al., 2022b). Moreover, exploration into the oral microbiome and cancers previously associated with oral health or changes in the gut microbiota are still lacking. Two cancers that fall into this category are prostate cancer and breast cancer. Prostate cancer has been not only associated with periodontal disease (Lee, Kweon, Choi, Kim, & Choi, 2017), but a handful of small pilot studies have identified various shifts within the gut microbiota of affected patients (Golombos et al., 2018; Liss et al., 2018). Similarity studies on breast cancer have identified shifts in the gut microbiome of patients (Flores et al., 2012; Goedert et al., 2018, 2015) as well as shifts in the microbial communities present in urine and breast tissue (Irfan et al., 2020). Due to these recent findings highlighted above, Chapter 5 of this thesis will examine the salivary microbiome of both prospective and retrospective cases of these diseases in large population cohorts.

1.2 - Microbiome Amplicon Sequencing and Experimental Methodology

Modern-day large-scale surveys of the human microbiota are accomplished using some variation of DNA sequencing technology. These technologies include both long and short-read sequencing platforms that have revolutionized the way we study microbial communities (Gehrig et al., 2022). Within these categories there are two general approaches to conducting a DNA based microbiome survey (Nearing, Comeau, & Langille, 2021). These methods differ in both the resources required to perform them but also in the types of information that can be gained from them.

Metagenomic shotgun sequencing is a resource intensive process where hundreds of millions of reads are generated from semi-randomly sheared DNA within a sample. This type of sequencing not only generates taxonomic information but also functional knowledge at the expense of high resource requirements (Douglas & Langille, 2021). On the other hand, marker gene sequencing, also known as amplicon sequencing, is a less resource intensive process that generates taxonomic information and can also be used to infer high level functional information (Douglas & Langille, 2021). Due to the lower resource requirements of marker gene sequencing, it is the primary choice in microbiome biomarker detection surveys. The prevalence of marker-gene sequencing is reflected in the literature with most large-scale surveys focusing on taxonomic biomarkers rather than microbial function. Although there is argument within the field that metagenomic shotgun sequencing may be a better way to characterize microbial communities because it provides insight into microbial function. Due to the reasons above along with the broad application and lower resource requirements, this thesis and all subsequent chapters within it will be focused on the discussion and use of marker gene sequencing. In particular I will highlight bioinformatic processing and statistical examination of marker gene sequencing data in Chapters 2 and 3.

Marker gene sequencing uses the amplification and sequencing of specific genes contained within the taxa of interest to infer community composition. Genes that are selected for this process are known typically as marker genes and have two general characteristics. First, genes that are selected should be present within all taxa of the community of interest. Second the amplified and sequenced gene should have enough sequence variation between different lineages that reasonable estimates of taxonomic identity can be determined. The most commonly sequenced marker gene in microbiome studies is the 16S rRNA gene (Goodrich, Di Rienzi, et al., 2014).

The use of the 16S rRNA gene (18S rRNA gene in eukaryotes) originates from pioneering work in bacterial and archaeal phylogenetics by Carl Woese and George E. Fox during the late 1970s (Carl R Woese & Fox, 1977). Woese and Fox identified the 16S rRNA gene as useful in the development of prokaryotic phylogenies due to its favorable phylogenetic characteristics. The 16S rRNA gene is ~1550 base pairs long and plays an essential role in the 30S ribosomal subunit; allowing it to bind to Shine-Dalgarno sequences on messenger RNA (mRNA). Binding of mRNA on various areas within the 16S rRNA gene has led to strong selection pressure (Carl Richard Woese et al., 1975). Areas under selection pressure from interaction requirements are known as conversed regions within the gene and are useful for the construction of primers to amplify different gene regions. Conserved regions within the 16S rRNA gene are flanked by nine regions of hypervariability. Hypervariable regions range from 30-100 base pairs in length. Due to their slow rate of base pair substitutions they not only can be used to detected the relatedness between species but also their taxonomic identity (Carl R Woese & Fox, 1977).

Due to the limited read length capacity of high throughput short read sequencing, several universal primers targeting different conserved regions of the 16S rRNA gene have been developed (Comeau, Douglas, & Langille, 2017). Despite substitutions within hypervariable regions no one region is sufficient to identify all bacteria at the species level (Chakravorty, Helb, Burday, Connell, & Alland, 2007). Furthermore, despite the high amount of conservation within conserved regions, not all 16S rRNA genes are bound at the same efficiency by universal primers currently in use (Comeau et al., 2017). Commonly used primers include those that target the first and third (V1-V3), fourth and fifth (V4-V5), or sixth and eighth hypervariable regions (V6-V8). Although variations from these primer sets do exist. Biases associated with these selections have previously been described by myself and lab colleagues (Nearing et al., 2021). It is now clear that choice of universal primer set can have a large impact on not only the taxa identified but also the biological conclusion drawn from a study. For example, recent work by Willis et al., found that the V6-V8 region was optimal for the identification of archaea in the North Atlantic Ocean when compared to the V4-V5 region (C. Willis, Desai, & LaRoche, 2019). Numerous taxonomic discrepancies between these regions have been identified. However, each of these primer sets do target a diverse group of taxa across the prokaryotic tree of life. As such the sequences generated from this approach can give a strong understanding of the underlying prokaryotic composition within a sample.

Several other marker genes have also been proposed and have been used in the study of microbial communities. For example, pioneering work in 1984 by Stahl et al., amplified the 120-base pair 5S rRNA gene which is present across all domains of life apart from some fungi, higher animals, and protists (Szymanski, Barciszewska, Erdmann, & Barciszewski, 2002). This allowed the identification of several different microbial lineages living within hydrothermal vents (Stahl et al., 1984). However due to its limited size, and thus limited phylogenetic information, this gene is no longer commonly used in sequence-based microbiome studies. Contrastingly, fungi are commonly identified by sequencing internal transcribed spacer (ITS) regions located between the small and large subunit rRNAs (Kõljalg et al., 2005). In bacteria there is a single ITS region, (ITS1) between the 16S and 23S rRNA genes, however in eukaryotes a second region (ITS2) is located between the 5.8S and 28S rRNA genes. Unlike the 16S rRNA gene ITS regions undergo higher rates of insertions and deletions which often allow for better taxonomic resolution within fungi than the 18S rRNA gene (Schoch et al., 2012). Typically, only one ITS region is amplified during marker gene studies each showing bias in its ability to classify specific taxonomic groupings (Blaalid et al., 2013).

Other genes that have been proposed to be useful targets for marker gene sequencing studies include Cpn60, gyrB, and rpoB each with their own benefits (Links, Dumonceaux, Hemmingsen, & Hill, 2012; Ogier, Pagès, Galan, Barret, & Gaudriault, 2019; Peeters & Willems, 2011). For example, Cpn60, a universal chaperonin protein found in bacteria, has recently been shown to be a useful marker gene because it has larger pairwise inter- and intra-species distances than the 16S rRNA gene. Highlighting that Cpn60 may be more useful to identify taxa that are closely related than 16S rRNA gene sequencing (Links et al., 2012). Other examples include *gyrB*, and *rpoB* which have been shown to have stronger taxonomic resolution for specific lineages (Ogier et al., 2019; Peeters & Willems, 2011). However, due to the mass adoption of the 16S rRNA gene as the primary target for bacterial marker gene studies, a larger number of reference sequences are available. Indeed, the SILVA rRNA database (Version, 138.1) has over 9.4 million rRNA gene sequences representing all three domains of life and 89 different bacterial phyla (Pruesse et al., 2007). The SILVA database along with others such as Greengenes is critical for the development and training of sequence based taxonomic classification tools (DeSantis et al., 2006). Furthermore, 16S rRNA gene databases, such as SILVA and Greengenes, also contain reference trees useful for the creation of phylogenetic measurements avoiding the need to create lower accuracy study specific *de novo* trees (Janssen et al., 2018).

Clearly the choice of marker gene used during microbiome studies plays a large role in the biological results obtained at the end of a study. For example, those focused on fungi will obtain different observed communities and possibly different biological conclusions when using primers targeting the ITS1 or ITS2 regions (Blaalid et al., 2013). However, there are numerous other choices that must be made during a typical marker gene sequencing experiment that can impact the biological conclusions that are obtained (Goodrich, Di Rienzi, et al., 2014; Nearing et al., 2021). This includes steps that occur both before and after DNA sequencing. For example, before sequencing researchers must determine the sample of interest, sampling device, DNA extraction method, library preparation method, and sequencing platform. After sequencing multiple computational choices must also be made such as how to deal with low quality sequences, choosing analytic units, taxonomic classification tools, and statistical analysis (Figure 1.1). Each one of these steps can alter the observed communities within their samples and thus impact their conclusions and the resulting biomarkers identified within a study. A previous review that I published highlights many of these choices and potential biases introduced into sequence based microbiome studies (Nearing et al., 2021). However, this thesis will be focused on steps that occur after sequencing is complete. Specifically, I will examine the steps involved in processing data into measurements of the observed microbial community.

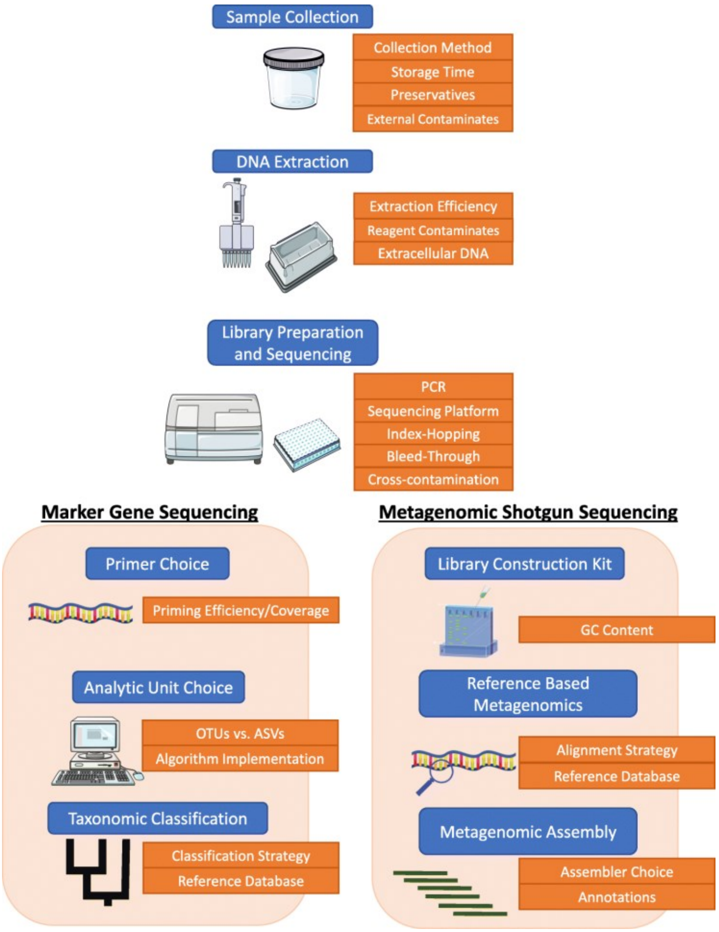


Figure 1.1: ***The various stages that can introduce bias in sequenced-based human microbiome studies****. Orange boxes represent the various areas within a stage that can result in the introduction of systemic bias. This figure was adapted from a figure within a review that I published* (Nearing et al., 2021). *"Bleed-Through" is a bias described on*

Illumina Miseq sequencers that can appear during sequencing when previous DNA is not fully washed out and is sequenced on future runs.

Sequencing data is processed in numerous steps before microbial community measurements are generated. While a general workflow for marker gene data processing is typically followed, large variations from protocol to protocol can exist (Knight et al., 2018). For example, sequences are often quality filter to remove low quality base scores that are susceptible to error. However, the basis on which a quality score threshold is chosen is often unclear or not well justified within studies. Indeed, a gold standard for an acceptable quality score does not exist in current microbiome literature and thus varies from study to study. This lack of 'gold standard' does not just stop at quality score thresholding. In fact, for each step within a typical marker gene sequencing pipeline numerous tools have been developed to address similar problems (Figure 1.2). This often presents to researchers as difficult choices they must make during data processing. Unfortunately, these choices are often unclear or given little or no rationale. This is a problem because analysis of the same sequencing data using different tools may result in remarkably different biological conclusions.

Undoubtedly, these issues spill over into microbial biomarker detection using sequenced based microbiome measurements. While there are several steps within marker gene data processing highlighted within Figure 1.2. Two areas we hypothesized as creating some of the largest study-to-study biases include dealing with sequencing and PCR errors after quality score thresholding and the use of different differential abundance (DA) tools to identify microbial biomarkers. We hypothesis that these steps led to some of the largest bias within microbiome studies because within each step, highly different tools with differing algorithms are used to accomplish similar goals. Yet within literature they are used interchangeably with one another. These two areas have been comprehensively examined in Chapters 2 and 3 of this thesis and present key information needed to compare previously published literature and produce robust biomarkers in future studies.

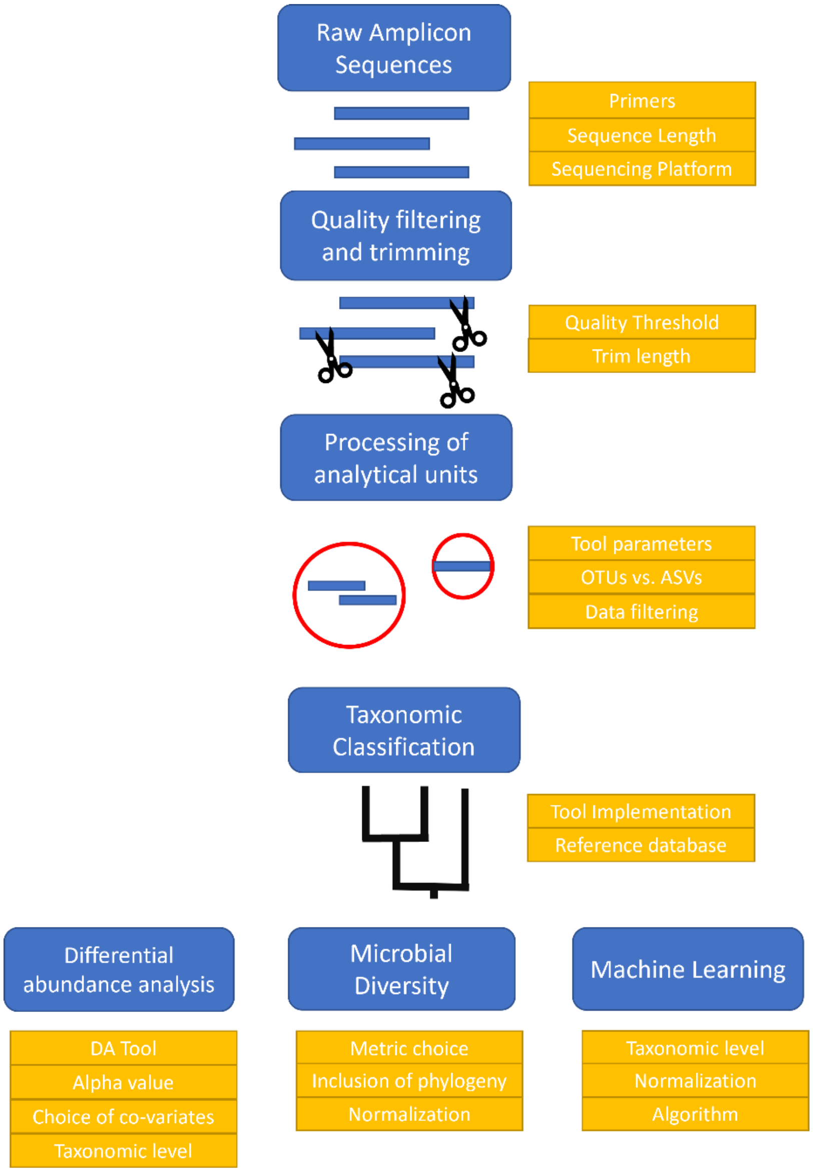


Figure 1.2: ***Processing of marker gene sequencing data and associated processing choices.*** *Blue boxes represent the major steps of a typical data processing pipeline for marker gene sequencing data. Yellow boxes represent some of the various choices researchers must make during data processing.*

1.3 - Dealing With Sequencing Error and the Creation of Analytic Units in Marker Gene Sequencing Experiments

During a sequencing run, base calls are assigned quality scores to represent the likelihood that the sequencer made an error. For sequencing done on Illumina based platforms, the most common short read sequencer, these quality scores are determined through several proprietary parameters. These parameters include the intensity of photons captured by the sensor, and the signal-to-noise ratio of the captured image. Estimates of quality represent the probability that an error was made and can range from 1 in 10 all the way to 1 in 10,000. While it is standard to remove low quality score reads no gold standard exists. Within literature quality score cut-offs can range from estimated error rates of 1 in 10,000 to 1 in 1,000 with little to no clear best option. While quality score thresholding removes problematic reads; base call errors will still be present within filtered data. This is because typical marker gene experiments can generate anywhere from 10,000 to 100,000 reads per sample.

Moreover, sequencing is not the only area where base pair errors can occur. The amplification of the marker gene of interest using polymerase chain reactions (PCR) can also lead to base pair substitutions. For example, Thermus aquaticus DNA polymerases have an estimated error rate of 2.1 x 10^{-4} during typical PCR conditions (Keohavong & Thilly, 1989). While modern day high-fidelity polymerases have improved error rates by an estimated 100 fold, errors can still exist in amplified DNA products (McInerney, Adams, & Hadi, 2014).

Fortunately, a number of different solutions were developed to help address this problem with operational taxonomic unit (OTU) clustering and sequencing denoising taking the forefront (Knight et al., 2018; Pollock, Glendinning, Wisedchanwet, & Watson, 2018). Both of which differing in their general approach to dealing with errors generated during marker gene sequencing protocols. At the time of starting my PhD no gold standard for these methods existed and little to no information existed on how their differences could impact biological conclusions. Furthermore, even within these two solutions various implementations existed using different heuristics, algorithms, or reference databases (Amir et al., 2017; Callahan et al., 2016; R. C. Edgar, 2016, 2017). These were the main motivations behind the analysis within Chapter 2 of this thesis.

Operational taxonomic units (OTUs) were first proposed by Peter Sneath and Robert Sokal in the 1960s to create a quantitative measurement for taxonomic classification. The method they proposed used trait tables with a numeric scoring system. Trait table scores could then be used to quantify the amount of observed difference between organisms and thus create taxonomic groupings (Sneath & Sokal, 1973). This idea has since been co-opted in marker gene sequencing experiments such as 16S rRNA gene sequencing. In this case DNA sequence identity represents the scoring between different taxa and various thresholds are used to identify new taxonomic units. Various nucleotide identity thresholds have been used to generate OTUs, however, the most common range between 97% and 99%. Indeed, early work within the field by Stackebrandt and Goebel coined the use of a 97% identity threshold in 16S rRNA gene sequencing as it was found to match with a 70% DNA reassociation value, which at the time was used to define bacterial species (Moore et al., 1987; Stackebrandt & Goebel, 1994). It should be noted that since this initial work and the revolution that next generation sequencing brought, the definition of bacterial species has constantly been evolving (Doolittle & Papke, 2006; Gevers et al., 2005). Correspondingly, various other OTU clustering thresholds have been proposed to better represent the underlying microbial communities within a sample. For example work by Robert Edgar in 2018 found that a threshold of 99% identity for full length 16S rRNA genes to be more in line with the current definitions of species found within GenBank (R. C. Edgar, 2018). Despite this OTU identity thresholds found within the literature continue to vary due to numerous reasons including differing benchmarks between variable regions (J. S. Johnson et al., 2019).

Identity thresholding, however, is not the only difference found between studies that generate analytic units using OTU clustering. In general, this procedure can be done in three different ways; *de novo* clustering, referenced based clustering, and open-reference clustering (Nearing et al., 2021). Each of these approaches differ in the way sequences are clustered with one another. In the case of *de novo* clustering sequences are grouped purely on sequencing identity using various greedy algorithms implemented in tools such as USEARCH and VSEARCH (R. C. Edgar, 2010; Rognes, Flouri, Nichols, Quince, & Mahé, 2016). Unlike other clustering options *de novo* clustering suffers from two key issues. The first being that its implementation is quadratic in run time, leading to highly intense computational requirements as sequence count expands (Westcott & Schloss, 2015). Secondarily, numerous studies have identified that not only are *de novo* OTUs not directly comparable across studies (Callahan, McMurdie, & Holmes, 2017) but can differ within the same study when sequence order changes (Y. He et al., 2015). On the other hand, *de novo* OTUs have been shown to generate higher quality consensus sequences that represent the underlying biological community in higher detail than other OTU clustering approaches (Westcott & Schloss, 2015).

Reference-based clustering groups sequences based on their similarity to those found within a database. Common databases used for 16S rRNA gene analysis include SILVA, Greengenes, and RDP (Cole et al., 2014; DeSantis et al., 2006; Quast et al., 2013). While computational requirements to generate reference-based OTUs are lower and they are more easily comparable between studies, they often lack the ability to accurately depict the biodiversity found within a community (Westcott & Schloss, 2015). This is due to several reasons, including reads from potentially novel taxa being disregarded due to them not being in reference databases. To address this a third approach was developed which is a hybrid between *de novo* and reference-based clustering.

Open reference clustering is a hybrid approach that first generates clusters in a reference-based manner and then uses remaining sequences to generate *de novo* OTUs (Navas-Molina et al., 2013). This approach has been commonly employed across marker gene sequencing studies as it highlights the strengths of both *de novo* and reference-based clustering. Although having two differing definitions of OTUs within the same analysis has been critiqued (Westcott & Schloss, 2015).

Regardless of various works being done to compare these methods all three are commonly used across microbiome studies with various identity thresholds. Indeed, work has shown that using different clustering strategies or precent identity thresholds within the same datasets can significantly alter diversity measures while having only subtle effects on overall composition (Sul et al., 2011; Westcott & Schloss, 2015). As such a clear choice in what identity threshold or clustering approach used often comes down to the biological question at hand and the need to compare results between studies.

While OTUs have served the microbiome community well, several deficiencies in their use have been noted including: their inability to identify microbes at high taxonomic resolution, their reproducibility between studies, changing reference data, and their over exaggeration of microbial richness (Callahan et al., 2017; R. C. Edgar, 2017). As such several new methods for addressing sequencing error and creating analytic units for marker gene sequencing have been proposed. These new methods, known as sequencing denoising, aim to identify error prone reads and either remove them from consideration or correct them (Callahan et al., 2017).

The initial development of sequencing denoising methods were motived by the fact that OTUs often fail to resolve ecologically important differences between closely related organisms such as differing species or strains (Eren et al., 2013). Indeed, Oligotyping was the first of these methods and was employed in various datasets; highlighting the ecological importance of uncovering exact sequences within various microbial communities (Eren, Borisy, Huse, & Mark Welch, 2014; Eren et al., 2013). Following Oligotyping, a new tool called Minimum Entropy Decomposition (MED) was later introduced as an unsupervised way of generating exact sequences from marker gene sequencing data. Following the development of MED further improvements in the generation of exact biological sequences referred to as amplicon sequence variants (ASVs) were made. This includes the use of error modeling and the inclusion of larger datasets when tuning parameters.

This led to three separate sequence denoising algorithms to be commonly used within the literature DADA2 (Callahan et al., 2016), Deblur (Amir et al., 2017), and UNOISE3 (R. C. Edgar, 2016). Interestingly, despite similar goals of generating ASVs, each of these methods used unique procedures to identify and correct sequence errors. For example DADA2 uses a data driven approach through the construction of a parametric error model, while both UNOISE3 and Deblur use predefined parameters based on previous Illumina sequencing runs to identify error prone reads (Amir et al., 2017; Callahan et al., 2016; R. C. Edgar, 2016). A greater discussion of these approaches and their differences can be found within the introduction of Chapter 2 of this thesis. However, critically within these descriptions is the point that no clear gold standard within the field was apparent. Indeed, often methods were chosen with little rational and were often compared interchangeable with one another. This fact along with not understanding how tool choice could impact biological conclusions, such as microbial biomarkers, were the main motivators in the research presented within Chapter 2. The goal of which was to identify whether the choice of sequence denoising method could play a significant role in study conclusions such as beta diversity, richness, and overall microbial profile.

1.4 - DNA Sequencing Based Microbiome Measurements

Investigating the drivers of diversity, as well as both microbial and functional composition is a central aspect of understanding microbial communities. Indeed, studying these microbial community measurements have allowed us to investigate how microbial communities change with respect to different environmental pressures and contribute to various aspects including human health. Fortunately, sequence-based microbiome studies are well suited to begin to address questions surrounding microbial diversity and composition. Below, I will discuss common ways of investigating microbial diversity, and composition, using marker gene sequencing. I will also highlight how these sequenced based measurements of microbial communities have been proposed as biomarkers. Finally, while the term species will be used throughout this section, similar concepts can also be applied to OTUs, ASVs, or any other taxonomic classifications generated during marker gene data processing.

Various ways of measuring biodiversity within ecosystems exist. Many of which have been adapted from the field of classical ecology to investigate microbiomes. For example, classical measurements of both alpha and beta diversity have been applied to sequence based microbiome surveys (Tipton, Darcy, & Hynson, 2019). Alpha diversity can be defined as the species diversity within a site of interest (Whittaker, 1972). In the case of microbiome studies sites are often defined as individual samples. There are several different ways of comparing alpha diversity between samples with the simplest form being richness - the number of species observed (Whittaker, 1972). Indeed, in microbiome studies measurements of richness have proven to be a powerful way of summarizing microbial communities. For example, in inflammatory bowel disease microbial richness has been repeatedly shown to be reduced within the gut compared to health controls (Glassner, Abraham, & Quigley, 2020). In fact reductions in microbial richness have been proposed as a useful biomarker of treatment susceptibility and disease flares (Douglas et al., 2018; Lloyd-Price et al., 2019).

Richness, however, is not the only parameter used to measure diversity within samples. Evenness, the spread of abundance across species, is also commonly considered when estimating alpha diversity from sequenced microbial communities. This is because microbial communities with the same richness can have considerably different distributions of abundance within them (Tipton et al., 2019). To address this issue the evenness of a microbiome sample can be measured, where higher scores are associated with samples where species share the same abundance values. By examining evenness researchers can identify communities that are dominated or skewed toward only a small number of species. Oftentimes both microbial richness and evenness can be considered within a single estimate of alpha diversity. Examples of this include both the Shannon index and Simpson index, which are commonly used alpha diversity measurements within microbiome studies (A. D. Willis, 2019).

In addition to both richness and evenness, phylogenetic information can also be included within alpha diversity measurements. Phylogenetic diversity can help account for the evolutionary history between species within a sample. Various forms of phylogenetic diversity measurements exist with Faith's Phylogenetic Diversity being one of the most common (Faith, 1992). Faith's Phylogenetic Diversity does not consider evenness within a sample and simply represents the sum of branch lengths between all species within a sample, given a phylogenetic tree. This metric represents how far the community of interest extends across the evolutionary tree of life (Matsen IV, 2015). It is commonly applied to microbiome studies. For example, measurements of Faith's Phylogenetic Diversity within the oral cavity have been associated with a number of outcomes including ancestry, diabetes, and SARS-CoV-2 infection (Saeb et al., 2019; Yongjian Wu et al., 2021; Yaohua et al., 2019). Similar to measurements of richness other forms of phylogenetic diversity exist that consider species abundance (Allen, Kon, & Bar-Yam, 2009; McCoy & Matsen IV, 2013). These weighted phylogenetic measurements of alpha diversity have proven useful as they are less impacted by low abundance contaminates or sequencing artefacts. Although in the end, which metric of alpha diversity used within a study often comes down to the question at hand due to their unique views of community structure.

Common to all measurements of diversity is the need to investigate each community with the same amount of observation effort. For example, ecologists counting lions on a savannah are likely to count more lions if they spend ten days in the savannah compared to five. While this is a macroscopic example, similar situations apply to sequence based microbial community surveys. This is because total read counts can vary arbitrarily from one sample to another and thus lead to unequal observation effort between samples (Gloor, Macklaim, Pawlowsky-Glahn, & Egozcue, 2017). Without correcting this issue, measurements of diversity such as the number of species within a sample will be biased toward samples with larger read counts. Highlighting the potential to identify spurious findings and or biomarkers (A. D. Willis, 2019).

Multiple solutions to unequal observation efforts in sequencing have been proposed with the issue traditionally being dealt with using rarefaction (S. Weiss et al., 2017). In the context of ecology, rarefaction was first suggested by Howard Sanders in 1968 who suggested randomly discarding samples until the observation effort was the same across all observed communities (Sanders, 1968). In the case of our savannah lion example this might involve randomly discarding data from five of the ten days of our field work. In the case of microbiome studies this means we must randomly discard reads from each sample so that all samples are of equal read depth (McMurdie & Holmes, 2014). This approach has been used to great success within the microbiome field uncovering differences in alpha diversity across not only the human body but also soil, water, and built environments. However, this practice has come under significant criticism due to the amount of data being unused and the disregard for measurements of uncertainty (McMurdie & Holmes, 2014; A. D. Willis, 2019). Indeed, others have argued that read counts represent the certainty in which we can estimate the diversity within a sample, a sensible conclusion given the relation between read depth and discovery of novel taxa. Methods that exploit read count information have been developed to model both the number of observed and unobserved species within a population (A. Willis, Bunge, & Whitman, 2017). While these methods are not as commonplace, improvements in their estimations could significantly benefit our understanding of microbiome diversity.

Beta diversity is another common category of diversity metrics used to describe and compare microbial communities. Beta diversity can be defined as measuring the diversity between groups of samples and is often represented by various dissimilarity and distance metrics (Tipton et al., 2019). In this case, samples with the exact same microbial profiles are measured with a distance or dissimilarity of 0 and reach closer to 1 as they differ from one another. Beta diversity metrics are commonly used to describe how groupings of microbiome samples differ in their overall community profile or structure (Pollock et al., 2018). Like alpha diversity metrics, beta diversity metrics can also consider both the presence/absence of species and/or their abundance profiles. For example, the Jaccard distance of a group of samples only considers whether a microbe is present or absent and is thus considered an unweighted beta diversity metric (Parks & Beiko, 2013). Contrastingly, Bray-Curtis dissimilarity considers both the presence/absence of species as well as their abundance. Because it considers both abundance and presence/absence it is thought of as a weighted beta diversity metric. Debate on which metrics better represent microbial communities have been ongoing within the field, however, each has its own benefits (Parks & Beiko, 2013). For example, unweighted measurements give a better understanding of how microbial communities differ when both are dominated by similar species. However, like measures of pure richness it is thought to be at higher risk for technical bias due to the equal weighting of low abundance contaminants or sequencing artifacts. Nonetheless, both have been useful for helping to identify microbial populations that could potentially harbor biomarkers of disease (Goodrich, Di Rienzi, et al., 2014).

It is common in microbiome literature to add phylogenetic information into beta diversity metrics. Both weighted and unweighted UniFrac were first proposed by Louzpone et al. to allow researchers to capture information on how similar microbial communities are in the context of evolutionary history (C Lozupone & Knight, 2005; Catherine Lozupone, Lladser, Knights, Stombaugh, & Knight, 2011). These techniques are particularly powerful in marker gene studies as common targets such as the 16S rRNA gene can capture a significant amount of phylogenetic information between species. Both weighted UniFrac and unweighted UniFrac are ubiquitously used across the microbiome field in conjunction with rarefaction to address issues of unequal sampling efforts as described above. For example, work has shown that gut microbiota unweighted UniFrac profiles of individuals with colon cancer are significantly different from health controls (Flemer et al., 2017).

Traditionally rarefaction is used to address differences in observation efforts when calculating beta diversity metrics. However this practice has recently come under scrutiny (Cameron et al., 2019; McMurdie & Holmes, 2014). Accordingly, new ways of measuring beta diversity without the use of rarefaction have been proposed. The majority of which depend on compositional data (CoDa) analysis techniques first proposed by Aitchison to resolve issues of compositionality (Aitchison, 1982). Indeed, microbiome profiles observed through DNA sequencing are inherently compositional in nature and as such are generally measured in relative abundances. This is because sequencing instruments have a maximum capacity in the number of unique DNA fragments, they can process. Accordingly, when this maximum is reached, reads sequenced from one species must take up space in the maximum capacity resulting in a lower number of available reads for the remainder (Gloor et al., 2017). This is important because increases in relative abundance do not necessarily mean that there is a corresponding increase in absolute abundance, a common issue highlighted within CoDa literature.

To help address this issue microbiome profiles have been represented using various different normalizations based on ratios (Morton et al., 2019). The most common of which is the center-log-ratio which compares the read counts of each species within a sample to the geometric mean read count across all species within that sample. In this scenario we can begin to interpret microbiome profiles by comparing how species increased or decreased compared to the “average microbe” rather than examining differences in proportions. Center-log-ratio normalization techniques along with others such as the iterative-log-ratio have been not only used to normalize microbiome profiles but also to generate various beta diversity measures (Cameron et al., 2019; Silverman, Washburne, Mukherjee, & David, 2017; Washburne et al., 2017).

Unfortunately, the use of ratio normalizations and CoDa analysis is not without its own faults. For example, center-log-ratio methods cannot deal with zeros in the numerator and as such must either use some form of a heuristic or a pseudo-count. Unfortunately, the use of pseudo-counts is not without fault and remains an active area of research (S. Weiss et al., 2017). Another example is the use of additive log ratios which focus on a single taxon as the reference frame or ratio denominator. In this case choosing a single taxon to do this can often be difficult due to the high sparsity of microbiome data and the uncertainty of the underlying microbial abundances. However, solutions to this issue have been proposed including the use of multiple reference frames or known spike-in quantities (McLaren, Nearing, Willis, Lloyd, & Callahan, 2022; Morton et al., 2019).

Overall, there is no one way of analyzing microbial diversity and often it is unclear which metric may be best suited for the question at hand. This leaves researchers with yet another choice in how to view the observed microbial communities within their sample.

A second approach to identify differences in microbial community structure is through the use of various machine learning algorithms. Machine learning has proven to be useful in the microbiome field in several instances. Chiefly machine learning has been used in microbiome literature to either identify different community types, or classify samples based on microbial profile. Although numerous other applications exist including in taxonomic classification (Qiong, M., M., & R., 2007), and protein structure prediction (Jumper et al., 2021).

Eco-typing or community typing is a common technique used in microbiome studies to identify whether samples fall within similar community profiles. For example, one community type may be characterized by the high abundance of several phyla while another may be dominated by a single species. Community typing is generally done using various unsupervised machine learning algorithms with k-means clustering being one of the most common. The use of community typing has led to several interesting results including the identification of specific microbial patterns associated with infant probiotic usage and microbiome maturation (Samara et al., 2022). Overall, community typing represents a powerful technique to identify whether broad community patterns are associated with environmental or host conditions.

Machine learning has also been successfully applied to microbiome data in numerous cases to classify host phenotypes (Marcos-Zambrano et al., 2021). Various different supervised machine learning algorithms have been applied to microbiome data including support vector machines, artificial neural networks, and Random Forests (Marcos-Zambrano et al., 2021). In practice these algorithms have been applied to both taxonomic data as well as functional data to classify various sample characteristics (Douglas & Langille, 2021). Common examples of supervised learning with microbiome data include the prediction of individuals with and without disease. For instant Flemer et al., used oral microbiome data and Random Forest modeling to classify individuals with and without colorectal cancer (Flemer et al., 2017). Underscoring their potential use in diagnostic applications. Moreover, machine learning algorithms such as Random Forest can also be used to identify species/functions or groups of species/function that carry signal with the host phenotype of interest. This is nicely illustrated by Thomas et al., who used Random Forest modelling across seven different studies to identify associations between colon cancer and choline trimethylamine-lyases gene abundance within the gut microbiome (Thomas et al., 2019).

In addition to using machine learning to identify taxa of interest, various bioinformatic tools have been developed with the single purpose of identifying taxa-metadata relations. Indeed, one of the most common questions that is asked during a microbiome study is whether particular taxa are associated with host characteristics (McLaren et al., 2022). Common examples of this might include comparing the abundances of microbes within the oral cavity of healthy individuals to individuals with disease to uncover microbial biomarkers. In the most general form these types of analyzes are referred to as differential abundance (DA) analysis. Indeed, DA analysis has been applied to numerous microbiome studies ranging from the association of specific microbes to pH levels, to whether taxa are associated with human cancers. Interestingly, despite how straightforward DA analysis may seem, a number of different tools have been developed specifically for this purpose (Morton et al., 2019). In fact various normalizations based on quantiles (Law, Chen, Shi, & Smyth, 2014; Love, Huber, & Anders, 2014), means (Calgaro, Romualdi, Waldron, Risso, & Vitulo, 2020; Robinson & Oshlack, 2010), rarefaction (S. J. Weiss et al., 2015), and ratios (Fernandes et al., 2014b;

Mandal et al., 2015) have been proposed to be used in microbiome DA analysis. Moreover, differing modelling strategies have also been proposed using various distributions including beta-binomial (B. D. Martin, Witten, & Willis, 2020), negative binomial (Love et al., 2014), normal, and zero-inflated normal (Paulson, Stine, Bravo, & Pop, 2013). These differences highlight the variability among microbiome DA methods, yet they are often used interchangeable within the literature. We believe that the interchangeable use of DA methods has contributed to difficulties in reproducing microbiome associations between studies of the same disease.

Despite these differences, in general DA tools fall within into two differing categories; those that adhere to CoDa principals using ratio normalizations and those that are count based. Tools that use CoDa principals include ALDEx2 (Fernandes et al., 2014a) and ANCOM (Mandal et al., 2015) which use either center-log-ratios or additive log ratios by default. Contrastingly, count based tools attempt to identify taxon associations by examining proportions or normalizing read library sizes across the samples of interest. Clearly these different strategies are attempting to address fundamentally different questions, with CoDa tools examining the abundance of taxa relative to another feature and count based tools examining either proportional changes or normalized counts. Whether a particular normalization is best for the identification of microbiome-based biomarkers is unclear. For example, multiple studies have examined the gut microbiome and colon cancer with differing results (Cheng, Ling, & Li, 2020). While there are several reasons as to why their results may not align such as differences in DNA extraction or variable region, another contributing factor could be the usage of differing DA tools. This highlights the need to identify how DA tool choice might impact the biomarkers being detected within the same study.

While previous works have already shown that the specificity and sensitivity could differ between tools during simulations of different microbiome profiles (Calgaro et al., 2020; Cappellato, Baruzzo, & Di Camillo, 2022; Ma et al., 2021; S. Weiss et al., 2017). They lacked comprehensive investigations into what degree DA tool choice could impact the results from data generated from true microbiome specimens. To help address this, myself and colleagues examined 38 16S rRNA gene sequencing datasets within Chapter 5 of this thesis. The overarching goal of this analysis was to determine how similar results were between DA tools and the consistency of those results.

1.5 - Outlook

Based on the number of decisions and the lack of clear gold standards within DNA based microbiome data analysis, there will clearly be at least some minor differences between studies. However, these differences have not been well characterized for numerous steps across experimental protocols, especially with regards to computational analysis. Indeed, several studies have now compared differing DNA extraction kits (Shaffer et al., 2021), sequencing primers (C. Willis et al., 2019), and even sequencer platforms (D'Amore et al., 2016) to begin to identify how they systematically differ from one another. Yet the number of computational choices, a researcher must make during microbiome investigations represents a plethora of combinations. Seemingly the hope of researchers is that the biomarkers and conclusions we draw from studies are robust to these computational choices.

The goal of this thesis was twofold. One to attempt to begin characterizing how critical bioinformatic tool choices are to microbial biomarker detection, and to then attempt to use this information to find more robust biomarkers within the salivary microbiome of breast, prostate, and colon cancer. To address the former goal, two different steps in microbiome biomarker inference were selected to be examined for impacts on the biological conclusions drawn. Both of which represent what we hypothesize as being some of the most critical choices made during the computational analysis of microbiome sequencing data. The first choice we examined was how researchers address sequencing error using OTUs or different sequence denoising methods. Results from this investigation are highlighted within Chapter 2. Following this we were also interested in DA analysis and accordingly examined DA tool choice in Chapter 3.

Following these investigations, we attempted to identify our own salivary biomarkers by first characterizing the microbial variation within the saliva of healthy adults (Chapter 4). We then applied this knowledge to identify biomarkers of cancer in Chapter 5.

My hope is that by reading this thesis researchers will be more aware of the impact of the choices they make on the biological conclusions that they draw from microbiome studies. I also hope that some of the biomarkers identified with Chapter 5 of this thesis can be applied on a broader scale to validate their findings and potentially lead to novel risk factors for the development of prostate, colon, and breast cancer.

Chapter 2 - Denoising the Denoisers: An Independent Evaluation of Microbiome Sequence Error-Correction Approaches

This chapter is a reproduction of the paper of the same named published in the journal PeerJ (Nearing, Douglas, Comeau, & Langille, 2018). I was first author on this work and was the primary contributor toward designing the project, collecting data, and analyzing the results. I wrote the paper while receiving constructive feedback from my co-authors. The co-authors on this paper were: Gavin M. Douglas, André M. Comeau, and Morgan G.I. Langille.

This paper was published under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, reproduction, and adaptation in any medium and for any purpose provided this it is properly attributed. Original declarations of competing interests, author contributions, funding, and data availability are available within the original publication. All supplemental figures and files referred to in this chapter are freely available as part of the publication on the PeerJ journal website. Access to these supplemental files can be found at the following link: https://doi.org/10.7717/peerj.5364.

2.1 - Abstract

High-depth sequencing of universal marker genes such as the 16S rRNA gene is a common strategy to profile microbial communities. Traditionally, sequence reads are clustered into operational taxonomic units (OTUs) at a defined identity threshold to avoid sequencing errors generating spurious taxonomic units. However, there have been numerous bioinformatic packages recently released that attempt to correct sequencing errors to determine real biological sequences at single nucleotide resolution by generating amplicon sequence variants (ASVs). As more researchers begin to use high resolution ASVs, there is a need for an in-depth and unbiased comparison of these novel "denoising" pipelines. In this study, we conduct a thorough comparison of three of the most widely-used denoising packages (DADA2, UNOISE3, and Deblur) as well as an open-reference 97% OTU clustering pipeline on mock, soil, and host-associated communities. We found from the mock community analyses that although they produced similar microbial compositions based on relative abundance, the approaches identified vastly different numbers of ASVs that significantly impact alpha diversity metrics. Our analysis on real datasets using recommended settings for each denoising pipeline also showed that the three packages were consistent in their per-sample compositions, resulting in only minor differences based on weighted UniFrac and Bray–Curtis dissimilarity. DADA2 tended to find more ASVs than the other two denoising pipelines when analyzing both the real soil data and two other host-associated datasets, suggesting that it could be better at finding rare organisms, but at the expense of possible false positives. The open-reference OTU clustering approach identified considerably more OTUs in comparison to the number of ASVs from the denoising pipelines in all datasets

tested. The three denoising approaches were significantly different in their run times, with UNOISE3 running greater than 1,200 and 15 times faster than DADA2 and Deblur, respectively. Our findings indicate that, although all pipelines result in similar general community structure, the number of ASVs/OTUs and resulting alpha-diversity metrics varies considerably and should be considered when attempting to identify rare organisms from possible background noise.

2.2 - Introduction

Microbiome studies often use an amplicon sequencing approach where a single genomic region is sequenced at a sufficient depth to provide relative abundance profiles of the microbes present in a sample. The 16S rRNA gene (16S) is usually chosen as a marker gene for sequencing of bacterial communities due to its unique structure that contains both conserved and variable regions and presence in all known Bacteria and Archaea species. This sequencing approach is often used to avoid the high cost of shotgun metagenomic sequencing or to avoid problems with sequencing non-microbial DNA from host contamination. However, sequencing errors make it difficult to distinguish biologically real nucleotide differences in 16S sequences from sequencing artefacts. To avoid this issue sequences are often clustered into operational taxonomic units (OTUs) at a particular identity threshold (e.g. 97%) to avoid the problem of differentiating biological from technical sequence variations; however, this comes at the cost of taxonomic resolution. Recently, many new bioinformatic sequence "denoising" approaches have been developed to address this issue by attempting to correct sequencing errors thus improving taxonomic resolution. These pipelines differ in how they correct sequencing errors. DADA2 generates a parametric error model that is trained on the entire sequencing run and then applies that model to correct and collapse the sequence errors into what the authors call amplicon sequence variants (ASVs) (Callahan et al., 2016). This approach is advantageous as it builds unique error models for each sequencing run. Deblur aligns sequences together into "sub-OTUs" and, based on an upper error rate bound along with a constant probability of indels and the mean read error rate, removes predicted error-derived reads from neighboring sequences (Amir et al., 2017). Deblur employs a sample-by-sample approach which reduces both memory requirements and computational demand. UNOISE3 employees a one-pass clustering strategy that does not depend on quality scores, but rather two parameters with pre-set values that were curated by its author to generate “zero-radius OTUs” (R. C. Edgar, 2016). The advantage of a one-pass clustering strategy is that it saves on the computational time required to analyze the sequences in the provided study. Note that ASVs, sub-OTUs, and zero-radius OTUs are synonymous and the term ASV will be used henceforth. Denoising approaches provide improved resolution and they avoid having to make a choice between various OTU strategies which may result in differing results (R. C. Edgar, 2017). In addition, ASVs can be identified by their unique biological sequences instead of relying on per-study IDs, which allows for easier comparison across datasets (Callahan et al., 2017).

Although there have been several bioinformatic comparisons of OTU-based approaches (Allali et al., 2017; Plummer & Twin, 2015), a thorough third-party comparison of denoising pipelines has yet to be conducted. In this chapter, we compare the strengths and weaknesses of the DADA2, UNOISE3, and Deblur packages along with a comparison to an open-reference 97% OTU-based approach (Rognes et al., 2016), while following the recommended default settings. We assess the accuracy of these approaches using several mock communities including both bacterial and fungal amplicons. In addition, we compare the results of the four pipelines on three previously-published real human, mouse, and soil datasets.

2.3 - Materials and Methods

2.3.1 - Sequence Acquisition

The HMP mock community (which has even expected abundances) and the ZymoBIOMICS Microbial Community Standard (referred to as the Zymomock community) were sequenced by the Integrated Microbiome Resource at Dalhousie University using an Illumina MiSeq on separate sequencing runs, as previously described using the V4-V5 16S rRNA gene region (Comeau et al., 2017). Reads were then uploaded to the European Nucleotide Archive (ENA) under accession number PRJEB24409. The Extreme dataset (mock-12) originally presented in the DADA2 paper and the fungal ITS1 dataset (mock-9) were retrieved from the Mockrobiota project (Bokulich et al., 2016). The Extreme dataset was sequenced using an Illumina MiSeq (Callahan et al., 2016) and the fungal mock community was sequenced using an Illumina HiSeq (Bokulich et al., 2016). The soil data set collected from blueberry fields (available under NCBI SRA PRJNA389786) and the exercise dataset collected from stool of mice that exercised plus controls (ENA accession PRJEB18615) (Lamoureux, Grandy, & Langille, 2017) as well as the human-associated dataset of intestinal biopsies of pediatric Crohn's disease patients plus controls (ENA accession PRJEB21933) (Douglas et al., 2018) were sequenced at the Integrated Microbiome Resource at Dalhousie University.

2.3.2 - Filtering

All sample data was filtered using the Microbiome Helper filtering scripts (Comeau et al., 2017). In summary, primers were trimmed off all reads using Cutadapt (v 1.14) (M. Martin, 2011) and GNU Parallel (Tange, 2011). Primer-free sequences were then input into the dada2_filter.R script available in Microbiome Helper. This script takes in the maximum expected number of errors allowed as well as a truncation length. The HMP mock community and the Zymomock community were truncated to read lengths of 270 and 210 base pairs for the forward and reverse reads respectively to remove low quality bases at the end of the reads. The shorter length for the reverse reads is a result of their lower overall quality compared to the forward reads. Note that for DADA2, and Deblur to work properly reads need to be of the same length. The single-end reads from the Extreme mock community and the fungal mock community were truncated to 80 base-pairs. The three real datasets: soil, mouse, and human-associated, were truncated to 270 and 210 base-pairs for the forward and reverse reads, respectively. The number of expected errors allowed were defined as three different filtering stringencies: 5 (low), 3 (medium), and 1 (high).

2.3.3 - DADA2 Pipeline

The DADA2 pipeline was run using Microbiome Helper scripts, which wraps the core algorithms of the DADA2 pipeline (Callahan et al., 2016). Filtered reads were input into the wrapper script dada2_inference.R which runs the DADA2 inference algorithm. Once ASVs are determined, they are passed into DADA2's chimera-checking algorithm which was run using the wrapper script dada2_chimera_taxa.R to screen out chimeric sequences. The output objects containing ASV sequences and abundances counts were then converted into BIOM table format using convert_dada2_out.R. All DADA2 wrapper scripts were run with default settings.

2.3.4 - UNOISE3 Pipeline

Filtered reads were input into USEARCH's (v 10) (R. C. Edgar, 2010) fastq_mergepairs command if they were paired-end reads or concatenated together into one FASTQ if they were single-end reads. Next, the merged FASTQ file was converted into a FASTA file using the Microbiome Helper script run_fastq_to_fasta.pl and then used as input for USEARCH's fastx_uniques command which generated a FASTA containing all the unique sequences and their abundance. Finally, the FASTA containing unique sequences was used as input into USEARCH's unoise3 (R. C. Edgar, 2016) command generating a BIOM table and representative ASVs that were used in subsequent analyses. All USEARCH scripts were run with default settings.

2.3.5 - Deblur Pipeline

Paired-end filtered reads were stitched together using the Microbiome Helper wrapper script run_pear.pl which wraps the program PEAR (v 0.9.10) (J. Zhang, Kobert, Flouri, & Stamatakis, 2014). This step was skipped for filtered single-end reads. Next, reads were renamed to match a format that was compatible with QIIME2 (Caporaso et al., 2010) and converted into a QIIME2 artifact. Samples were then run through QIIME2's built-in deblur command using the 16S setting, which uses Greengenes 13_8 (DeSantis et al., 2006) for positive filtering. A separate analysis to test the effect of positive filtering was conducted by using the non-positive filtered output files ("all.biom" and "all.seqs.fa") from the stand alone version of Deblur. Fungal reads were run using the "other" setting and the UNITE 10.10.2017 database (Kõljalg et al., 2013). Finally, the representative

ASV sequences and a BIOM table were exported from the QIIME2 artifact. All other Deblur scripts were run with default settings.

2.3.6 - Open Reference OTU Pipeline

Filtered reads were stitched and imported into a QIIME2 artifact as previously described in the Deblur pipeline. The imported sequences were then dereplicated using the QIIME2 VSEARCH plugin. Dereplicated sequences were filtered for chimeras using the QIIME2 VSEARCH UCHIME reference-based chimera filtering (R. C. Edgar, Haas, Clemente, Quince, & Knight, 2011) using the Greengenes13_8 97% OTU database for reference. Chimera-checked sequences were then clustered at 97% in an open-reference fashion using the QIIME2 VSEARCH cluster-features-open-reference plugin. The Greengenes 13_8 97% OTU database was used as the reference database during open-reference OTU picking. Finally, singleton OTUs were removed from the OTU table and the table was exported in BIOM format along with its representative sequences in FASTA format.

2.3.7 - Run Time and Memory Analysis

The soil dataset was filtered using the low-stringency filter and then individual samples were rarefied to either 5000, 10000, 20000 or 30000 reads per sample. The different read-depth sets were then run through the three denoising pipelines and user time and maximum memory usage was determined using the GNU time (v 1.7) command.

2.3.8 - ASV and OTU Analysis of Mock Communities

ASVs and OTUs were compared against the expected sequences provided with each of the mock communities. This comparison was done using the command-line BLASTN (v 2.7.10) (Altschul, Gish, Miller, Myers, & Lipman, 1990) tool against the expected sequences from each community. This allowed us to determine the number of full length 100% matches and 97% matches. All ASVs and OTUs that did not match these criteria were then compared against the SILVA 16S rRNA gene database (v 128) (Pruesse et al., 2007) or the UNITE 10.10.2017 database to find all full length 100% and 97% matches. Any ASVs or OTUs that did not match these databases were then labeled as "Unmatched". The number of unique expected sequences for each mock community was determined by slicing out the amplified regions using a custom Python (v 3.6.1) script (slice_amplified_region.py, available on GitHub) from the expected sequences of each mock community and then finding the number of unique sequences from this output using USEARCH10's fastx_uniques command. In addition to analysis done with no abundance filtering, we also compared how filtering of low abundance ASVs/OTUs affected the type and amount of ASVs/OTUs called by each pipeline. This was done by applying a 0.1% minimum abundance filter to each approach and dataset except for the Extreme mock community where a 0.0004% minimum abundance filter was applied. The 0.1% minimum abundance filter was chosen based on the known 0.1% bleed through between Illumina MiSeq runs. A lower minimum abundance filter of 0.0004% was chosen for the Extreme mock community since some of the expected sequences had a lower expected abundance than 0.1%. Histograms with sequence identity to the expected sequences were generated for each pipeline and each mock community except for the fungal mock community due to the community missing some expected sequences from its expected sequence list. All sequences that showed less than 75% identity were binned together.

2.3.9 - Abundance Data Analysis of Mock Communities

For the HMP, Zymomock and Extreme datasets all ASVs/OTUs that matched at 97% identity or greater with the provided expected sequences based on a BLASTN search were added to the abundance of the corresponding expected taxa. Stacked bar charts of expected taxa relative abundances were created using the ggplot2 (v 2.2.1) (Wickham, 2009) R (v 3.4.3) (R Development Core Team, 2008) package and the cowplot (v 0.9.2) R package (see Data Deposition for exact scripts).

Due to the incomplete nature of the expected sequences for the fungal mock community, UNITE database hits at 97% or greater to an expected species within the community were considered as expected ASVs/OTUs. All other ASVs/OTUs were classified as “Non-Reference” hits.

2.3.10 - Analysis of Real Datasets

Sequences from the three real datasets; soil, human-associated, and exercise were filtered using medium stringencies (allowing up to 3 expected errors per sequence) for each package and rarified to 5215, 3000 and 4259 reads, respectively. These rarefaction levels were chosen as they were the lowest read count above 2000, after the DADA2 pipeline was complete. ASV abundance tables output by all approaches were combined into a single table where each biological sample was represented multiple times (once for each pipeline). ASVs/OTUs not called by a specific pipeline were given an abundance of zero in their column (e.g. ASVs only called by Deblur for sampleA were given zero abundances in the columns for DADA2's and UNOISE3's outputs of sampleA). Representative sequences were aligned against the Greengenes 13_5 99% OTU aligned sequences and placed on the Greengenes 13_5 99% OTU tree using SEPP (Mirarab et al., 2011). Weighted UniFrac and unweighted UniFrac distances at the ASV/OTU level were then generated using the QIIME1 beta_diversity.py command and principal coordinates were generated using the QIIME1 prinicpal_cordinates.py command.

The Bray-Curtis distance matrix at the genus level was generated by assigning taxonomy to the resulting ASV/OTUs from each package using the RDP classifier (Cole et al., 2014) with the assignTaxonomy function available in the DADA2 package and the rdp_train_set_16 database. Distances were then generated using the summarize_taxa.py and beta_diversity.py commands in QIIME1. Ordination was generated using the metaMDS function in the vegan R package (Oksanen et al., 2018).

Mantel correlations between each distance matrix for each pipeline and each beta-diversity metric (weighted/unweighted UniFrac and Bray-Curtis) were generated using the vegan R package. The number of observed OTUs/ASVs per sample for each real dataset were generated using the rarified combined OTU/ASV tables and the alpha_diversity.py command in QIIME1 using observed OTUs as the metric option.

2.4 – Results

2.4.1 - Differences in Methodology and Availability Between Different Denoising Pipelines

There are important aspects other than accuracy that need to be considered when determining which denoising package a researcher should use for their project. Both DADA2 and UNOISE3 are suggested to be run in a pooled-sample workflow, where all sequences are pooled together during the denoising process (Table 2.1). This allows them to better account for batch errors across multi-run experiments. Deblur, on the other hand, runs its denoising process sample-by-sample. This approach helps lower Deblur's computational requirements, but at the cost of reducing its ability to correct multi-run batch effects. Both DADA2 and Deblur are open source projects, whereas UNOISE3 is a closed-source project which has a free 32-bit academic version with a 4 Gb memory cap and a full 64-bit version that costs between $885-1485 USD (Table 2.1). Another major difference is that the built-in Deblur function in QIIME2 has a positive filtering process. This default setting causes Deblur to discard reads that do not reach a length-scaled bit score and an e-value threshold to any sequence in the 88% representative sequences Greengenes database. Note the default database can be changed using the "other" version of the Deblur plugin in QIIME2, an important feature when working with fungal or eukaryotic data. It is also important to note that the stand-alone version of Deblur outputs both positively filtered and non-filtered results by default, unlike the QIIME2 plugin which, only outputs the positively filtered results. Currently, the functionality of both DADA2 and Deblur can be accessed through a graphical user interface as plugins in

QIIME2 Studio, whereas UNOISE3 does not support a graphical user interface (Table 2.1).

Table 2.1: ***Qualitative comparison of DADA2, Deblur, and UNOISE3***

Pipeline	Implemented In	Open Source	*Pooled Sampling	**Positive Filtering	Version Tested	GUI via QIIME2	Publication Date
DADA2	R	Yes	Yes	No	1.6	Yes	April 13 2016
Deblur	Python	Yes	No	Yes	1.0.2	Yes	March 7, 2017
UNOISE3	C++	No	Yes	No	3	No	Oct 15, 2016

** When all sequences from all samples are denoised at the same time (in contrast to running each sample separately).*

*** Compares resulting ASVs to a database (Greengenes for Deblur) and discards reads if they do not match a certain identity threshold (88% for Deblur).*

2.4.2 - Total Number of ASVs/OTUs Vary Across Approaches in Mock Communities

We processed four different mock communities with the DADA2, UNOISE3, and Deblur denoising pipelines as well as an open-reference 97% OTU clustering pipeline to compare the resulting ASVs/OTUs from each approach. Focusing on the number of called ASVs from each denoising pipeline, we found no approach consistently called more ASVs. DADA2 called the most ASVs in two mock communities (HMP: 42, Extreme: 78) and UNOISE3 called the most ASVs in the other two mock communities (Fungal: 38, Zymomock: 43) under medium stringency filtering (Figure 2.1, Supplemental Table 1). Overall, open-reference OTU clustering output the most ASVs/OTUs for all four mock communities (HMP: 453, Extreme: 8891, Fungal: 96, Zymomock: 294). None of the approaches were capable at outputting all expected sequences at 100% identity in any of the mock communities that were processed and, in all datasets, at least one denoising pipeline output more ASVs than expected sequences. All four approaches output at least one ASV/OTU at 97% or greater identity from all organisms in the HMP mock community and the Zymomock community (Supplemental Tables 2-3). DADA2 output nine more ASVs with 97% or greater identity matches to expected sequences in the Extreme dataset (read depth of 11.6 million reads) than the other two denoising pipelines (Supplemental Table 4). Six of the nine taxa that DADA2 called, and the other pipelines did not call, had expected relative abundances of only 0.000427% (Supplemental Table 4). Of the other three taxa not found by Deblur or UNOISE3, one was at an expected abundance of 0.00427% and two had an expected abundance of 0.0427% (Supplemental Table 4). Interestingly, open reference OTU clustering called OTUs corresponding to these nine expected taxa along with two more expected taxa that DADA2 missed, both of which were in the lowest possible expected abundance range (0.000427%). One organism was missed by all pipelines in the Extreme community (*C. methylpentusum*), which was also in the lowest possible expected abundance range. Sequence identity histograms were constructed for the HMP, Zymomock and Extreme communities (Supplemental Figures 1-3). In the HMP community, no ASV was found to have below 80% identity by the denoising pipelines, but open reference OTU clustering found 21 OTUs below 80% identity. Furthermore, the OTU pipeline called 324 OTUs below 99% identity whereas all denoising approaches called less than 15 ASVs (Supplemental Figure 1). In the Extreme community, most ASVs were found in the 99-100% identity range whereas the majority of OTUs found by open-reference OTU clustering were in the 95-97% identity range (Supplemental Figure 2).

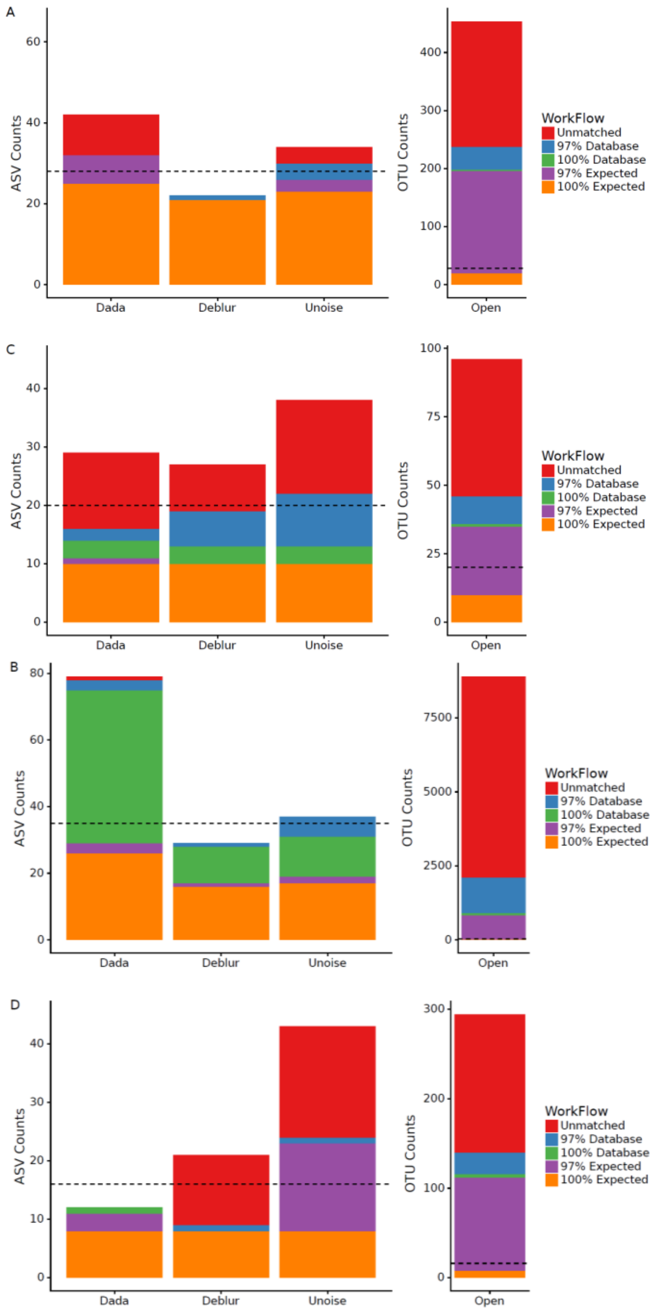


Figure 2.1: ***Total number of ASVs/OTUs identified by each sequence processing method for four different mock communities.*** *Amplicon sequence variants/Operational*

taxonomic units (ASVs/OTUs) were compared to a database of full-length amplicon sequences for just the microbes supposedly in the community ("Expected") and against the full SILVA or ITS databases ("Database") using BLASTN at 97% and 100% identity cut-offs. "Unmatched" sequences did not match an expected sequence or the SILVA/ITS databases at 97% identity or greater. Dotted lines indicate the total number of ASVs/OTUs expected, accounting for 16S copy variation within genomes. Note that the y-axis for open-reference OTU clustering is different than the y-axis on the denoising methods. (A) Human Microbiome Project equal abundance mock community; (B) Extreme dataset; (C) Fungal ITS1 mock community; (D) Zymomock community.

Due to the expected sequences for the fungal community being incomplete, we included any UNITE database hits as expected sequences if the matching sequence was from a species that was included in the fungal community. This resulted in almost all the fungi present in the community to be found by all four pipelines except for *Penicillium allii* and *P. commune*. Open-reference OTU clustering did find *P. allii*, although at an Extremely low abundance of 0.004% compared to its expected abundance of 6.25%.

Given that some of the above potential spurious ASVs would be removed by sequence bleed-through (Illumina, n.d.) or low-abundance filters in typical workflows, we applied an abundance cutoff filter of 0.1% to each mock community except for the Extreme community, where an abundance filter of 0.0004% was applied (Supplemental Figure 4). This lower abundance filter was chosen due to some expected sequences being in abundance of 0.000427%. Application of the abundance filter to the HMP community resulted in all 10 unmatched ASVs (those that did not match either the expected or SILVA by 97% or greater) called by DADA2 to be discarded, but none of the four unmatched reads in UNOISE3 to be discarded (Supplemental Figure 4). A similar phenomenon was seen in the Zymomock community with all 12 of Deblur's unmatched reads being discarded (along with one database hit) and UNOISE3 only discarding one of 19 unmatched reads it called. The largest effect that the application of the filters had was the removal of a significant amount of OTUs found in each mock community by the open-reference OTU pipeline. The most drastic change was seen in the Extreme community that started with 8891 OTUs and was reduced to 1248 OTUs (Supplemental Figure 4). In the fungal community the number of OTUs found by open-reference OTU clustering (27) became less than the number of ASVs found by UNOISE3 (36).

To determine how read quality filtering affects the number of ASVs called by each pipeline, we ran each denoising pipelines using two additional quality filtering stringencies, low and high (see Methods). The different quality filter stringencies used made only small impacts on the numbers of ASVs called by each pipeline for the HMP, Extreme and fungal datasets (Supplemental Table 1). A difference of six ASVs was the largest between the high and medium stringencies on the HMP community and was output by UNOISE3 (Supplemental Table 1). In the Zymomock community, the number of ASVs called by DADA2 only varied by one for all three stringencies, but Deblur varied by as much as 12 ASVs and UNOISE3 varied by as much as 16 ASVs being outputted between the high and medium filter stringencies (Supplemental Table 1).

2.4.3 - Denoising Pipelines are Consistent in Determining Mock Community Composition

Despite the different ASV counts between each pipeline in the mock community, the relative abundances of the expected taxa are strikingly similar (Figure 2.2). In both the

HMP and Zymomock communities, only a small portion of ASVs called by DADA2 and Deblur did not match the expected sequences by 97% identity or greater. In contrast, UNOISE3 identified multiple (8 in HMP, 20 in Zymomock) sequences that summed together to make up 2.9% and 4.8% of the relative abundance in the HMP and Zymomock communities, respectively (Supplemental Tables 1-3). Open-reference OTU clustering found 6783 non-reference OTUs (Supplemental Table 1) in the Extreme community that summed together to make up 2.6% of the community, whereas the denoising approaches all found non-reference abundances less than or equal to 0.3% (Supplemental Table 4). None of the approaches were good at distinguishing the proper abundances of the two *Parabacteroides distasonis* strains with denoising pipelines finding similar proportions of both strains and the open-reference OTU pipeline finding dominance of the 13400 strain and not the 13401 strain (Figure 2.2B, Supplemental Table 4). None of the approaches performed well at matching the expected abundance of the Zymomock community or the fungal community (Figure 2.2). All three denoising pipelines called over-abundances of *Lactobacillus fermentum* in the Zymomock community, with Deblur calling the most (44.5%) and UNOISE3 calling the least (37.6%) (Supplemental Table 3). Similarly, all approaches called non-reference hits in greater than 9% abundance in the fungal community (Supplemental Table 5), with UNOISE3 calling the most non-reference hits (12.4%) and Deblur calling the least (9.8%). Due to all four pipelines producing similar proportions of non-reference hits in the fungal community, this could suggest that either the mock compositions are not in the expected proportions, the four pipelines are similarly biased, or that an upstream process during sequencing caused the introduction of unexpected sequences.

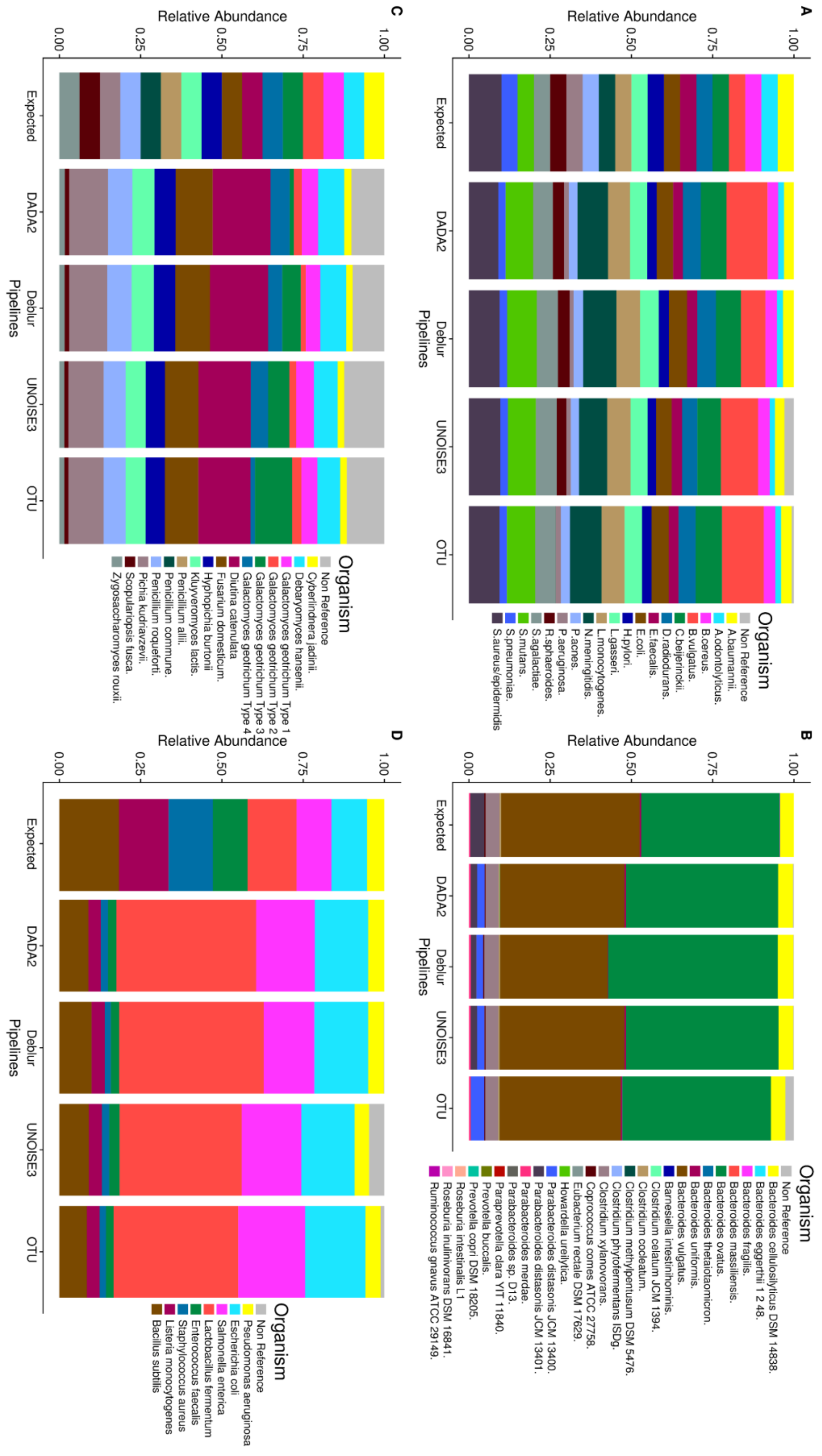
A
Relative Abundance
0.00
0.25
0.50
0.75
1.00
Expected
DADA2
Deblur
UNOISE3
OTU
Pipelines
Organism
Non Reference
A.baumannii.
A.odontolyticus.
B.cereus.
B.vulgatus.
C.beijerinckii.
D.radiodurans.
E.faecalis.
E.coli.
H.pylori.
L.gasseri.
L.monocytogenes.
N.meningitidis.
P.acnes.
P.aeruginosa.
R.sphaeroides.
S.agalactiae.
S.mutans.
S.pneumoniae.
S.aureus/epidermidis
B
Relative Abundance
0.00
0.25
0.50
0.75
1.00
Expected
DADA2
Deblur
UNOISE3
OTU
Pipelines
Organism
Non Reference
Bacteroides cellulosilyticus DSM 14838.
Bacteroides eggerthii 1 2 48.
Bacteroides fragilis.
Bacteroides massiliensis.
Bacteroides ovatus.
Bacteroides thetaiotaomicron.
Bacteroides uniformis.
Bacteroides vulgatus.
Barnesiella intestinihominis.
Clostridium celatum JCM 1394.
Clostridium cocleatum.
Clostridium methylpentusum DSM 5476.
Clostridium phytofermentans ISDg.
Clostridium xylanovorans.
Coprococcus comes ATCC 27758.
Eubacterium rectale DSM 17629.
Howardella ureilytica.
Parabacteroides distasonis JCM 13400.
Parabacteroides distasonis JCM 13401.
Parabacteroides merdae.
Parabacteroides sp. D13.
Paraprevotella clara YIT 11840.
Prevotella buccalis.
Prevotella copri DSM 18205.
Roseburia intestinalis L1
Roseburia inulinivorans DSM 16841.
Ruminococcus gnavus ATCC 29149.
C
Relative Abundance
0.00
0.25
0.50
0.75
1.00
Expected
DADA2
Deblur
UNOISE3
OTU
Pipelines
Organism
Non Reference
Cyberlindnera jadinii.
Debaryomyces hansenii.
Galactomyces geotrichum Type 1
Galactomyces geotrichum Type 2
Galactomyces geotrichum Type 3
Galactomyces geotrichum Type 4
Diutina catenulata
Fusarium domesticum.
Hyphopichia burtonii
Kluyveromyces lactis.
Penicillium allii.
Penicillium commune.
Penicillium roqueforti.
Pichia kudriavzevii.
Scopulariopsis fusca.
Zygosaccharomyces rouxii.
D
Relative Abundance
0.00
0.25
0.50
0.75
1.00
Expected
DADA2
Deblur
UNOISE3
OTU
Pipelines
Organism
Non Reference
Pseudomonas aeruginosa
Escherichia coli
Salmonella enterica
Lactobacillus fermentum
Enterococcus faecalis
Staphylococcus aureus
Listeria monocytogenes
Bacillus subtilis

Figure 2.2: **Relative abundances of taxa generated by each sequence processing method for four different mock communities.** *All ASVs/OTUs that matched with expected sequences at 97% or greater identity were assigned taxonomy using a BLASTN search against the expected sequences provided for the Extreme, Human Microbiome Project, and Zymomock mock communities. All ASVs/OTUs that matched an expected species with 97% or greater identity to the UNITE database were classified as expected sequences in the fungal mock community. Non-reference refers to the abundance of ASVs/OTUs that did not match expected sequences with 97% or greater identity. (A) Human Microbiome Project equal abundance mock community; (B) Extreme dataset—it is important to note that some organisms are not displayed in this figure due to their very low abundances; (C) Fungal ITS1 mock community; (D) Zymomock community.*

2.4.4 - Weighted Beta-Diversity Results from Different Approaches are Indistinguishable in Real Soil and Host-Associated Communities

After comparing each pipeline using mock communities, we next wanted to investigate how comparable the results between pipelines were for real 16S datasets. We compared the pipelines on a soil dataset (soil) due to its high diversity (Fierer & Jackson, 2006) a mouse exercise stool dataset (exercise) and a gut biopsy sample dataset from Crohn's Disease patients plus controls (human-associated). The intra-sample distances based on weighted UniFrac measurements were comparable between each approach among all three datasets (Figure 2.3). In general, we found the intra-samples distances to be small, ranging between median values of ~0.09 and ~0.15. There were no consistent differences seen between all three datasets. Looking at the soil dataset, we found that all three denoising pipelines had similarly small intra-sample distances (~0.11) (Figure 2.3A), while the open-reference OTU clustering pipeline was slightly further away from UNOISE3 and Deblur (~0.13). Despite this observation, it is important to note that this difference is still relatively small. Furthermore, when each sample was plotted onto a PCoA we found that samples tended to group by biological origin rather than the approach used to process the sequences, suggesting that a relatively similar PCoA plot would be generated by each pipeline (Figure 2.3D-F). The Mantel correlations between each weighted UniFrac distance matrix were all highly correlated (correlation values ranging from 0.764 to 0.975), which suggests similar weighted UniFrac profiles between each approach (Supplemental Tables 6-8).

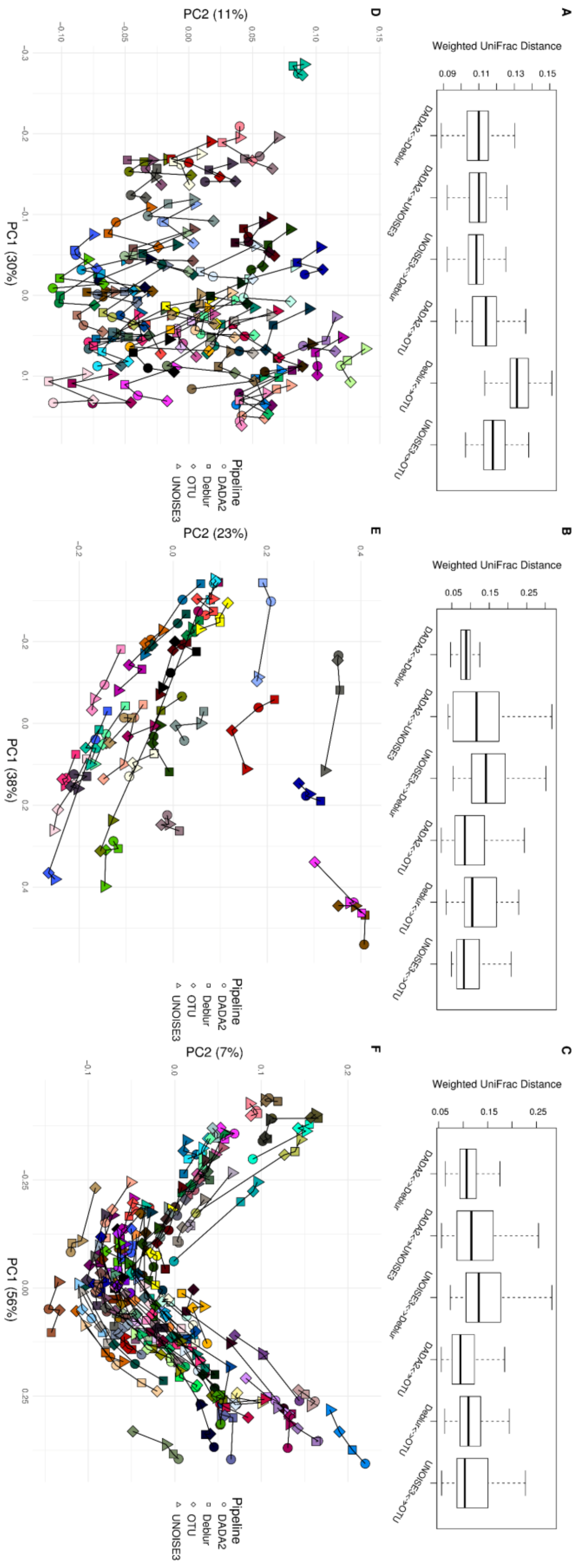

A
B
C
Weighted UniFrac Distance
0.09
0.11
0.13
0.15
0.05
0.15
0.25
DADA2<->Deblur
DADA2<->UNOISE3
UNOISE3<->Deblur
DADA2<->OTU
Deblur<->OTU
UNOISE3<->OTU
D
E
F
PC2 (11%)
PC1 (30%)
PC2 (23%)
PC1 (38%)
PC2 (7%)
PC1 (56%)
-0.10
-0.05
0.00
0.05
0.10
0.15
-0.3
-0.2
-0.1
0.0
0.1
-0.2
0.0
0.2
0.4
-0.1
0.0
0.1
0.2
-0.25
0.00
0.25
Pipeline
○ DADA2
□ Deblur
◇ OTU
△ UNOISE3

Figure 2.3: ***Weighted UniFrac intra-sample distances between sequence processing methods based on three real datasets****. (A–C) The weighted UniFrac distances between the same biological samples based on ASVs/OTUs outputted by each of the different sequence processing methods on the soil, human associated and exercise datasets, respectively. (D–F) Principal coordinates analysis of the weighted UniFrac distances of all the samples in the soil, human associated, and Exercise datasets, respectively. The four different sample profiles generated for each biological sample are colour-coded and are joined by an interconnecting line.*

Looking at Bray-Curtis dissimilarity, another metric that takes in abundance information, we found that intra-sample distances based on genus-level assignment were similar among different approaches (Supplemental Figure 5). In general, the intra-sample distances were relatively low, with median values ranging from ~0.04 to ~0.20, indicating a high amount of agreement between approaches. Interestingly, we found that the intra-sample distances seemed to increase across datasets based on how diverse the samples within them were (Supplemental Figure 5). In two of the three datasets (soil and BISCUIT), we found that DADA2 and UNOISE3 tended to be closer together in distance than Deblur was to DADA2 or UNOISE3, indicating a slightly higher amount of agreement between DADA2 and UNOISE3 at the genus level. Again, these differences were relatively small and would have minimal impacts on biological results obtained from them. Plotting the Bray-Curtis dissimilarity matrices onto an NMDS plot resulted in similar findings as weighted UniFrac, with samples grouping by biological origin rather than the pipeline used to process them. Furthermore, we found that the Bray-Curtis dissimilarity matrices were extremely well correlated with each other, ranging in values between 0.956-0.995 (Supplemental Tables 6-8).

We next compared unweighted UniFrac distances, which is a metric that considers the presence or absence of ASVs/OTUs and their phylogenetic distance. We found that the median intra-sample distance between each pipeline was much greater, ranging between ~0.40 to ~0.79 (Supplemental Figure 6). Similar to our results focused on Bray-Curtis dissimilarity, we also found that the median intra-sample distance between samples tended to increase with sample diversity (0.72-0.79 for soil; 0.40-0.55 for human-associated; 0.60-0.70 for exercise. The PCoA plot based on these distances resulted in samples grouping together based on pipeline, rather than biological origin, indicating large differences between the different approaches (Supplemental Figure 6).

To look at how much of a difference the positive filtering process had on the profiles generated by Deblur and to see if the slight difference in Bray-Curtis dissimilarity among the denoising approaches was caused by this filter, we ran the stand-alone version of Deblur and examined the non-positive filtered results. We found that the intra-sample distances based on any of the three metrics tested (weighted/unweighted UniFrac and Bray-Curtis dissimilarity) between any of the pipelines did not vary (Supplemental Figure 7).

2.4.5 - Alpha-Diversity Metrics Vary Between Denoising Pipelines

We next investigated how the number of ASVs/OTUs called in a real dataset differed between processing pipelines. In the soil dataset, DADA2 called 16609 ASVs, UNOISE3 called 11613 ASVs, Deblur called 8270 ASVs and open-reference OTU clustering called 21297 OTUs (Figure 2.4A). Across all datasets, we found that DADA2 called the most

ASVs among the denoising pipelines and that open-reference OTU clustering called the most ASVs/OTUs overall (Figure 2.4A, 2.4C, 2.4E). On average, DADA2 called 727 more ASVs than Deblur and 532 more ASVs than UNOISE3 while open-reference OTU clustering called 3135 more OTUs/ASVs than DADA2. In all datasets, Deblur called the least amount of ASVs/OTUs. Looking at the number of OTUs called per sample, we found that DADA2 did not correlate well with UNOISE3 or open-reference OTU clustering and in the soil dataset the correlation between DADA2 and UNOISE3 was not found to be significant (*p*=0.054) (Figure 2.4B, 2.4D, 2.4F). This is a concerning result as it indicated the possibility of different biological results based on the pipeline that was chosen to process the data.

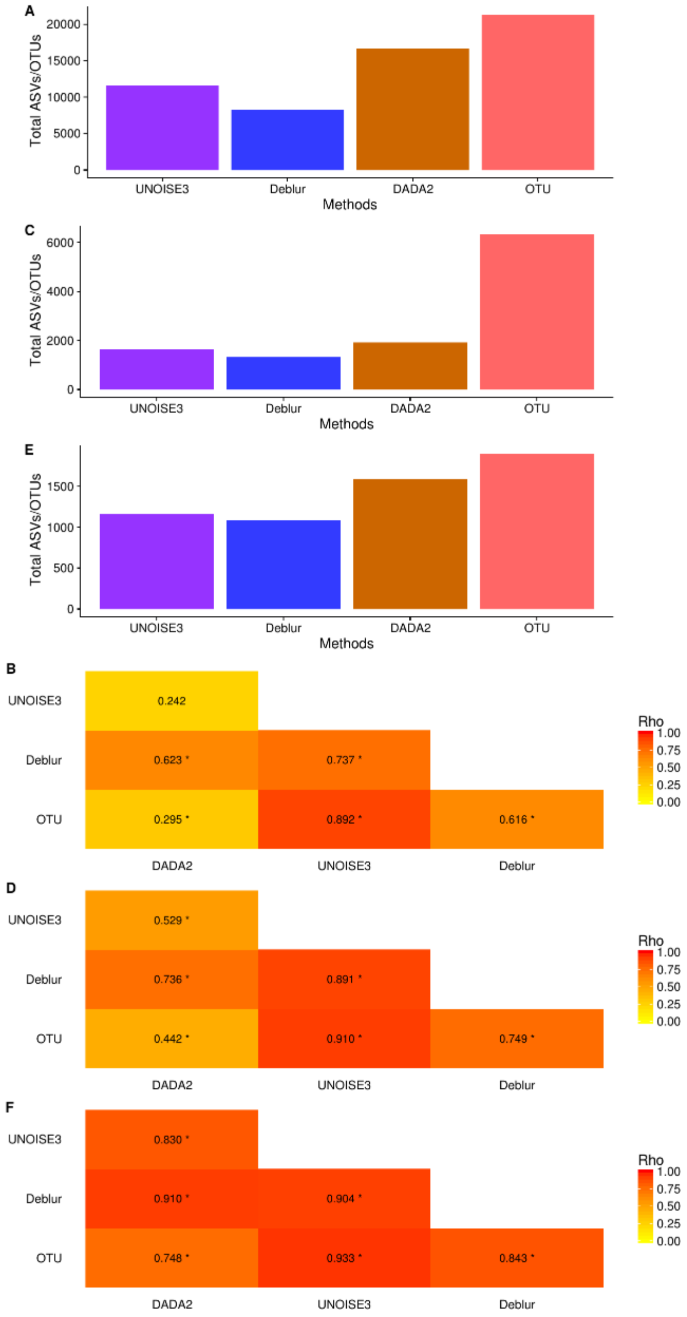


Figure 2.4: **Total number of ASVs/OTUs called by each processing method and the per-sample observed ASV/OTUs correlation between each sequence processing method.** *(A, C, E) The total numbers of ASVs/OTUs determined by each method on the soil, exercise, and human associated datasets, respectively. (B, D, F) Heatmaps of the Spearman correlations between the numbers of observed ASVs/OTUs per sample between*

different sequence processing methods. Significant p-values ($p < 0.05$) are indicated by *.

One interesting result was that, despite DADA2 calling the most total ASVs among the denoising pipelines, it called the least amount per sample (Supplemental Figure 8-10) and, among denoising pipelines, UNOISE3 tended to call the most (Supplemental Fig 8-10). However, open-reference OTU clustering called the most OTUs/ASVs overall per sample among all approaches that were tested with the exception of the soil dataset.

2.4.6 - Computational Requirements are Vastly Different Across Denoising Pipelines

Knowing that all three of these pipelines resulted in similar relative abundance profiles on mock communities and small intra-sample distances on real 16S communities, we next investigated how the run time and memory usage differed between the denoising approaches. We found that UNOISE3 (4.6 minutes) was 1273 times faster than DADA2 (5834 minutes) and 15 times faster than Deblur (69.3 minutes) at a total read count of 1,926,000 reads evenly distributed across 103 samples (Figure 2.5A). Run times for all pipelines increased as the number of reads per sample increased. Deblur used a static amount of memory (611 Mb) as reads per sample increased, whereas in general the other two pipelines increased in memory usage as the number of reads per sample increased, with the exception of DADA2 run at 1,926,000 reads (Figure 2.5B). Deblur used the smallest amount of memory at the maximum read count of 1,926,000 reads. We found that DADA2 had the highest amount of memory usage (4071 Mb at 1,287,000 reads) among the three pipelines. Interestingly, this usage was more than the amount used at the maximum read count (3600 Mb). In addition, none of the runs exceeded the 4 Gb memory cap on the 32-bit free academic version of USEARCH10.

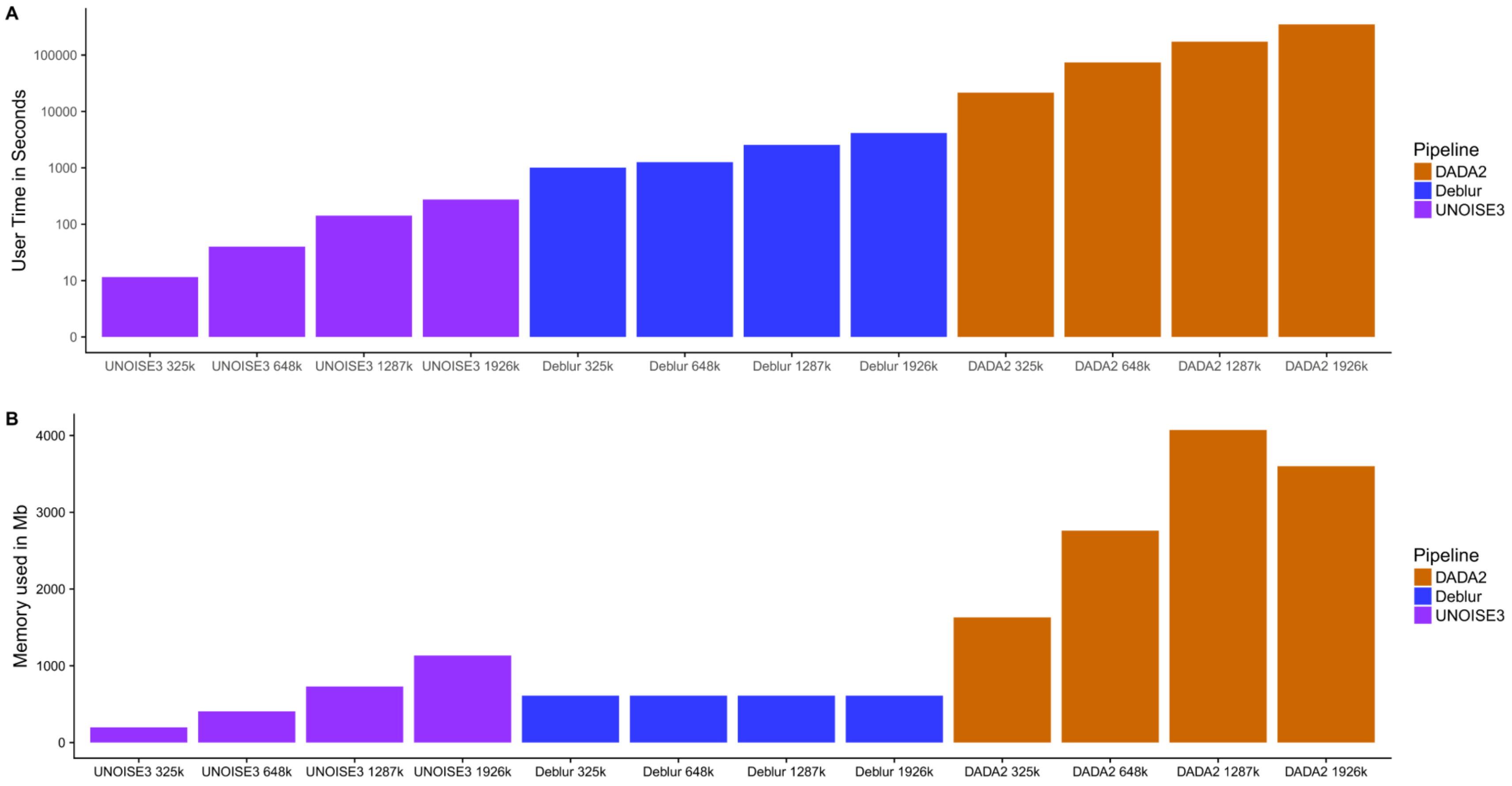

A
User Time in Seconds
0
10
100
1000
10000
100000
UNOISE3 325k
UNOISE3 648k
UNOISE3 1287k
UNOISE3 1926k
Deblur 325k
Deblur 648k
Deblur 1287k
Deblur 1926k
DADA2 325k
DADA2 648k
DADA2 1287k
DADA2 1926k
Pipeline
DADA2
Deblur
UNOISE3
B
Memory used in Mb
0
1000
2000
3000
4000
UNOISE3 325k
UNOISE3 648k
UNOISE3 1287k
UNOISE3 1926k
Deblur 325k
Deblur 648k
Deblur 1287k
Deblur 1926k
DADA2 325k
DADA2 648k
DADA2 1287k
DADA2 1926k
Pipeline
DADA2
Deblur
UNOISE3

Figure 2.5: ***Run time and memory usage of each denoising pipeline on a dataset of varying size.*** *The time in seconds (A) and memory in megabytes (B) to run varying amounts of reads through the three different denoising methods. Note time is on a log_10 scale.*

2.5 – Discussion

Despite the differences between the underlying algorithms, all four approaches were comparable based on weighted UniFrac and Bray-Curtis distances. However, the tools varied greatly when looking at unweighted or raw number of ASVs/OTUs found. It should be noted that all denoising pipelines were run using their recommended settings, limiting our comparison to only looking at the default settings for each pipeline. Adjusting of parameters within each sequence analysis pipeline would allow users to tailor towards more sensitivity and specificity as desired. Furthermore, each package uses individual chimera filtering methods which may affect the overall accuracy of the entire pipeline.

During mock community data processing, no denoising approach consistently called more ASVs than another, however open-reference OTU clustering found more OTUs than any of the three denoising approaches before abundance filtering (Figure 2.1). In many cases, the number of OTUs found was vastly greater than the number of expected sequences, which is consistent with previous literature reporting that OTU clustering tends to exaggerate the number of unique organisms found within a sample (R. C. Edgar, 2017). No pipeline was able to call all expected sequences for any of the mock communities at 100% identity (Figure 2.1), which indicates that some may not have been present or that preparation steps show large bias towards specific 16S rRNA gene variants within an organism. Each sequence processing pipeline was able to detect every organism in the HMP community (note *S. aureus* and *S. epidermidis* are collapsed together as they have the same sequenced region) and the Zymomock community. However, one odd result was that observed abundances within the HMP community were comparable to the expected abundances within the HMP community, but not the

Zymomock community (Figure 2.2A, 2.2D)**.** This again most likely indicates bias during library preparation but could also be due to similar bias among pipelines. In the Extreme dataset, all denoising pipelines missed *P. buccalis*, *C. methylpentusum* and *P. sp._D13* (Figure 2.2B, Supplemental Table 4). All three of these organisms had very low expected abundances (less than 0.00427%) which may explain why they were difficult to detect (Supplemental Table 4). We do not believe it was an issue with sequence depth as the single sample in the mock community was sequenced at a depth of 11.6 million reads and open-reference OTU clustering only missed one organism (*C. methylpentusum*). Deblur and UNOISE3 both failed to detect 9 of the 27 expected taxa in the Extreme dataset at 97% identity, which were all detected by DADA2 and open-reference OTU clustering. Again, these nine organisms were at very low abundances (less than 0.05%) (Supplemental Table 4). As mentioned above, open-reference OTU clustering did find 2 more expected taxa than DADA2, however, this came at the cost of significantly increased numbers of non-reference sequences being detected. Differences in detection between DADA2 and the other two denoising pipelines suggests that it is better at detecting organisms that are very rare without the cost of finding significantly more non-reference hits that plagues the open-reference OTU clustering pipeline. Whether finding rare organisms is truly advantageous is debatable, as many of these low-abundance organisms would be removed by typical filtering cut-offs and/or contribute little to weighted beta-diversity metrics such as UniFrac or Bray-Curtis dissimilarity.

In all mock datasets, a large number of Unmatched OTUs were found by open-reference OTU clustering and, in some of the datasets, different denoising pipelines also found a relatively large amount of Unmatched ASVs (Figure 2.1). This could have been due to errors in sequencing or mutations during PCR amplification of the 16S rRNA gene as most sequences did show over 90% identity to the expected sequences, indicating that they originated from a 16S rRNA gene (Supplemental Figure 1-3). It is also possible that many of these sequences are contaminants from the environment, while others may have been from chimeras that were not detected by each pipeline's individual chimera-checking algorithms. We also found a high proportion of matches to the reference database but not the expected sequences. There are several reasons why this could have occurred. Some of these sequences may be contaminants introduced during mock community preparation or sequencing as they match with 100% identity to a sequence within the Greengenes database indicating that they are a real 16S rRNA gene sequence. However, due to different pipelines identifying different numbers of these unexpected sequences, it could also suggest that some of these contaminants are likely either chimeric sequences that can be found within the Greengenes database, or sequencing errors/mutations during PCR amplification that could not be corrected for by the respective pipelines. Unsurprisingly, the highest sequence depth mock community (Extreme) had the most Unmatched OTUs (Figure 2.1), most likely due to the increased absolute abundance of sequencing errors and the increased likelihood of finding minor contaminants. Furthermore, DADA2 found a significantly greater number of sequences under or at 75% sequence identity, when compared to the other denoising pipelines, highlighting the trade-off that it makes to find lower-abundance ASVs (Supplemental Figure 2).

In all cases, the open-reference OTU clustering had greatly increased numbers of OTUs found when compared to ASVs found by denoising pipelines (Figure 2.1). To try and address this issue, we applied an abundance filter of 0.1% minimum abundance (except for the Extreme community 0.0004% minimum abundance) to all four different pipelines over all the datasets (Supplemental Figure 4). The largest effect that this filter had was the reduction of the number of OTUs called by the open-reference OTU clustering. These significant changes in the number of OTUs by open-reference OTU clustering highlight the importance of removing low-abundance OTUs during analysis and demonstrate the stability of denoising approaches after abundance filtering. The filter cutoff did have an impact on the number of Unmatched ASVs called by DADA2 in the HMP mock community (Supplemental Figure 4A) and the Unmatched ASVs called by Deblur in the Zymomock community (Supplemental Figure 4E) but had little effect on the number of ASVs called by UNOISE3 on these communities, indicating UNOISE3's tendency to call ASVs of higher abundance. No differences were seen in the Extreme community on ASV output by denoising approaches, most likely due the relatively low cutoff of 0.0004%. Overall, results on the mock communities showed that Deblur tended to call the least amount of ASVs/OTUs among all pipelines and open-reference OTU clustering called the most.

The relative abundances determined for each study on the mock communities were similar to each other irrespective of which pipeline processed the data (Figure 2.2). This finding suggests that biological conclusions based on microbial relative abundance data should be unaffected by the choice of denoising pipeline. One trend that was noticed in the relative abundance data was that UNOISE3 tended to call higher abundances of non-reference ASVs/OTUs in each mock community except for the Extreme community where open-reference OTU clustering found the highest abundance of non-reference hits.

Interestingly, the lowest identity match for any of these ASVs called by UNOISE3 in both the Zymomock and HMP mock communities was still found to be above 90% identity to the expected sequences (Supplemental Figure 1,3) and was classified as Gammaproteobacteria by the RDP classifier using a 70% confidence threshold, suggesting it is a 16S rRNA sequence that may have been introduced by contamination, sequencing bleed-through or acquired an error early on in PCR amplification.

The relative abundances determined within the Zymomock and fungal communities were highly similar between pipelines, but markedly differed from the expected result. This finding suggests that either the expected abundances of sequences from these communities may be incorrect or all four pipelines are similarly biased. This non-agreement could also be due to steps during the sequencing processes such as PCR amplification, which may be causing primer bias (Aird et al., 2011) or the inclusion of contaminant organisms. In the case of the fungal community, it is possible that none of these approaches work well with ITS1 data which are more variable than 16S data. Additional fungal mock communities should be analyzed in the future to better explore this issue.

Benchmarking relative abundance profiles from different pipelines with mock communities can be useful, however, they tend to lack the diversity that is found in many real sample datasets. To address this issue, we compared resulting microbial compositions from each pipeline across three real datasets (exercise, human-associated, and soil). Weighted UniFrac, unweighted UniFrac and Bray-Curtis dissimilarity distances between the same biological samples for each approach were examined. In both cases the weighted UniFrac and Bray-Curtis intra-sample distances between all pipelines for all three datasets were small (less than a median of 0.21) (Figure 2.3, Supplemental Figure 5). This complemented our previous results, showing that each pipeline had comparable microbial compositions for the mock communities. Furthermore, plotting the samples on a PCoA or NMDS resulted in the same biological samples from each pipeline grouping together (Figure 2.3, Supplemental Figure 5). This indicated that a similar plot would be observed whether the researcher was using the Deblur, UNOISE3, DADA2 or open-reference OTU clustering. Interestingly, Deblur did not agree with the DADA2 or UNOISE3 as much as they agreed with each other on multiple occasions, based on Bray Curtis dissimilarity at the genus level (Supplemental Figure 5A,5B). This result is interesting, as one of the main differences between Deblur and the other two denoising pipelines is its positive filtering feature, and so we expected this feature to be driving this difference. However, when we compared the other three approaches to Deblur and Deblur without positive filtering, we found no difference (Supplemental Figure 7). Due to the similar weighted UniFrac results (Figure 2.3) between the denoising pipelines, we believe that this difference is most likely due to highly similar sequences being classified into slightly different genera.

Comparing the three denoising pipelines to the open-reference OTU picking pipeline, we found that Deblur agreed the least with the OTU pipeline of all the denoising pipelines (Supplemental Figure 5). In most cases, OTUs had similar intra-sample distances between pipelines in both weighted UniFrac and Bray Curtis dissimilarity as denoising pipelines had amongst each other. The relatively small differences seen in the abundance-based metrics indicates that choice of pipeline would have minimal impact when looking at weighted beta-diversity metrics. Although, due to denoising pipelines capability of single nucleotide resolution, they may provide more strain information than OTU clustering at 97% would.

Unweighted UniFrac beta-diversity metrics were highly variable between pipelines indicating different bias between pipelines when determining low abundance sequence (Supplemental Figure 6A-C). When pipelines were plotted together on a PCoA plot we found that the samples separated by approach, rather than sample, indicating that interpretation from unweighted UniFrac data would most likely be impacted based on the pipeline chosen to process any set of given data. We wish to highlight, however, that recent evidence has emerged that the unweighted version of UniFrac analysis can give misleading results (Wong, Wu, & Gloor, 2016), therefore these patterns should be interpreted with caution.

To follow up on the vastly different unweighted UniFrac profiles of samples, we looked at the alpha-diversity between the same samples run by each pipeline and the total number of ASVs/OTUs called by each pipeline in the real datasets. We found that DADA2 called the most ASVs among all denoising approaches, but overall open-reference OTU clustering found the most (Figure 2.4). This was not surprising, as it agreed with the analysis of the mock communities. Interestingly, despite DADA2 finding the most ASVs overall for all three denoising pipelines, it found the least amount of ASVs per sample (Supplemental Figure 8-10). We suspect this is due to DADA2's ability to create pooled error profiles and then pick ASVs sample by sample. Overall, open-reference OTU clustering found more OTUs per sample than ASVs found by any denoising pipelines. This indicated that the number of different organisms found within a sample is directly impacted by the choice of processing pipeline, emphasizing the difficulty in determining the true number of different organisms a sample contains. It should be noted that all of the denoising pipelines provide parameters that could be altered to increase or decrease sensitivity in identifying rare/spurious ASVs depending on a user's targeted application. One alarming result we found was the poor correlation between the number of observed OTUs/ASVs found by DADA2 and both open-reference OTU clustering and UNOISE3 (Figure 2.4). This is concerning as major differences in biological signal would be seen depending on the approach that was chosen to process the data (i.e. a sample could have relatively low numbers of ASVs based on DADA2 analysis, but have relatively high numbers of ASVs/OTUs based on UNOISE3 or an OTU analysis). This issue highlights that the approaches are all very good at identifying highly abundant sequences but vary when identifying low-abundance sequences which will impact metrics that do not take into account the abundance of OTUs/ASVs.

A major difference between the three denoising pipelines was their computational run time. UNOISE3 was magnitudes faster than both DADA2 and Deblur. This is most likely due to both the programming language that UNOISE3 is implemented in (C++), as well as its simple one-pass denoising pipeline. DADA2 was the slowest pipeline and, although computation time could be inconvenient for those with limited computational power, it did not reach times that were impractical even when running almost 2 million total reads. Memory usage for each program also did not reach impractical amounts when running close to 2 million reads, with DADA2 using a maximum amount of 1024 Mb of memory which is a reasonable amount for modern computers. Memory usage by UNOISE3 did not come close to reaching the 4 Gb memory cap on the 32-bit version, even after running 103 samples at a total read depth of 2 million reads. This suggests that for most moderately sized 16S datasets the 32bit version of UNOISE3 should be sufficient.

2.6.1 – Conclusion

In conclusion, all four pipelines are comparable when looking at weighted results that are based on the relative abundances of ASVs/OTUs while using default settings. However, the approaches do vary when looking at the number of ASVs/OTUs found and unweighted metrics such as unweighted UniFrac, while using the default settings for each denoising pipeline. The number of ASVs called did not differ between denoising pipelines in a consistent way across mock communities, suggesting that determining species richness within low-diverse samples could be problematic. However, we did find that open-reference OTU clustering consistently called more OTUs than ASV-calling pipelines. Analysis of the real datasets showed that DADA2 consistently called more ASVs than the other two denoising pipelines and that, again, open-reference OTU clustering called the most overall. More importantly, in the soil dataset and in the Extreme dataset DADA2 and the open-reference OTU clustering pipeline were capable of finding more low-abundance organisms, but DADA2 could do this without the cost of significantly increasing the number of non-reference hits. The most alarming result was the poor correlation in the number of ASVs/OTUs per sample found between DADA2 and UNOISE3 or open-reference OTU clustering. From this, we believe that choice of approach will have large impacts on determining the alpha-diversity between different samples. Looking at computational run time, we found that DADA2 was by far the slowest denoising pipeline, whereas UNOISE3 was the fastest, processing datasets more than 1200 times faster than DADA2. In the end, the choice of approach did not play a large role in weighted analyses based on microbial abundances, but did have implications on unweighted results and alpha-diversity metrics. We believe this is promising, as it indicates that no matter the choice of approach, a similar weighted biological signal will be seen. On the other hand, extreme caution is required when looking at unweighted results and alpha-diversity metrics between different pipelines.

Chapter 3 – Microbiome Differential Abundance Methods Produce Different Results Across 38 Datasets

This chapter is a reproduction of the paper of the same name published in the journal Nature Communications (Nearing et al., 2022). I was co-first author on this work with Dr. Gavin M. Douglas, a former PhD student supervised by Dr. Morgan Langille. My contributions to this project including the initial project proposal, data collection along with other co-authors, data processing, and the code to run each dataset through each differential abundance tool. Gavin, participated in the construction of preliminary code to run a subset of tools. Data analysis was conducted in a joint fashion with Gavin taking lead in section 3.4.4 and myself taking lead in the remaining sections. All results sections received critical feedback and code from both authors. All sections of the paper were jointly written by both me and Gavin while receiving feedback from all other co-authors.

Additional co-authors on this manuscript include Molly G. Hayes, Jocelyn MacDonald, Dhwani K. Desai, Nicole Allward, Casey M.A. Jones, Robyn J. Wright, Akhilesh S. Dhanani, André M. Comeau, and Morgan G.I. Langille.

This paper was published under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution, and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. To view a copy of this license visit: http://creativecommons.org/licenses/by/4.0/

The original author funding statements, author contributions breakdown, acknowledgements and data availability statements are also available in the original publication. All code used to generate results for this chapter can be found at: https://github.com/nearinj/Comparison_of_DA_microbiome_methods. All supplemental and additional files referred to in this chapter are freely available as part of the publication on the Nature Communications website and can be found at this link: https://doi.org/10.1038/s41467-022-28034-z.

3.1 – Abstract

Identifying differentially abundant microbes is a common goal of microbiome studies. Multiple methods are used interchangeably for this purpose in the literature. Yet, there are few large-scale studies systematically exploring the appropriateness of using these tools interchangeably, and the scale and significance of the differences between them. Here, we compare the performance of 14 differential abundance testing methods on 38 16S rRNA gene datasets with two sample groups. We test for differences in amplicon sequence variants (ASVs) and operational taxonomic units (OTUs) between these groups. Our findings confirm that these tools identified drastically different numbers and sets of significant ASVs, and that results depend on data pre-processing. For many tools the number of features identified correlate with aspects of the data, such as sample size, sequencing depth, and effect size of community differences. ALDEx2 and ANCOM-II produce the most consistent results across studies and agree best with the intersect of results from different approaches. Nevertheless, we recommend that researchers should use a consensus approach based on multiple differential abundance methods to help ensure robust biological interpretations.

3.2 – Introduction

Microbial communities are frequently characterized by DNA sequencing. Marker gene sequencing, such as 16S rRNA gene sequencing, is the most common form of microbiome profiling and enables the relative abundances of taxa to be compared across different samples. A frequent and seemingly simple question to investigate with this type of data is: which taxa significantly differ in relative abundance between sample groupings? Newcomers to the microbiome field may be surprised to learn that there is little consensus on how best to approach this question. Indeed, there are numerous ongoing debates regarding the best practices for differential abundance (DA) testing with microbiome data (Allaband et al., 2019; Pollock et al., 2018).

One area of disagreement is whether read count tables should be rarefied (i.e., subsampled) to correct for differing read depths across samples (S. Weiss et al., 2017). This approach has been heavily criticized because excluding data could reduce statistical power and introduce biases. In particular, using rarefied count tables for standard tests, such as the t-test and Wilcoxon test, can result in unacceptably high false positive rates (McMurdie & Holmes, 2014). Nonetheless, microbiome data is still frequently rarefied because it can simplify analyses, particularly for methods that do not control for variation in read depth across samples. For example, LEfSe (Segata et al., 2011) is a popular method for identifying differentially abundant taxa that first converts read counts to percentages. Accordingly, read count tables are often rarefied before being input into this tool so that variation in sample read depth does not bias analyses. Without addressing the variation in depth across samples by some approach, the richness can drastically differ between samples due to read depth alone.

A related question to whether data should be rarefied is whether rare taxa should be filtered out. This question arises in many high-throughput datasets, where the burden of correcting for many tests can greatly reduce statistical power. Filtering out potentially uninformative features before running statistical tests can help address this problem, although in some cases this can also have unexpected effects such as increases in false positives (Bourgon, Gentleman, & Huber, 2010). Importantly, this filtering must be independent of the test statistic evaluated (referred to as Independent Filtering). For instance, hard cut-offs for the prevalence and abundance of taxa across samples, and not within one group compared with another, are commonly used to exclude rare taxa (Schloss, 2020). This data filtering could be especially important for microbiome datasets because they are often extremely sparse. Nonetheless, it remains unclear whether filtering rare taxa has much effect on DA results in practice.

Another contentious area is regarding which statistical distributions are most appropriate for analyzing microbiome data. Statistical frameworks based on a range of distributions have been developed for modelling read count data. For example, DESeq2 (Love et al., 2014) and edgeR (Robinson & Oshlack, 2010) are both tools that assume read counts follow a negative binomial distribution. To identify differentially abundant taxa, a null and alternative hypothesis are compared for each taxon. The null hypothesis states that the same setting for certain parameters of the negative binomial solution explain the distribution of taxa across all sample groupings. The alternative hypothesis states that different parameter settings are needed to account for differences between sample groupings. If the null hypothesis can be rejected for a specific taxon, then it is considered differentially abundant. This idea is the foundation of distribution-based DA tests, including other methods such as corncob (B. D. Martin et al., 2020) and metagenomeSeq (Paulson et al., 2013), which model microbiome data with the beta-binomial and zero-inflated Gaussian distributions, respectively.

Finally, it has recently become more widely appreciated that DNA sequencing data is compositional in nature (Gloor et al., 2017). Highlighting the fact that read counts only provide relative abundance information rather than absolute quantities. This is because a sequencer can only read a maximum number of reads during any single sequencing run. This results in read counts being proportional to an arbitrary maximum read count for each sample that is sequenced. A characteristic that can led to false inferences when standard statistical methods, intended for absolute abundances, are used with taxonomic relative abundances. Compositional data analysis (CoDa) methods circumvent this issue by reframing the focus of analysis to ratios of read counts between different taxa within a sample (Aitchison, 1982; Morton et al., 2019).The difference among CoDa methods considered in this thesis is what abundance value is used as the denominator, or the reference, for the transformation. The centered log-ratio (CLR) transformation is a CoDa approach that uses the geometric mean of the read counts of all taxa within a sample as the reference/denominator for that sample. In this approach all taxon read counts within a sample are divided by this geometric mean and the log fold changes in this ratio between samples are compared. An extension of this approach is implemented in the tool ALDEx2 (Fernandes et al., 2014a) . The additive log-ratio transformation is an alternative approach where the reference is the count abundance of a single taxon, which should be present with low variance in read counts across samples. In this case the ratio between the reference taxon chosen (denominator) and each taxon in that sample are compared across different sample groupings. An additive log ratio strategy is implemented by the tool ANCOM by testing all possible taxonomic denominators for each taxa of interest (Mandal et al., 2015).

Regardless of the above choices, evaluating the numerous options for analyzing microbiome data has proven difficult. This is largely because there are no gold standards to compare DA tool results. Simulating datasets with specific taxa that are differentially abundant is a partial solution to this problem, but it is imperfect. For example, it has been noted that parametric simulations can result in circular arguments for specific tools, making it difficult to assess their true performance (Hawinkel, Mattiello, Bijnens, & Thas, 2019). It is unsurprising that distribution-based methods perform best when applied to simulated data based on that distribution. Nonetheless, simulated data with no expected differences has been valuable for evaluating the false discovery rate (FDR) of these methods. Based on this approach it has become clear that many of the methods produce unacceptably high numbers of false positive identifications (Calgaro et al., 2020; Thorsen et al., 2016; S. Weiss et al., 2017). Similarly, based on simulated datasets with spiked taxa it has been shown that these methods can drastically vary in statistical power (Hawinkel et al., 2019; Thorsen et al., 2016).

Although these general observations have been well substantiated, there is less agreement regarding the performance of tools across evaluation studies. Certain observations have been reproducible, such as the higher FDR of edgeR and metagenomeSeq. Similarly, ALDEx2 has been repeatedly shown to have low power to detect differences (Calgaro et al., 2020; Hawinkel et al., 2019). In contrast, both ANCOM and limma voom (Law et al., 2014; Ritchie et al., 2015) have been implicated as both accurately and poorly controlling the FDR, depending on the study (Calgaro et al., 2020; Hawinkel et al., 2019; S. Weiss et al., 2017). To further complicate comparisons, different sets of tools and dataset types have been analyzed across evaluation studies. This means that, on some occasions, the best performing method in one evaluation is missing from another. In addition, certain popular microbiome-specific methods, such as MaAsLin2 (Mallick et al., 2021), have been missing from past evaluations. Finally, many evaluations limit their analysis to a small number of datasets that do not represent the breadth of datasets found in 16S rRNA gene sequencing studies.

Given the inconsistencies across these studies it is important that additional, independent evaluations be performed to elucidate the performance of current DA methods. This is particularly important as these tools are typically used interchangeably in microbiome research. Accordingly, herein we have conducted additional evaluations of common DA tools across 38 two-group 16S rRNA gene datasets. We first present the concordance of the methods on these datasets to investigate how consistently the methods cluster and perform in general, with and without the removal of rare taxa. Next, based on artificially subsampling the datasets into two groups where no differences are expected, we present the observed FDR for each DA tool. Lastly, we present an evaluation of how consistent biological interpretations would be across diarrheal and obesity datasets depending on which tool was applied. Our work enables improved assessment of these DA tools and highlights which key recommendations made by previous studies hold in an independent evaluation. Furthermore, our analysis shows various characteristics of DA tools that authors can use to evaluate published literature within the field.

3.3 – Methods

3.3.1 - Dataset Processing

Thirty-eight different datasets were included in our main analyses for assessing the characteristics of microbiome differential abundance tools. Three additional datasets were also included for a comparison of differential abundance consistency across diarrhea and obesity-related microbiome datasets. All datasets presented herein have been previously published or are publicly available (Alkanani et al., 2015; Baxter, Ruffin, Rogers, & Schloss, 2016; Chase et al., 2016; De Tender et al., 2015; Dinh et al., 2015; Douglas et al., 2018; Dranse et al., 2018; Duvallet, Gibbons, Gurry, Irizarry, & Alm, 2017; Frère et al., 2018; Gonzalez et al., 2018; Goodrich, Waters, et al., 2014; Hoellein et al., 2017; Ji, Parks, Edwards, & Pruden, 2015; Kesy, Oberbeckmann, Kreikemeyer, & Labrenz, 2019; Lamoureux et al., 2017; C. A. Lozupone et al., 2013; McCormick et al., 2016; Mejía-León, Petrosino, Ajami, Domínguez-Bello, & de la Barca, 2014; Nearing et al., 2019; Noguera-Julian et al., 2016; Oberbeckmann, Osborn, & Duhaime, 2016; Oliveira et al., 2018; Papa et al., 2012; Pop et al., 2014; Rosato et al., 2020; Ross et al., 2015; Scheperjans et al., 2015; Scher et al., 2013; Schneider et al., 2017; Schubert et al., 2014; Singh et al., 2015; Son et al., 2015; Turnbaugh et al., 2009; Vincent et al., 2013; L. Wu et al., 2019; Yurgel et al., 2017; Zeller et al., 2014; L. Zhu et al., 2013) (Supplemental Data 1). Most datasets were already available in table format with ASV or operational taxonomic unit abundances while a minority needed to be processed from raw sequences. These raw sequences were processed with QIIME 2 version 2019.7 (Bolyen et al., 2019) based on the Microbiome Helper standard operating procedure (Comeau et al., 2017). Primers were removed using cutadapt (M. Martin, 2011) and stitched together using the

QIIME 2 VSEARCH (Rognes et al., 2016) join-pairs plugin. Stitched reads were then quality filtered using the quality-filter plugin and reads were denoised using Deblur (Amir et al., 2017) to produce amplicon sequence variants (ASVs). Abundance tables of ASVs for each sample were then output into tab-delimited files. Rarefied tables were also produced for each dataset, where the rarefied read depth was taken to be the lowest read depth of any sample in the dataset over 2000 reads (with samples below this threshold discarded).

Chimeric ASVs were identified with the UCHIME2 and UCHIME3 chimera-checking algorithms (R. Edgar, 2016) implemented in VSEARCH (v2.17.1) (Rognes et al., 2016). Both the UCHIME2 and UCHIME3 de novo approaches were applied in addition to the UCHIME2 reference-based chimera checking approach. For this latter approach we used the SILVA v138.1 short-subunit reference database (Quast et al., 2013). We used the default options when running these algorithms.

3.3.2 - Differential Abundance Testing

We created a custom shell script (run_all_tools.sh) that ran each differential abundance tool on each dataset within this study. As input the script took a tab-delimited ASV abundance table, a rarefied version of that same table, and a metadata file that contained a column that split the samples into two groups for testing. This script also accepted a prevalence cut-off filter to remove ASVs below a minimum cut-off, which was set to 10% (i.e., ASVs found in fewer than 10% of samples were removed) for the filtered data analyses we present. Note that in a minority of cases a genus abundance table was input instead, in which case all options were kept the same. When the prevalence filter option was set, the script also generated new filtered rarefied tables based on an input rarefaction depth.

Following these steps, each individual differential abundance method was run on the input data using either the rarefied or non-rarefied table, depending on which is recommended for that tool. Rarefaction was performed using GUniFrac (version 1.1) (J. Chen et al., 2012).The workflow used to run each differential abundance tool (with run_all_tools.sh) is described below. The first step in each of these workflows was to read the dataset tables into R (version 3.6.3) with a custom script and then ensure that samples within the metadata and feature abundance tables were in the same order. An alpha-value of 0.05 was chosen as our significance cutoff and FDR adjusted p-values (using Benjamini-Hochberg adjustment) were used for methods that output p-values (with the exception of LEfSe which does not output all p-values by default) (Benjamini & Hochberg, 1995).

3.3.2.1 - ALDEx2

We passed the non-rarefied feature table and the corresponding sample metadata to the *aldex* function from the ALDEx2 R package (version 1.18.0) (Fernandes et al., 2014a) which generated Monte Carlo samples of Dirichlet distributions for each sample, using a uniform prior, performed CLR transformation of each realization, and then performed Wilcoxon tests on the transformed realizations. The function then returned the expected Benjamini-Hochberg (BH) FDR-corrected p-value for each feature based on the results the different across Monte Carlo samples.

3.3.2.2 - ANCOM-II

We ran the non-rarefied feature table through the R ANCOM-II (Kaul, Mandal, Davidov, & Peddada, 2017; Mandal et al., 2015) (https://github.com/FrederickHuangLin/ANCOM) (version 2.1) function feature_table_pre_process, which first examined the abundance table to identify outlier zeros and structural zeros (Kaul et al., 2017). The following packages were imported by ANCOM-II: exactRankTests (version 0.8.31), nlme (version 3.1.149), dplyr (version 0.8.5), ggplot2 (version 3.3.0) and compositions (version 1.40.2). Outlier zeros, identified by finding outliers in the distribution of taxon counts within each sample grouping, were ignored during differential abundance analysis, and replaced with NA. Structural zeros, taxa that were absent in one grouping but present in the other, were ignored during data analysis and automatically called as differentially abundant. A pseudo count of 1 was then applied across the dataset to allow for log transformation. Using the main function ANCOM, all additive log-ratios for each taxon were then tested for significance using Wilcoxon rank-sum tests, and p-values were FDR-corrected using the BH method. ANCOM-II then applied a detection threshold as described in the original paper (Mandal et al., 2015), whereby a taxon was called as DA if the number of corrected p-values reaching nominal significance for that taxon was greater than 90% of the maximum possible number of significant comparisons.

3.3.2.3 - corncob

We converted the metadata and non-rarefied feature tables into a phyloseq object (version 1.29.0) (McMurdie & Holmes, 2013), which we input to corncob's differentialTest function (version 0.1.0) (B. D. Martin et al., 2020). This function fits each taxon count abundance to a beta-binomial model, using logit link functions for both the mean and overdispersion. Because corncob models each of these simultaneously and performs both differential abundance and differential variability testing (B. D. Martin et al., 2020), we set the null overdispersion model to be the same as the non-null model so that only taxa having differential abundances were identified. Finally, the function performed significance testing, for which we chose Wald tests (with the default non-bootstrap setting), and we obtained BH FDR-corrected p-values as output.

3.3.2.4 - DESeq2

We first passed the non-rarefied feature tables to the DESeq function (version 1.26.0) (Love et al., 2014) with default settings, except that instead of the default relative log expression (also known as the median-of-ratios method) the estimation of size factors was set to use “poscounts”, which calculates a modified relative log expression that helps account for features missing in at least one sample. The function performed three steps: (1) estimation of size factors, which are used to normalize library sizes in a model-based fashion; (2) estimation of dispersions from the negative binomial likelihood for each feature, and subsequent shrinkage of each dispersion estimate towards the parametric (default) trendline by empirical Bayes; (3) fitting each feature to the specified class groupings with negative binomial generalized linear models and performing hypothesis testing, for which we chose the default Wald test. Finally, using the results function, we obtained the resulting BH FDR-corrected p-values.

3.3.2.5 - edgeR

Using the phyloseq_to_edgeR function (https://joey711.github.io/phyloseq-extensions/edgeR.html), we added a pseudocount of 1 to the non-rarefied feature table and used the function calcNormFactors from the edgeR R package (version 3.28.1) (Robinson & Oshlack, 2010) to compute relative log expression normalization factors. Negative binomial dispersion parameters were then estimated using the functions estimateCommonDisp followed by estimateTagwiseDisp to shrink feature-wise dispersion estimates through an empirical Bayes approach. We then used the exactTest for negative binomial data (Robinson & Oshlack, 2010) to identify features that differ between the specified groups. The resulting p-values were then corrected for multiple testing with the BH method with the function topTags.

3.3.2.6 - LEfSe

The rarefied feature table was first converted into LEfSe format using the LEfSe script format_input.py (Segata et al., 2011). We then ran LEfSe on the formatted table using the run_lefse.py script with default settings and no subclass specifications. Briefly, this command first normalized the data using total sum scaling, which divides each feature count by the total library size. Then it performed a Kruskal-Wallis (which in our two-group case reduces to the Wilcoxon rank-sum) hypothesis test to identify potential differentially abundant features, followed by linear discriminant analysis (LDA) of class labels on abundances to estimate the effect sizes for significant features. From these, only those features with scaled LDA analysis scores above the threshold score of 2.0 (default) were called as differentially abundant. This key step is what distinguished LEfSe from the Wilcoxon test approach based on relative abundances that we also ran. In addition, no multiple-test correction was performed on the raw LEfSe output as only the p-values of significant features above-threshold LDA scores are returned by this tool.

3.3.2.7 - limma voom

We first normalized the non-rarefied feature table using the edgeR calcNormFactors function, with either the trimmed mean of M-values (TMM) or TMM with singleton pairing (TMMwsp) option. We choose to run this tool with two different normalization functions as we found the standard TMM normalization technique to struggle with highly spare datasets despite it previously being shown to perform preferentially in DA testing. Furthermore, the TMMwsp method is highlighted within the edgeR package as an alternative for highly sparse data. During either of these normalization steps a single sample was chosen to be a reference sample using upper-quartile normalization. This step failed in some highly sparse abundance tables; in these cases, we instead chose the sample with the largest sum of square-root transformed feature abundances to be the reference sample. After normalization, we used the limma R package (version 3.42.2) (Ritchie et al., 2015) function voom to convert normalized counts to $\log_2$-counts-per-million and assign precision weights to each observation based on the mean-variance trend. We then used the functions lmFit, eBayes, and topTable in the limma R package to fit weighted linear regression models, perform tests based on an empirical Bayes moderated t-statistic (Phipson, Lee, Majewski, Alexander, & Smyth, 2016) and obtain BH FDR-corrected p-values.

3.3.2.8 - MaAsLin2

We entered either a rarefied or non-rarefied feature table into the main Maaslin2 function within the MaAsLin2 R package (version 0.99.12) (Mallick et al., 2021). We specified arcsine square-root transformation as in the package vignette (instead of the default log) and total sum scaling normalization. For consistency with other tools, we specified no random effects and turned off default standardization. The function fit a linear model to each feature's transformed abundance on the specified sample grouping, tested significance using a Wald test, and output BH FDR-corrected p-values.

3.3.2.9 - metagenomeSeq

We first entered the counts and sample information to the function newMRexperiment from the metagenomeSeq R package (version 1.28.2) (Paulson et al., 2013). Next, we used cumNormStat and cumNorm to apply cumulative sum-scaling normalization, which attempts to normalize sequence counts based on the lower-quartile abundance of features. We then used fitFeatureModel to fit normalized feature counts with zero-inflated log-normal models (with pseudo-counts of 1 added prior to $\log_2$ transformation) and perform empirical Bayes moderated t-tests, and MRfulltable to obtain BH FDR-corrected p-values.

3.3.2.10 - t-test

We applied total sum scaling normalization to the rarefied feature table and then performed an unpaired Welch's t-test for each feature to compare the specified groups. We corrected the resulting p-values for multiple testing with the BH method.

3.3.2.11 - Wilcoxon test

Using raw feature abundances in the rarefied case, and CLR-transformed abundances (after applying a pseudocount of 1) in the non-rarefied case, we performed Wilcoxon rank-sum tests for each feature to compare the specified sample groupings. We corrected the resulting p-values with the BH method.

3.3.3 - Comparing Numbers of Significant Hits Between Tools

We compared the number of significant ASVs each tool identified in 38 different datasets. Each tool was run as described above using default settings with some modifications suggested by the tool authors, as noted above. A heatmap representing the number of significant hits found by each tool was constructed using the pheatmap R package (version 1.0.12) (Kolde, 2012). Spearman correlations between the percent of significant ASVs identified by a tool and the following dataset characteristics were computed using the cor.test function in R: sample size, Aitchison's distance effect size as computed using a PERMANOVA test (adonis; vegan, version 2.5.6) (Dixon, 2003), sparsity, mean sample ASV richness, median sample read depth, read depth range between samples and the coefficient of variation for read depth within a dataset. In addition, for the unfiltered analyses, we also computed Spearman correlations with the percent of ASVs below 10% prevalence in each dataset (i.e., the percent of ASVs that would be removed to produce the filtered datasets). Correlations were displayed using the R package corrplot (version 0.85) and gridExtra (version 2.3). Dataset manipulation for plotting and reshaping were conducting using the following R packages: doMC (version 1.3.5), doParallel (version 1.0.15), matrixStats (version 0.56.0), reshape2 (version 1.4.4), plyr (version 1.8.6) and tidyverse (version 1.3.0).

3.3.4 - Cross-Tool, Within-Study Differential Abundance Consistency Analysis

We compared the consistency between different tools within all datasets by pooling all ASVs identified as being significant by at least one tool in the 38 different datasets. The number of methods that identified each ASV as differentially abundant were then tallied. A second way of examining the between method consistency without choosing a specific significance threshold was to examine the overlap between the top 20 ASVs identified by each DA method. To do this ASVs were ranked for each DA method depending on their significance value apart from ANCOM-II, where its W statistic was used for ranking. Like the above analysis, we than tallied the number of methods that identified each ASV as being in its top 20 most differentially abundant ASVs. Multi-panel figures were combined using the R package cowplot (version 1.0.0) and ggplotify (version 0.0.5). To get another view of the data principal coordinate analysis plots were constructed using the mean inter-tool Jaccard distance across the 38 main datasets. Distances were computed by averaging over the inter-tool distance matrices for all individual datasets to weight each dataset equally using the R packages vegan (version 2.5.6) and parallelDist (version 0.2.4). Labels were displayed using the R package ggrepel (version 0.8.1).

3.3.5 - False Positive Analysis

To evaluate the false positive rates of each DA method, eight datasets were selected for analysis based on having the largest sample sizes, while also being from diverse environment types. In each dataset, only the most frequent sample group was chosen for analysis to help ensure similar composition among samples tested. Within this grouping, random labels of either case or control were assigned to samples and the various differential abundance methods were tested on them. This was replicated 100 times for each dataset and tool combination aside from ALDEx2, ANCOM-II, and Corncob. These were run using 100 replicates in only 3 of the 8 datasets (Freshwater – Arctic, Soil – Blueberry, Human - OB (1)) with 100 ALDEx2 replications also being run in the Human - HIV (3) dataset. This was due to the long computational time required to run these tools on all datasets. The remaining datasets were replicated 10 times for each of these three tools. After this analysis was completed, the number of differentially abundant ASVs identified by each tool was assessed at an alpha value of 0.05. Boxplots for this data was constructed using the R packages ggplot2 (version 3.3.0), ggbeeswarm (version 0.6.0), and scales (version 1.1.0).

3.3.6 - Cross-Study Differential Abundance Consistency Analysis

For this analysis we acquired two additional pre-processed datasets that were not used for other analyses, which are the GEMS1 (Pop et al., 2014) and the dia_schneider (Schneider et al., 2017), datasets (Supplemental Data 1). The processed data for these datasets was acquired from the MicrobiomeDB (Oliveira et al., 2018) and the microbiomeHD (Duvallet et al., 2017) databases, respectively. These datasets were combined with three of the datasets used elsewhere in this chapter (Human – C. diff [1 and 2] and Human – Inf.), to bring the number of diarrhea-related datasets to five. These three pre-existing datasets all related to enteric infections that had all been previously demonstrated to show a distinct signal of microbial differences driven by diarrhea in patient samples (Duvallet et al., 2017).

For the obesity cross-study analysis we leveraged four datasets that were part of the core 38 datasets: Human – OB (1-4) (Goodrich, Waters, et al., 2014; Ross et al., 2015; Turnbaugh et al., 2009; L. Zhu et al., 2013). We also included an additional obesity dataset, ob_zupancic (Zupancic et al., 2012), that we acquired from the microbiomeHD database.

The ASVs in each of these datasets were previously taxonomically classified and so we used these classifications to collapse all feature abundances to the genus level. Note that taxonomic classification was performed using several different methods, which represents another source of technical variation. We excluded unclassified and sensu stricto-labelled genus levels. We then ran all differential abundance tools on these datasets at the genus level. These comparisons were between the diarrhea and non-diarrhea sample groups. The same processing workflow was used for the supplementary obesity dataset comparison as well.

For each tool and study combination, we determined which genera were significantly different at an alpha of 0.05 (where relevant). For each tool we then tallied the number of times each genus was significant, i.e., how many datasets each genus was significant in based on a given tool. The null expectation distributions of these counts per tool were generated by randomly sampling genera from each dataset. The probability of sampling a genus (i.e., calling it significant) was set to be equal to the proportion of actual significant genera. This procedure was repeated 1000 times, with genus replicates equal to the actual number of tested genera (218 and 116 for the diarrhea and obesity datasets, respectively). For each replicate we tallied the number of times the genus was sampled across datasets. Note that to simplify this analysis we ignored the directionality of the significance (e.g., whether it was higher in case or control samples). We also excluded genera never found to be significant. We computed the mean of these 1000 distributions to generate an empirical distribution of the expected mean number of studies where a genus would be called as significant, given random sampling. We determined where the observed mean values lay on each corresponding distribution to calculate statistical significance.

3.3.7 - Discriminatory Analysis

We calculated the discriminatory value of each ASV (i.e., the extent to which the ASV can be used to distinguish the sample groups) based on the area under the curve (AUC) of the receiver operator curve (ROC) for that ASV. This was performed independently for both non-rarefied relative and CLR abundances. For each ASV in a dataset the abundance of that ASV along with metadata groupings was used as input into the prediction function in the ROCR R package (Sing, Sander, Beerenwinkel, & Lengauer, 2005). Multiple different optimal abundance cut-offs were then used to classify samples based on the input ASVs abundance. Classifications were then compared to the true sample groupings to generate ROCs for each ASV within the 38 tested datasets. For each tool the mean

AUC of all ASVs identified as being differentially abundant in each dataset was computed, based on both relative and CLR abundances separately. We then calculated the precision, recall and F1 scores of each tool for the tested datasets when AUC cut-offs of 0.7 or 0.9 were used. In each case the "true positives" were treated as features that were above the specified AUC threshold.

3.4 – Results

3.4.1 - Microbiome Differential Abundance Methods Produce a Highly Variable Number of Significant ASVs Within the Same Microbiome Datasets

To investigate how different DA tools impact biological interpretations across microbiome datasets, we tested 14 different DA testing approaches (Table 3.1) on 38 different microbiome datasets with a total of 9,405 samples. These datasets corresponded to a range of environments, including the human gut, plastisphere, freshwater, marine, soil, wastewater, and built environments (Supplemental Data 1). The features in these datasets corresponded to both ASVs and clustered operational taxonomic units, but we refer to them all as ASVs below for simplicity.

Table 3.1: ***Differential abundance tools compared in this study.***

Tool (version)	Input	Norm.	Trans.	Distribution	Covariates	Random Effects	Hypothesis test	FDR Corr.	CoDa	Dev. For
ALDEx2 (1.18.0)	Counts	None	CLR	Dirichlet-multinomial	Yes*	No	Wilcoxon rank-sum	Yes	Yes	RNA-seq, 16S, MGS
ANCOM-II **(2.1)**	Counts	None	ALR	Non-parametric	Yes	Yes	Wilcoxon rank-sum	Yes	Yes	MGS
Corncob (0.1.0)	Counts	None	None	Beta-binomial	Yes	No	Wald (default)	Yes	No	16S, MGS
DESeq2 (1.26.0)	Counts	Modified RLE (default is RLE)	None	Negative binomial	Yes	No	Wald (default)	Yes	No	RNA-seq, 16S, MGS
edgeR (3.28.1)	Counts	RLE (default is TMM)	None	Negative binomial	Yes*	No	Exact	Yes	No	RNA-seq
LEFse	Rarefied Counts	TSS	None	Non-parametric	Sub-class factor only	No	Kruskal-Wallis	No	No	16S, MGS
MaAsLin2 (1.0.0)	Counts	TSS	AST (default is log)	Normal (default)	Yes	Yes	Wald	Yes	No	MGS
MaAsLin2 (rare) **(1.0.0)**	Rarefied counts	TSS	AST (default is log)	Normal (default)	Yes	Yes	Wald	Yes	No	MGS
metagenomeSeq **(1.28.2)**	Counts	CSS	Log	Zero-inflated (log-) Normal	Yes	No	Moderated t	Yes	No	16S. MGS
limma voom (TMM) **(3.42.2)**	Counts	TMM	Log; Precision weighting	Normal (default)	Yes	Yes	Moderated t	Yes	No	RNA-seq
limma voom (TMMwsp) **(3.42.2)**	Counts	TMMwsp	Log; Precision weighting	Normal (default)	Yes	Yes	Moderated t	Yes	No	RNA-seq
t-test (rare)	Rarefied Counts	None	None	Normal	No	No	Welch's t-test	Yes	No	N/A
Wilcoxon (CLR)	CLR abundances	None	CLR	Non-parametric	No	No	Wilcoxon rank-sum	Yes	Yes	N/A
Wilcoxon (rare)	Rarefied counts	None	None	Non-parametric	No	No	Wilcoxon rank-sum	Yes	No	N/A

** The tool supports additional covariates if they are provided. ANCOM-II automatically performs ANOVA in this case, ALDEx2 requires that users select the test, and edgeR requires use of a different function (glmFit or glmQLFit instead of exactTest).*

Abbreviations: ALR, additive log-ratio; AST, arcsine square-root transformation; CLR, centered log-ratio; CoDa, compositional data analysis; CSS, cumulative sum scaling; FDR Corr., false-discovery rate correction; MGS, metagenomic sequencing; RLE, relative log expression; TMM, trimmed mean of M-values; Trans., transformation; TSS, total sum scaling

We also investigated how prevalence filtering each dataset prior to analysis impacted the observed results. We chose to either use no prevalence filtering (Figure 3.1A) or a 10% prevalence filter that removed any ASVs found in fewer than 10% of samples within each dataset (Figure 3.1B).

We found that in both the filtered and unfiltered analyses the percentage of significant ASVs identified by each DA method varied widely across datasets, with means ranging from 3.8-32.5% and 0.8-40.5%, respectively. Interestingly, we found that many tools behaved differently between datasets. Specifically, some tools identified the most features in one dataset while identifying only an intermediate number in other datasets. This was especially evident in the unfiltered datasets (Figure 3.1A).

Despite the variability of tool performance between datasets, we did find that several tools tended to identify more significant hits (Supplemental Figure 1C-D). In the unfiltered datasets, we found that limma voom (TMMwsp; mean: 40.5%; SD: 41% / TMM; mean: 29.7%; SD: 37.5%), Wilcoxon (CLR; mean: 30.7%; SD: 42.3%), LEfSe (mean: 12.6%; SD: 12.3%), and edgeR (mean: 12.4%; SD: 11.4%) tended to find the largest number of significant ASVs compared with other methods. Interestingly, in a few datasets, such as the Human-ASD and Human-OB (2) datasets, edgeR found a higher proportion of significant ASVs than any other tool. In addition, we found that limma voom (TMMwsp) found the majority of ASVs to be significant (73.5%) in the Human-HIV (3) dataset while the other tools found 0-11% ASVs to be significant (Figure 3.1A). Disturbingly, we found that both limma voom methods identified over 99% of ASVs to be significant in several cases such as the Built-Office and Freshwater-Arctic datasets. This is most likely due to the high sparsity of these datasets causing the tools' reference sample selection method (upper-quartile normalization) to fail. Such extreme findings were also seen in the Wilcoxon (CLR) output, where more than 90% of ASVs were called as significant in eight separate datasets. We found similar, although not as extreme, trends with LEfSe where in some datasets, such as the Human-T1D (1) dataset, the tool found a much higher percentage of significant hits (3.5%) compared with all other tools (0-0.4%). This observation is most likely a result of LEfSe filtering significant features by effect size rather than using FDR correction to reduce the number of false positives. We found that two of the three compositionally aware methods we tested identified fewer significant ASVs than the other tools tested. Specifically, ALDEx2 (mean: 1.4%; SD: 3.4%) and ANCOM-II (mean: 0.8%; SD: 1.8%) identified the fewest significant ASVs. We found the conservative behavior of these tools to be consistent across all 38 datasets we tested.

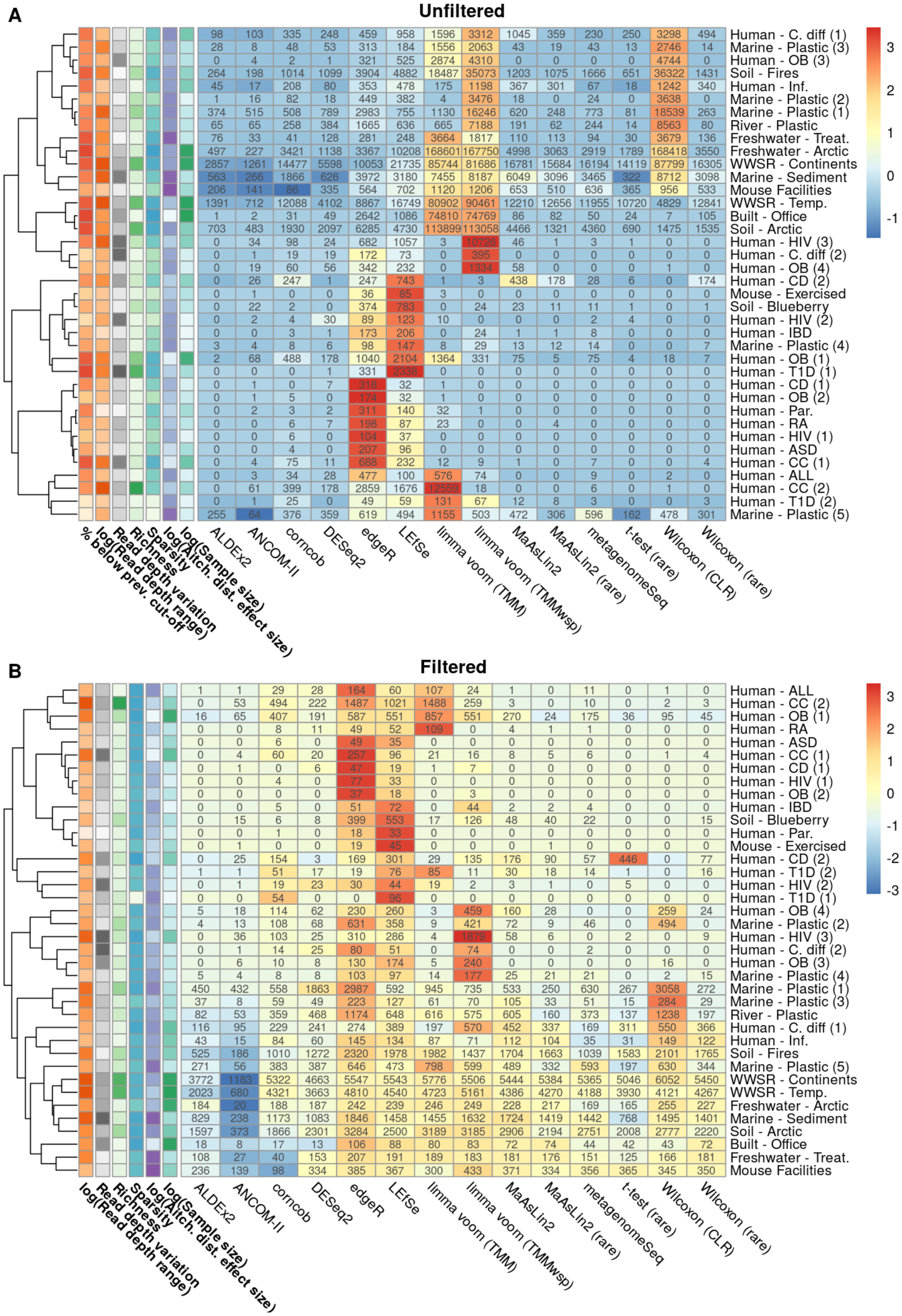


Figure 3.1: ***Variation in the proportion of significant features depending on the differential abundance method and dataset.*** *Heatmaps indicate the numbers of significant amplicon sequence variants (ASVs) identified in each dataset by the*

corresponding tool based on ***a*** *unfiltered data and* ***b*** *10% prevalence-filtered data. Cells are colored based on the standardized (scaled and mean centered) percentage of significant ASVs for each dataset. Additional colored cells in the left-most six columns indicate the dataset characteristics we hypothesized could be driving variation in these results (darker colors indicate higher values). Datasets were hierarchically clustered based on Euclidean distances using the complete method. Abbreviations: prev., previous; TMM, trimmed mean of M-values; TMMwsp, trimmed mean of M-values with singleton pairing; rare, rarefied; CLR, center-log-ratio. Source data are provided as a Source Data file.*

Overall, the results based on the filtered tables were similar, although there was a smaller range in the number of significant features identified by each tool. All tools except for ALDEx2 found a lower number of total significant features when compared with the unfiltered dataset (Supplemental Figure 1C-D). As with the unfiltered data, ANCOM-II was the most stringent method (mean: 3.8%; SD: 5.9%), while edgeR (mean: 32.5%; SD: 28.5%), LEfSe (mean: 27.5%; SD: 25.0%), limma voom (TMMwsp; mean: 27.3%; SD: 30.1% / TMM; mean: 23.5%; SD: 27.7%), and Wilcoxon (CLR; mean: 25.4%; SD: 31.7%) tended to output the highest numbers of significant ASVs (Figure 3.1B).

To investigate possible factors driving this variation we examined how the number of ASVs identified by each tool correlated with several variables. These variables included dataset richness, variation in sequencing depth between samples, dataset sparsity, and Aitchison's distance effect size (based on PERMANOVA tests). As expected, we found that the number of ASVs identified by all tools positively correlated with the effect size between test groups with Spearman correlation coefficient values ranging between 0.35-0.72 with unfiltered data (Figure 3.2A) and 0.31-0.56 for filtered data (Figure 3.2B). We also found in the filtered datasets that the number of ASVs found by all tools significantly correlated with the median read depth, range in read depth, and sample size. There was much less consistency in these correlations across the unfiltered data. For instance, only the t-test, both Wilcoxon methods, and both limma voom methods correlated significantly with the range in read depth (Figure 3.2B). We also found that edgeR was negatively correlated with mean sample richness in the unfiltered analysis. The percentage of ASVs in the unfiltered datasets that were lower than 10% prevalence was also significantly associated with the output of several tools. We also investigated if chimeras could influence the number of significant ASVs detected with results showing very limited impact (Supplemental Figure 2).

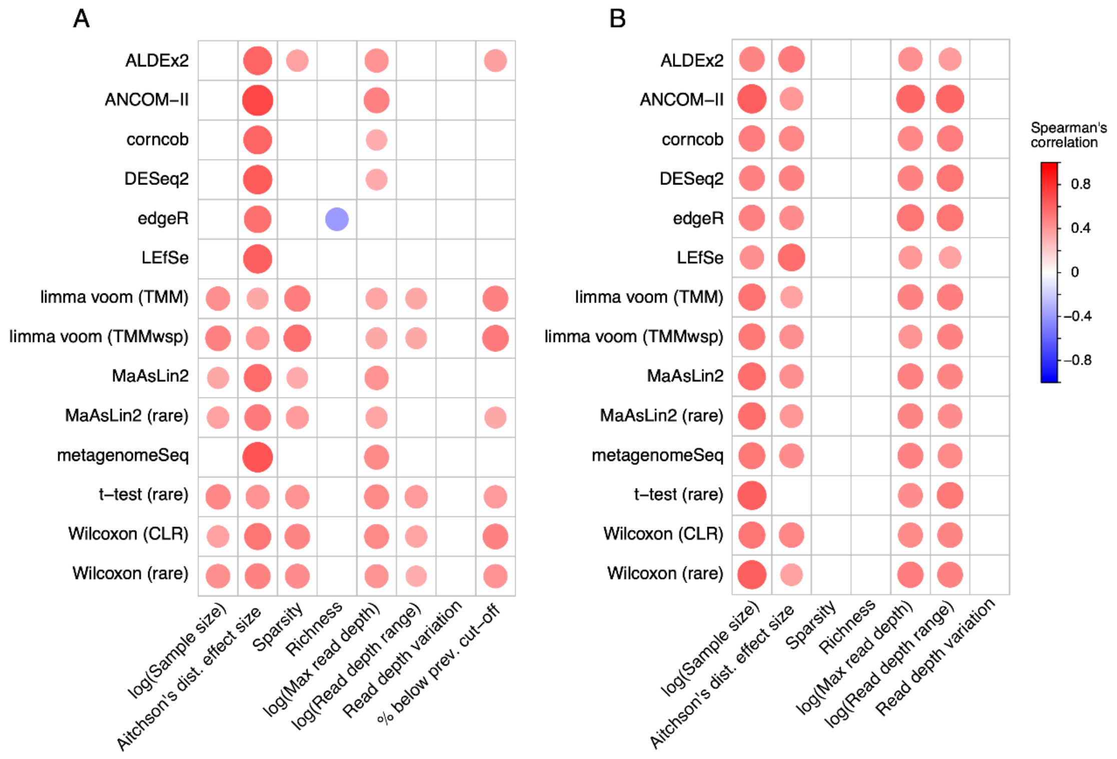


Figure 3.2: **Dataset characteristics associated with percentage of significant amplicon sequence variants.** *The correlation coefficients (Spearman's rho) are displayed by size and color for the* ***a*** *unfiltered and* ***b*** *prevalence-filtered data. Each dots size and color both correspond to its correlation coefficient. These correspond to the dataset characteristics correlated with the percentage of significant amplicon sequence variants identified by that tool per dataset. Only significant correlations before multiple comparison correction (p < 0.05) are displayed. Abbreviations: prev., previous; TMM, trimmed mean of M-values; TMMwsp, trimmed mean of M-values with singleton pairing; rare, rarefied; CLR, center-log-ratio. Source data are provided as a Source Data file.*

We next investigated whether the significant ASVs identified by the tested DA tools were, on average, at different relative abundances. The clearest outliers were ALDEx2 (median relative abundance of significant ASVs: 0.013%), ANCOM-II (median: 0.024%), and to a lesser degree DESeq2 (median: 0.007%), which tended to find significant features that were at higher relative abundance in the unfiltered datasets (Supplemental Figure 1A; the medians for all other tools ranged from 0.00004-0.003%). A similar trend for ALDEx2 (median: 0.011%) and ANCOM-II (median: 0.029%) was also apparent in the filtered datasets (Supplemental Figure 1B; the medians for all other tools ranged from 0.005-0.008%).

Finally, we also examined the discriminatory value of the significant ASVs identified by each tool in the filtered datasets. By discriminatory value we are referring to how well the sample groups can be delineated by a single ASV using hard cut-off abundance values. For this analysis we used either the relative abundances (Supplemental Figure 3A) or the CLR abundances (Supplemental Figure 3B) of each significant ASV input into receiver operator curves (ROC) predicting the groups of interest. Raw abundance values were used as input and multiple optimal cut-off points were selected to produce ROCs comparing sensitivity to specificity. We then measured the area under this curve (AUC) of each significant ASV and calculated the mean value across all ASVs identified by each tool. We found that ASVs identified by either ALDEx2 or ANCOM-II had the highest mean AUROC across the tested datasets using both relative abundances and CLR abundances as input (Supplemental Figure 3). Despite this trend, there were instances where these tools failed to identify any significant ASVs despite other tools achieving relatively high mean AUROCs for the ASVs they identified. For example, in the human-IBD dataset several tools found mean AUROCs of the ASVs they identified ranging from 0.8-0.9 using either CLR or relative abundances as input while both ALDEx2 and ANCOM-II failed to identify any significant ASVs.

As the above analysis has the potential to penalize tools that call a higher number of ASVs that are of lower discriminatory values we also investigated the ability of the tested DA methods to identify ASVs above specific AUC thresholds (Supplemental Figure 4). An important assumption of this analysis is that to accept the exact performance values all ASVs above and below the selected AUROC threshold must be true and false positives, respectively. Although this strict assumption is almost certainly false, it is likely that ASVs above and below the AUC threshold are at least enriched for true and false positives, respectively. We found that at an AUROC threshold of 0.7 both ANCOM-II and ALDEx2 had the highest precision for both relative abundance (medians: 0.99; SD: 0.36 and median: 0.82, SD:0.39) and CLR data (medians: 1.0; SD:0.35 and median: 0.83, SD:0.35). However, they suffered lower recall values based on both relative (medians: 0.17 and 0.02) and CLR-based (medians: 0.06 and 0.01) abundances when compared to the recall scores of tools such as LEfSe and edgeR on relative abundance (median: 0.96 and 0.69) and CLR data (medians: 0.50 and 0.34). When examining CLR data as input we found that limma voom (TMMwsp) had one of the highest F1 scores (median: 0.47) only being outcompeted by the Wilcoxon (CLR) test (mean: 0.70). Examining the data at a higher AUC threshold of 0.9 showed that all tools had relatively high recall scores, apart from some tools such as ANCOM-II, corncob, and t-test (rare) on CLR data (medians: 0.5, 0.5, and 0.20). The precision score of all tools at this threshold was low on both relative abundance (range: 0-0.01) and CLR data (range: 0-0.2). This result is unsurprising as in practice we would expect DA tools to identify features that are below a discriminatory threshold of 0.9 AUC.

3.4.2 - High Variability of Overlapping Significant ASVs

We next investigated the overlap in significant ASVs across tools within each dataset. These analyses provided insight into how similar the interpretations would be depending on which DA method was applied. We hypothesize that tools that produce significant ASVs that highly intersect with the output of other DA tools are the most accurate approaches. Conversely, this may not be the case if tools with similar approaches produce similar sets of significant ASVs and end up identifying the same spurious ASVs. Either way, identifying overlapping significant ASVs across methods provides insights into their comparability.

Based on the unfiltered data, we found that limma voom methods identified similar sets of significant ASVs that were different from those of most other tools (Figure 3.3A). However, we also found that many of the ASVs identified by the limma voom methods were also identified as significant based on the Wilcoxon (CLR) approach, despite these being highly methodologically distinct tools. Furthermore, the two Wilcoxon test approaches had highly different consistency profiles, which highlights the impact that CLR-transforming has on downstream results. In contrast, we found that both MaAsLin2 approaches had similar consistency profiles, although the non-rarefied method found slightly lower-ranked features. We also found that the most conservative tools, ALDEx2 and ANCOM-II, primarily identified features that were also identified by almost all other methods. In contrast, edgeR and LEfSe, two tools that often identified the most significant ASVs, output the highest percentage of ASVs that were not identified by any other tool: 11.4% and 9.6%, respectively. Corncob, metagenomeSeq, and DESeq2 identified ASVs at more intermediate consistency profiles.

The overlap in significant ASVs based on the prevalence-filtered data was similar overall to the unfiltered data results (Figure 3.3B). One important exception was that the limma voom approaches identified a much higher proportion of ASVs that were also identified by most other tools, compared with the unfiltered data. Nonetheless, similar to the unfiltered data results, the Wilcoxon (CLR) significant ASVs displayed a bimodal distribution and a strong overlap with limma voom methods. We also found that overall, the proportion of ASVs consistently identified as significant by more than 12 tools was much higher in the filtered data (mean: 38.6%; SD: 15.8%) compared with the unfiltered data (mean: 17.3%; SD: 22.1%). In contrast with the unfiltered results, corncob, metagenomeSeq, and DESeq2 had lower proportions of ASVs at intermediate consistency ranks. However, ALDEx2 and ANCOM-II once again produced significant ASVs that largely overlapped with most other tools.

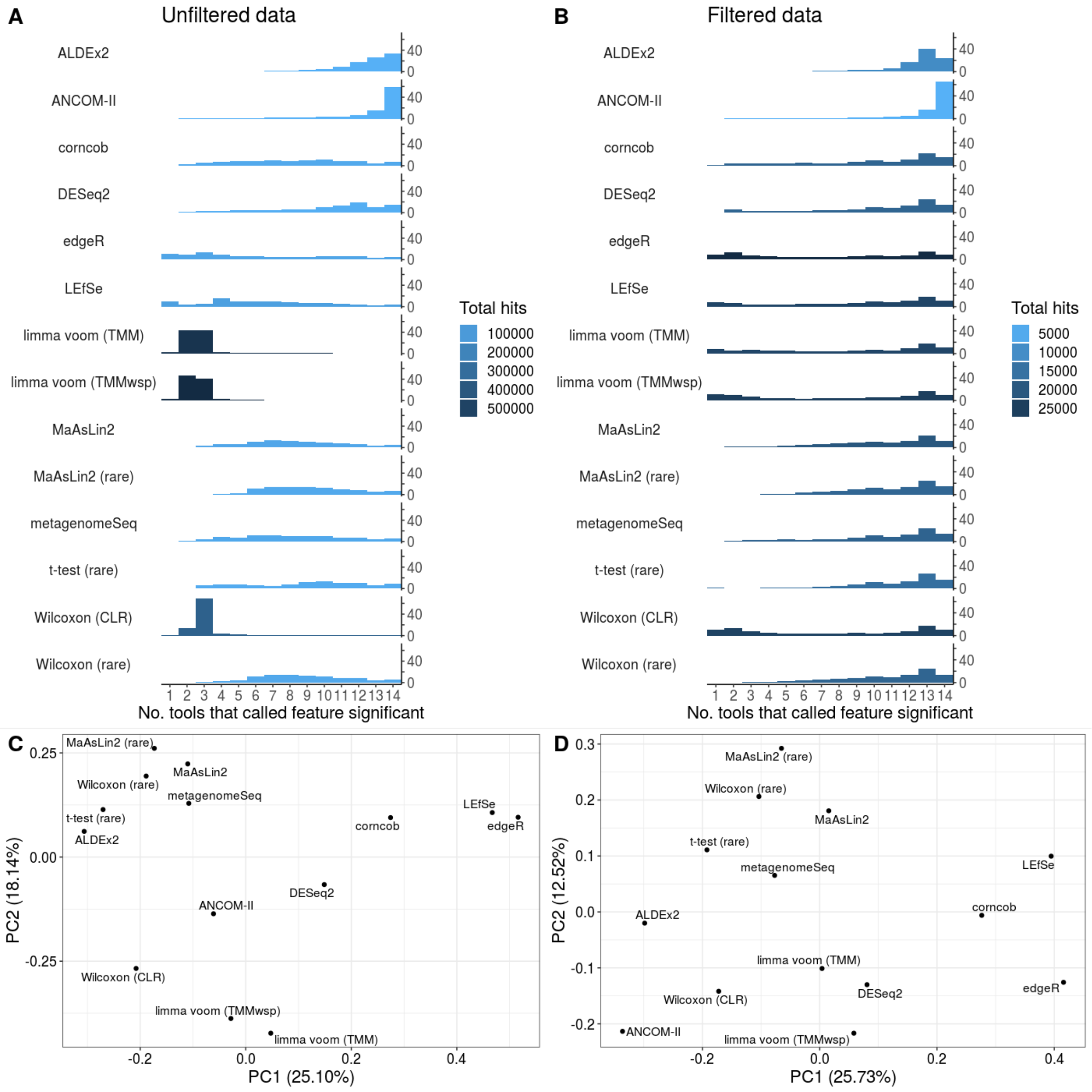


Figure 3.3: ***Overlap of significant features across tools and tool clustering.*** *a, b The number of tools that called each feature significant, stratified by features called by each individual tool for the a unfiltered and b 10% prevalence-filtered data. Results are shown as a percentage of all ASVs identified by each tool. The total number of significant features identified by each tool is indicated by the bar colors. For example, based on the unfiltered data these bars indicate that almost 40% of significant ASVs identified by ALDEx2 were shared across all other tools, while ALDEx2 did not identify any*

significant ASVs shared by fewer than eight tools. Note that when interpreting these results that they are dependent on which methods were included, and whether they are represented multiple times. For instance, two different workflows for running MaAslin2 are included, which produced similar outputs. c, d Plots are displayed for the first two principal coordinates (PCs) for both c non-prevalence-filtered and d 10% prevalence-filtered data. These plots are based on the mean inter-tool Jaccard distance across the 38 main datasets that we analyzed, computed by averaging over the inter-tool distance matrices for all individual datasets to weight each dataset equally. Abbreviations: TMM, trimmed mean of M-values; TMMwsp, trimmed mean of M-values with singleton pairing; rare, rarefied; CLR, center-log-ratio. Source data are provided as a Source Data file.

A major caveat of the above analysis is that each DA tool produced substantially different numbers of ASVs in total. Accordingly, in principle all of the tools could be identifying the same top ASVs and simply taking varying degrees of risk when identifying less clearly differential ASVs. To investigate this possibility, we identified the overlap between the 20 top-ranked ASVs per dataset (Supplemental Figure 5), which included non-significant (but relatively highly ranked) ASVs in some cases. The distribution of these ASVs in the 20 top-ranked hits for all tools was similar to that of all significant ASVs described above. For example, in the filtered data we found that, t-test (rare) (mean: 6.2; SD: 3.8), edgeR, (mean: 6.5; SD: 4.6), corncob (mean: 6.0; SD: 4.1), metagenomeSeq (mean: 5.2; SD: 4.7) and DESeq2 (mean: 4.7; SD: 4.5) had the highest number of ASVs only identified by that particular tool as being in the top 20. On average only a small number of ASVs were amongst the top 20 ranked of all tools in both the filtered (mean: 0.21; SD: 0.62) and unfiltered (mean: 0.11; SD: 0.31) datasets

(Supplemental Figure 5).The above analyses summarized the consistency in tool outputs, but it is difficult to discern which tools performed most similarly from these results alone. To identify overall similarly performing tools we conducted principal coordinates analysis based on the Jaccard distance between significant sets of ASVs (Figure 3.3C-D). These analyses provide insight into how similar the results of different tools are expected to be, which could be due to methodological similarities between them. However, this does not provide clear evidence for which tools are the most accurate. One clear trend for both unfiltered and filtered data is that edgeR and LEfSe cluster together and are separated from other methods on the first principal coordinate. Interestingly, corncob, which is a methodologically distinct approach, also clusters relatively close to these two methods on the first principal component. This may reflect that the distributions that these two methods rely upon become similar when considering the parameter values often associated with microbiome data.

The major outliers on the second principal coordinate differ depending on whether the data was prevalence-filtered. For the unfiltered data, the main outliers are the limma voom methods, followed by Wilcoxon (CLR; Figure 3.3C). In contrast, ANCOM-II is the sole major outlier on the second principal component based on filtered data (Figure 3.3D). These visualizations highlight the major tool clusters based on the resulting sets of significant ASVs. However, the percentage of variation explained by the top two components is relatively low in each case, which means that substantial information regarding potential tool clustering is missing from these panels (Supplemental Figure 6 and Supplemental Figure 7). For instance, ANCOM-II and corncob are major outliers on the third and fourth principal coordinates, respectively, of the unfiltered data analysis, which highlights the uniqueness of these methods.

3.4.3 - False Discovery Rate of Microbiome Differential Abundance Tools Depends on the Dataset

We next evaluated how the DA tools performed in cases where no significant hits were expected. These cases corresponded to sub-samplings of eight of the 38 datasets presented above. For each dataset we selected the most frequently sampled group and within this sample grouping we randomly reassigned them as case or control samples. Each DA tool was then run on this subset of randomly assigned samples and results were compared. Due to the random nature of assignment and the similar composition of samples from the same metadata grouping (e.g. healthy humans) we would expect tools to not identify any ASVs as being differential abundant. Through this approach we were able to infer the false positive characteristics of each tool. In other words, we determined what percentage of tested ASVs was called as significant by each tool even when there is no difference expected between the sample groups.

The clearest trend for both unfiltered and filtered data is that certain outlier tools have relatively high FDRs in this context while most others identify few false positives (Figure 3.4). Most strikingly, both limma voom methods output highly variable percentages of significant ASVs, especially based on the unfiltered data (Figure 3.4A). In 5/8 of the unfiltered datasets, the limma voom methods identified more than 5% of ASVs as significant on average due to many high value outliers. Only ALDEx2 and t-test (rare) consistently identified no ASVs as significantly different in the unfiltered data analyses. However, both MaAsLin2 and Wilcoxon (rare) found no significant features in the majority of tested datasets (6/8 and 7/8 respectively). Two clear outliers in the filtered data analyses were edgeR (mean: 0.69% - 27.9%) and LEfSe (mean: 3.4% - 5.1%) which consistently identified more significant hits compared with other tools (Figure 3.4B). However, it should be noted that in some datasets Corncob, DESeq2 and the limma methods also performed poorly.

Overall, we found that the raw numbers of significant ASVs were lower in the filtered dataset than in the unfiltered data (as expected due to many ASVs being filtered out), and that most tools identified only a small percentage of significant ASVs, regardless of filtering procedure. The exceptions were the two limma voom methods, which had high FDRs with unfiltered data, and edgeR and LEfSe, which had high FDRs on the filtered data. Although these tools stand out on average, we also observed that in several replicates on the unfiltered datasets, the Wilcoxon (CLR) approach identified almost all features as statistically significant (Figure 3.4A). This was also true for both limma voom methods, which highlights that a minority of replicates are driving up the average FDR of these methods.

We investigated the outlier replicates for the Wilcoxon (CLR) approach and found that the mean differences in read depth between the two tested groups were consistently higher in replicates in which 30% or more of ASVs were significant (Supplemental Figure 8). These differences were associated with similar differences in the geometric mean abundances per-sample (i.e., the denominator of the CLR transformation) between the test groups. Specifically, per dataset, these outlier replicates commonly displayed the most extreme mean difference in geometric mean between the test groups and were otherwise amongst the top ten most extreme replicates.

Interestingly, the pattern of differential read depth was absent when examining outlier replicates for the limma voom methods (Supplemental Figure 8).

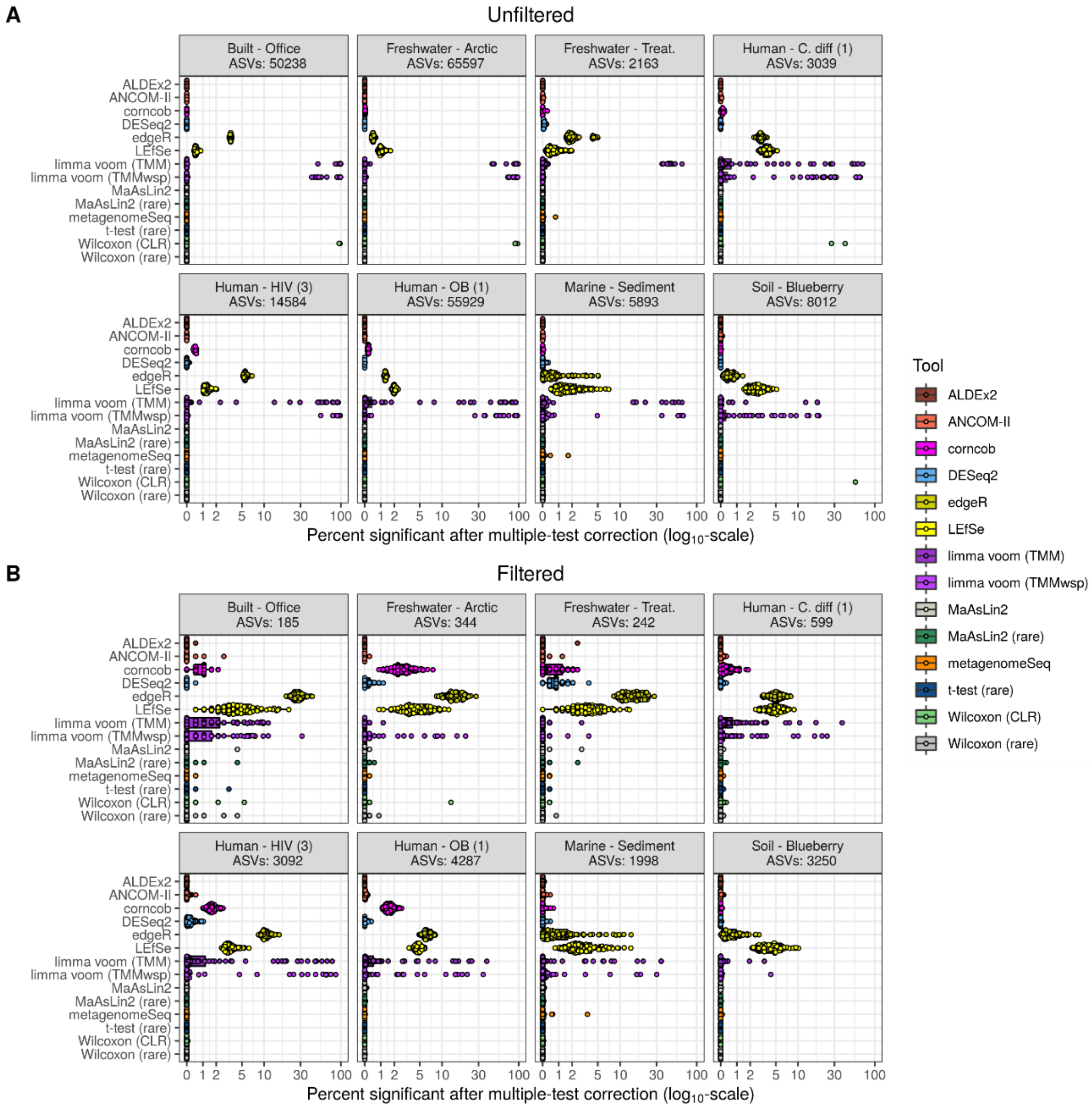


Figure 3.4: ***Distribution of false discovery rate simulation replicates for both unfiltered and filtered data****. The percentage of amplicon sequence variants that are significant after performing Benjamini–Hochberg correction of the p-values (using a cut-off of 0.05) are shown for each separate dataset and tool. Interquartile range (IQR) of boxplots represent the 25th and 75th percentiles while maxima and minima represent the maximum and*

minimum values outside 1.5 times the IQR. Notch in the middle of the boxplot represent the median. Note that the x-axis is on a pseudo-log10 scale. a Represents unfiltered datasets while b represents datasets filtered using a 10% prevalence requirement for each ASV. Datasets and tools were run 100 times while randomly assigning samples from the same environment and original groupings to one of two new randomly selected groupings. Differential abundance analysis was then performed on the two random groupings. Note that in the unfiltered datasets 100 replicates was only run 3 of the 8 datasets (Freshwater—Arctic, Soil—Blueberry, Human—OB (1)) with 100 ALDEx2 replications also being run in the Human - HIV (3) dataset. All other unfiltered datasets were run with 10 replicates due to computational limitations. Abbreviations: TMM, trimmed mean of M-values; TMMwsp, trimmed mean of M-values with singleton pairing; rare, rarefied; CLR, center-log-ratio. Source data are provided as a Source Data file.

3.4.4 - Tools Vary in how Consistently They Identify the Same Significant Genera Within Diarrhea Case-Control Datasets

Separate from the above analysis comparing consistency between tools on the same dataset, we next investigated whether certain tools provide more consistent signals across datasets of the same disease. This analysis focused on the genus-level across tools to help limit inter-study variation. We specifically focused on diarrhea as a phenotype, which has been shown to exhibit a strong effect on the microbiome and to be relatively reproducible across studies (Duvallet et al., 2017).

We acquired five datasets for this analysis representing the microbiome of individuals with diarrhea compared with individuals without diarrhea (see Methods). We ran all DA tools on each individual filtered dataset and restricted our analyses to the 218 genera found across all datasets. Like our ASV-level analyses, the tools substantially varied in terms of the number of significant genera identified. For instance, ALDEx2 identified a mean of 17.6 genera as significant in each dataset (SD: 17.4), while edgeR identified a mean of 46.0 significant genera (SD: 12.9). Tools that generally identify more genera as significant are accordingly more likely to identify genera as consistently significant compared with tools with fewer significant hits. Accordingly, inter-tool comparisons of the number of times each genus was identified as significant would not be informative.

Instead, we analyzed the observed distribution of the number of studies that each genus was identified as significant in compared with the expected distribution given random data. This approach enabled us to compare the tools based on how much more consistently each tool performed relative to its own random expectation. For instance, on average edgeR identified significant genera more consistently across studies compared with ALDEx2 (mean numbers of datasets that genera were found in across studies were 1.67 and 1.54 for edgeR and ALDEx2, respectively). However, this observation was simply driven by the increased number of significant genera identified by edgeR. Indeed, when compared with the random expectation, ALDEx2 displayed a 1.35-fold increase ($p < 0.001$) of consistency in calling significant genera in the observed data. In contrast, edgeR produced results that were only 1.10-fold more consistent compared with the random expectation ($p = 0.002$).

ALDEx2 and edgeR represent the extremes of how consistently tools identify the same genera as significant across studies, but there is a large range (Figure 3.5). Notably, all tools were significantly more consistent than the random expectation across these datasets (p < 0.05) (Table 2). In addition to ALDEx2, the other top performing approaches based on this evaluation included limma voom (TMM), both MaAsLin2 workflows, and ANCOM-II.

Table 3.2: ***Comparison of observed and expected consistency in differentially abundant genera across five diarrhea datasets.***

Tool	No. sig. genera	Max overlap	Mean exp.	Mean obs.	Fold diff.	p
ALDEx2	57	3	1.141	1.544	1.353	< 0.001
limma voom (TMM)	76	4	1.22	1.618	1.326	< 0.001
MaAsLin2 (rare)	74	3	1.204	1.595	1.325	< 0.001
ANCOM-II	15	3	1.033	1.333	1.29	< 0.001
MaAsLin2	79	3	1.215	1.557	1.281	< 0.001
Wilcoxon (rare)	88	4	1.269	1.625	1.281	< 0.001
metagenomeSeq	66	3	1.164	1.485	1.276	< 0.001
Wilcoxon (CLR)	82	3	1.22	1.549	1.27	< 0.001
limma voom (TMMwsp)	85	4	1.239	1.565	1.263	< 0.001
t-test (rare)	62	3	1.145	1.403	1.225	< 0.001
corncob	87	5	1.275	1.552	1.217	< 0.001
DESeq2	82	4	1.246	1.512	1.213	< 0.001

Tool	No. sig. genera	Max overlap	Mean exp.	Mean obs.	Fold diff.	p
LEfSe	119	5	1.408	1.613	1.146	< 0.001
edgeR	138	5	1.509	1.667	1.105	0.002

Column descriptions:

No. sig. genera: Number of genera significant in at least one dataset

Max overlap: Maximum number of datasets where a genus was called significant by this tool

Mean exp.: Mean number of datasets that each genus is expected to be significant in (of the genera that are significant at least once)

Mean obs.: Mean number of datasets that each genus was observed to be significant in (of the genera that are significant at least once)

Fold diff.: Fold difference of mean observed over mean expected number of times significant genera are found across multiple datasets

p: p-value based on one-tailed permutation test that used the "Mean obs." as the test statistic. Note that < 0.001 is indicated instead of exact values, because 0.001 was the minimum non-zero p-value we could estimate based on our permutation approach.

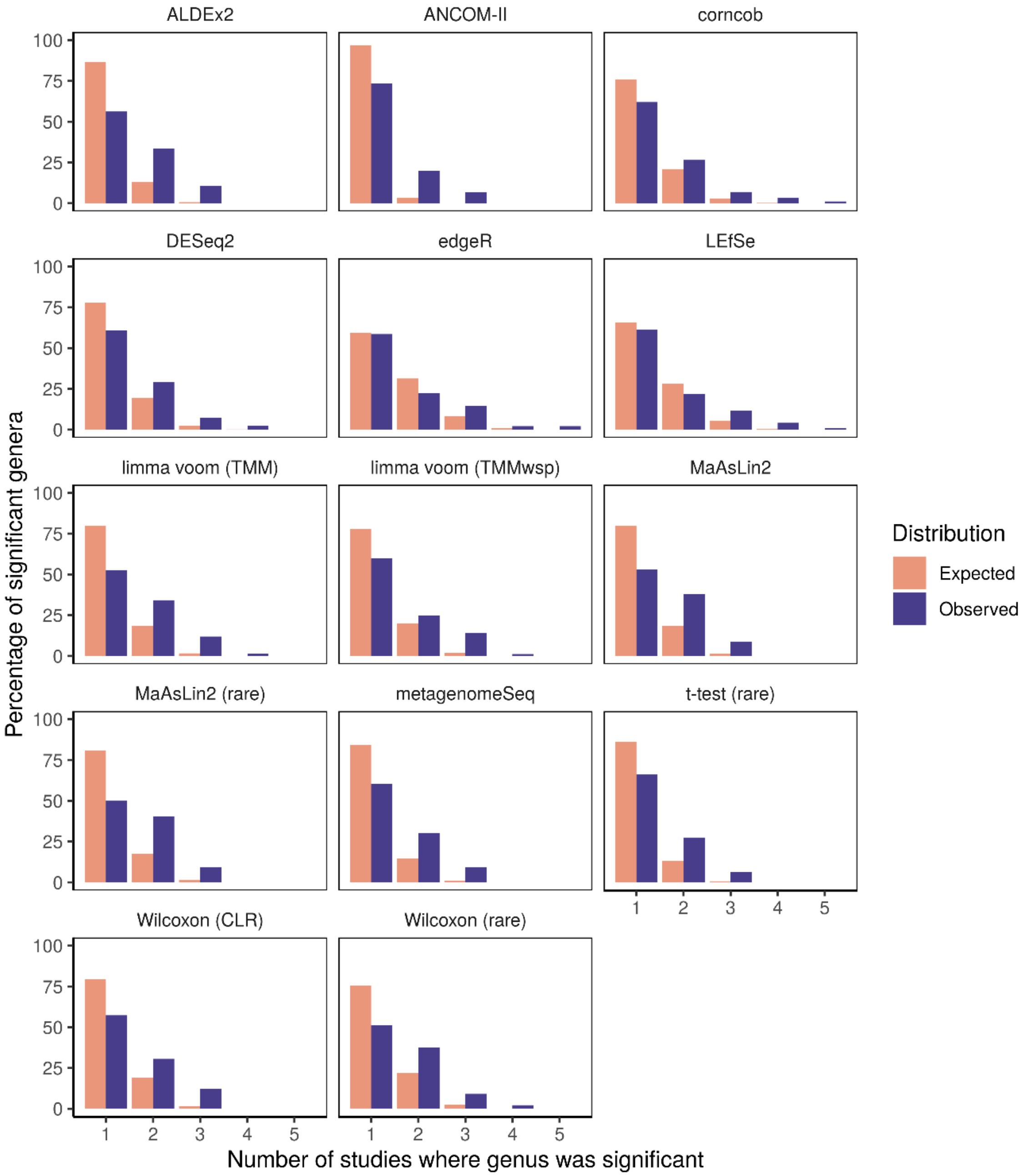


Figure 3.5: ***Observed consistency of significant genera across diarrhea datasets is higher than the random expectation overall.*** *These barplots illustrate the distributions of the number of studies for which each genus was identified as significant (excluding genera never found to be significant). The random expectation distribution is based on replicates of randomly selecting genera as significant and then computing the*

consistency across studies. Abbreviations: TMM, trimmed mean of M-values; TMMwsp, trimmed mean of M-values with singleton pairing; rare, rarefied; CLR, center-log-ratio. Source data are provided as a Source Data file.

We conducted a similar investigation across five obesity 16S rRNA gene datasets, which was more challenging to interpret due to the lower consistency in general (Supplemental Table 1). Specifically, most significant genera were called in only a single study and only MaAslin2 (both with non-rarefied and rarefied data), the t-test (rare) approach, ALDEx2, and the limma voom (TMMwsp) approach performed significantly better than expected by chance ($p < 0.05$). The MaAsLin2 (rare) approach produced by far the most consistent results based on these datasets (fold difference: 1.23; $p = 0.003$).

3.5 – Discussion

Herein we have compared the performance of commonly used DA tools on 16S rRNA gene datasets. While it might be argued that differences in tool outputs are expected given that they test different hypotheses, we believe this perspective ignores how these tools are used in practice. In particular, these tools are frequently used interchangeably in the microbiome literature. Accordingly, an improved understanding of the variation in DA method performance is crucial to properly interpret microbiome studies. We have illustrated here that these tools can produce substantially different results, which highlights that many biological interpretations based on microbiome data analysis are likely not robust to DA tool choice. These results might partially account for the common observation that significant microbial features reported in one dataset only marginally overlap with significant hits in similar datasets. However, it should be noted that this could also be due to numerous other biases that affect microbiome studies (Nearing et al., 2021). Our findings should serve as a cautionary tale for researchers conducting their own microbiome data analysis and reinforce the need to accurately report the findings of a representative set of different analysis options to ensure robust results are reported. Importantly, readers should not misinterpret our results to mean that 16S rRNA gene data is less reliable than other microbiome data types, such as shotgun metagenomics. We expect similar issues are affecting analyses with these data types, as they have similar or even higher levels of sparsity and inter-sample variation. Nonetheless, despite the high variation across DA tool results, we were able to characterize several consistent patterns produced by various tools that researchers should keep in mind when assessing both their own results and results from published work.

Two major groups of DA tools could be distinguished by how many significant ASVs they tended to identify. We found that limma voom, edgeR, Wilcoxon (CLR), and LEfSe output a high number of significant ASVs on average. In contrast, ALDEx2 and ANCOM-II tended to identify only a relatively small number of ASVs as significant. We hypothesize that these latter tools are more conservative and have higher precision, but with a concomitant probable loss in sensitivity. This hypothesis is related to our observation that significant ASVs identified by these two tools tended to also be identified by almost all other differential abundance methods, which we interpret to be ASVs that are more likely to be true positives. Furthermore, it was clear that in most, but not all, cases these two methods tended to identify the most discriminatory ASVs that were found by other tools. We believe that the lower number of ASVs identified by these approaches could be due to multiple reasons. ALDEx2's conservative nature is most likely due to its Monte Carlo Dirichlet sampling approach which down weights low abundance ASVs. On the other hand, ANCOM-II's conservative nature could be attributed to its large multiple testing burden. Furthermore, the variance in the number of features identified could also be attributed to pre-processing steps several tools use to remove potential ASVs for testing. This includes corncob that does not report significance values for ASVs only found in one group or ANCOM-II's ability to remove structural zeros. While it is unclear why some methods have much higher significance rates than other tools it could be due to several reasons. These include LEfSe's choice not to correct significance values for false discovery or Wilcoxon (CLR)'s inability to consider differences in sequencing depth between metadata groupings. It should be noted that in some cases authors haven chosen to apply FDR p-value correction to LEfSe output, when not including a subclass, however, this is not the default behavior of this tool (Rooks et al., 2014).

Given that ASVs commonly identified as significant using a wide range of approaches are likely more reliable, it is noteworthy that significant ASVs in the unfiltered data tended to be called by fewer tools. This was particularly evident for both limma voom approaches and the Wilcoxon (CLR) approach. Although it is possible that many of these significant ASVs are incorrectly missed by other tools, it is more likely that these tools are simply performing especially poorly on unfiltered data due to several reasons, such as data sparsity.

This issue with the limma voom approaches was also highlighted by high false positive rates on several unfiltered randomized datasets, which agrees with a past FDR assessment of this approach (Hawinkel et al., 2019). We believe that this issue is most likely driven by the inability for TMM normalization methods to deal with highly sparse datasets as filtering the data resulted in performance more in line with other DA methods (Robinson & Oshlack, 2010). It is important to acknowledge that our randomized approach for estimating FDR is not a perfect representation of real data; that is, real sample groupings will likely contain some systematic differences in microbial abundances—although the effect size may be very small—whereas our randomized datasets should have none. Accordingly, identifying only a few significant ASVs under this approach is not necessarily proof that a tool has a low FDR in practice. However, tools that identified many significant ASVs in the absence of distinguishing signals likely also have high FDR on real data.

Two additional particularly problematic tools based on this analysis were edgeR and LEfSe. The edgeR method has been previously found to exhibit a high FDR on several occasions (Hawinkel et al., 2019; Thorsen et al., 2016) Although metagenomeSeq also has been flagged as such (Hawinkel et al., 2019; Thorsen et al., 2016), that was not the case in our analysis. This agrees with a recent report that metagenomeSeq (using the zero-inflated log-normal approach, as we did) appropriately controlled the FDR, but exhibited low power (Lin & Peddada, 2020). There have been mixed results previously regarding whether ANCOM appropriately controls the FDR (Hawinkel et al., 2019; S. Weiss et al., 2017), but the results from our limited analysis suggest that this method is conservative and controls the FDR while sometimes potentially missing true positives, as evident in our discriminatory analysis.

Related to this point, we found that ANCOM-II performed better than average at identifying the same genera as significantly DA across five diarrhea-related datasets despite only identifying a mean of four genera as significant per dataset. Nonetheless, the ANCOM-II results were less consistent than ALDEx2, both MaAsLin2 workflows, and limma voom (TMM). The tools that produced the least consistent results across datasets (relative to the random expectation) included the t-test (rare) approach, LEfSe, and edgeR. The random expectation in this case was quite simplistic; it was generated based on the assumption that all genera were equally likely to be significant by chance. This assumption must be invalid to some degree simply because some genera are more prevalent than others across samples. Accordingly, it is surprising that the tools produced only marginally more consistent results than expected.

Although this cross-data consistency analysis was informative, it was interesting to note that not all environments and datasets are appropriate for this comparison. Specifically, we found that the consistency of significant genera across five datasets comparing obese and control individuals was no higher than expected by chance for most tools. This observation does not necessarily reflect that there are few consistent genera that differ between obese and non-obese individuals; it could instead simply reflect technical and/or biological factors that differ between the particular datasets we analyzed (Pollock et al., 2018).

We believe the above observations regarding DA tools are valuable, but many readers are likely primarily interested in hearing specific recommendations. Indeed, the need for standardized practices in microbiome analysis have recently become better appreciated (Hill, 2020). One goal of our work was to validate the recommendations of another recent DA method evaluation paper, which found that limma voom, corncob, and DESeq2 performed best overall of the tools they tested (Calgaro et al., 2020). Based on our results, we do not recommend these tools as the sole methods used for data analysis, and instead would suggest that researchers use more conservative methods such as ALDEx2 and ANCOM-II. Although these methods have lower statistical power (Calgaro et al., 2020; Hawinkel et al., 2019), we believe this is an acceptable trade-off given the higher cost of identifying false positives as differentially abundant. However, MaAsLin2 (particularly with rarefied data) could also be a reasonable choice for users looking for increased statistical power at the potential cost of more false positives. We can clearly recommend that users avoid using edgeR (a tool primarily intended for RNA-seq data) and LEfSe (without p-value correction) for conducting DA testing with 16S rRNA gene data. Users should also be aware that limma voom and the Wilcoxon (CLR) approaches may perform especially poorly on unfiltered data that is highly sparse. This is particularly true for the Wilcoxon (CLR) approach when read depths greatly differ between groups of interest.

More generally, we recommend that users employ several methods and focus on significant features identified by most tools, while keeping in mind the characteristics of the tools presented within this chapter. For example, authors may want to present identified taxonomic markers in categories based on the tool characteristics presented within this thesis or the number of tools that agree upon its identification. Importantly, applying multiple DA tools to the same dataset should be reported explicitly. Clearly this approach would make results more difficult to biologically interpret, but it would provide a clearer perspective on which differentially abundant features are robust to reasonable changes in the analysis.

A common counterargument to using consensus approaches with DA tools is that there is no assurance that the intersection of the tool outputs is more reliable; it is possible that the tools are simply picking up the same noise as significant. Although we think this is unlikely, in any case running multiple DA tools is still important to give context to reporting significant features. For example, researchers might be using a tool that produces highly non-overlapping sets of significant features compared with other DA approaches. Even if the researchers are confident in their approach, these discrepancies should be made clear when the results are summarized. This is crucial for providing accurate insight into how robust specific findings are expected to be across independent studies, which often use different DA approaches. Similarly, if researchers are most interested in determining if signals from a specific study are reproducible, then they should ensure that they use the same DA approach to help make their results more comparable.

How and whether to conduct independent filtering of data prior to conducting DA tests are other important open questions regarding microbiome data analysis (Schloss, 2020). Although statistical arguments regarding the validity of independent filtering are beyond the scope of this work, intuitively it is reasonable to exclude features found in only a small number of samples (regardless of which groups those samples are in). The basic reason for this is that otherwise the burden of multiple-test correction becomes so great as to nearly prohibit identifying any differentially abundant features. Despite this drawback, many tools identified large numbers of significant ASVs in the unfiltered data. However, these significant ASVs tended to be more tool-specific in the unfiltered data and there was much more variation in the percentage of significant ASVs across tools. Accordingly, we would suggest performing prevalence filtering (e.g., at 10%) of features prior to DA testing, although we acknowledge that more work is needed to estimate an optimal cut-off rather than just arbitrarily selecting one (McMurdie & Holmes, 2014).

Another common question is whether DA tools that require input data to be rarefied should be avoided. It is possible that the question of whether to rarefy data has received disproportionate attention in the microbiome field: there are numerous other factors affecting an analysis pipeline that likely affect results more. Indeed, tools that took in rarefied data in our analyses did not perform substantially worse than other methods on average. More specifically, the most consistent inter-tool methods, ANCOM-II and ALDEx2, are based on non-rarefied data, but MaAsLin2 based on rarefied data produced the most consistent results across datasets of the same phenotype. Accordingly, we cannot definitively conclude that DA tools that require input data to be rarefied are less reliable in general. It should be noted that we are referring only to rarefying in the context of DA testing: whether rarefying is advisable for other analyses, such as prior to computing diversity metrics, is beyond the scope of this work (McMurdie & Holmes, 2014; S. Weiss et al., 2017).

Others have investigated the above questions by applying simulations to various DA methods, which has yielded valuable insights (Calgaro et al., 2020; Hawinkel et al., 2019; Thorsen et al., 2016; S. Weiss et al., 2017). However, we believe this does not provide the full picture of how these DA tools perform. This is because it has been highlighted that in many scenarios simulations can led to circular arguments where tools that are designed around specific parameters perform favorably on simulations using those parameters (Hawinkel et al., 2019). Without better knowledge of the range of data structures that microbiome sequencing can result in these types of simulation analyses can be difficult to interpret. As such we believe that it was important to test these methods on a wide range of different real-world datasets in order to gain an understanding of how they differed from one another. By doing so we have highlighted the issues of using these tools interchangeably within the literature. Indeed, the question “which taxa significantly differ in relative abundance between sample groupings?” may be too simple and need further parameterization before it can be answered. This includes information such as what type of abundance the authors are comparing and the tools they plan to use within their analysis. Unfortunately, the variation across tools implies that biological interpretations based on these questions will often drastically differ depending on which DA tool is considered.

In conclusion, the high variation in the output of DA tools across numerous 16S rRNA gene sequencing datasets highlights an alarming reproducibility crisis facing microbiome researchers. While we cannot make a direct simple recommendation of a specific tool based on our analysis, we have highlighted several issues that authors should be aware of while interpreting DA results. This includes that several tools have inappropriately high false discovery rates such as edgeR and LEfSe and as such should be avoided when possible. It is also clear from our analysis that some tools designed for RNA-seq such as limma voom methods cannot deal with the much higher sparsity of microbiome data without including a data filtration step. We have also highlighted that these tools can significantly differ in the number of ASVs that they identify as being significantly different and that some tools are more consistent across datasets than others. Overall, we recommend that authors use the same tools when comparing results between specific studies and otherwise use a consensus approach based on several DA tools to help ensure results are robust to DA

Chapter 4 – Assessing the Variation Within the Oral Microbiome of Healthy Adults

This chapter is a near reproduction of the paper of the same named published in the journal mSphere (Nearing, DeClercq, Van Limbergen, & Langille, 2020). I was first author on this work and was the primary contributor to the data analysis, figure creation, and writing. I also participated in study design along with my co-authors Vanessa DeClercq, Johan Van Limbergen, and Morgan G.I. Langille.

This paper was published under a Creative Commons Attribution 4.0 International License and allows for the reproduction and adaptation with attribution (https://creativecommons.org/licenses/by/4.0/). The original acknowledgements of this paper can also be found within its original publication. All additional files and supplemental information referred to in this chapter are freely available as part of the original publication on the mSphere journal website and can be found at the following link: https://doi.org/10.1128/mSphere.00451-20.

4.1 – Abstract

Over 1000 different species of microbes have been found to live within the human oral cavity where they play important roles in maintaining both oral and systemic health. Several studies have identified the core members of this microbial community, however, the factors that determine oral microbiome composition are not well understood. In this study we exam the salivary oral microbiome of 1049 Atlantic Canadians using 16S rRNA gene sequencing to determine which dietary, lifestyle, and anthropometric features play a role in shaping microbial community composition. Features that were identified as being significantly associated with overall composition were then additionally examined for genera, amplicon sequence variants, and predicted pathway abundances that were associated with these features. Several associations were replicated in an additional secondary validation dataset. Overall, we found that several anthropometric measurements including waist-hip ratio (WHR), height, and fat free mass, as well as age and sex, were associated with overall oral microbiome structure in both our exploratory and validation cohorts. We were unable to validate any dietary impacts on overall taxonomic oral microbiome composition but did find evidence to suggest potential contributions from factors such as the number of vegetable and refined grain servings an individual consumes. Interestingly, each one of these factors on their own was associated with only minor shifts in the overall taxonomic composition of the oral microbiome suggesting that future biomarker identification for several diseases associated with the oral microbiome may be undertaken without the worry of confounding factors obscuring biological signal.

4.1.1 – Importance

The human oral cavity is inhabited by a diverse community of microbes known as the human oral microbiome. These microbes play a role in maintaining both oral and systemic health and as such have been proposed to be useful biomarkers of disease. However, to identify these biomarkers, we first need to determine the composition and variation of the healthy oral microbiome. Within this report we investigate the oral microbiome of 1049 healthy individuals to determine which genera and amplicon sequence variants are commonly found between individual oral microbiomes. We then further investigate how lifestyle, anthropometric, and dietary choices impact overall microbiome composition. Interestingly, the results from this investigation showed that while many features were significantly associated with oral microbiome composition no single biological factor explained a variation larger than 2%. These results indicate that future work on biomarker detection may be encouraged by the lack of strong confounding factors.

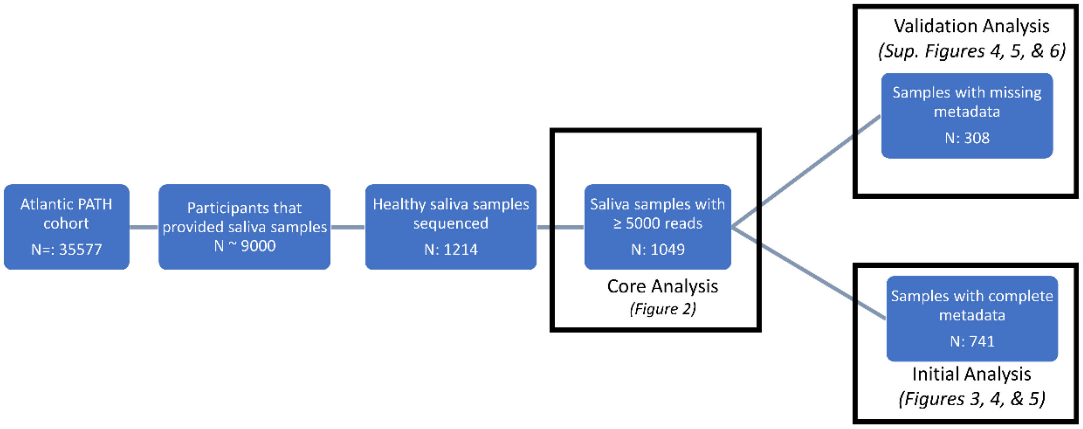
4.2 – Introduction

The human oral cavity is colonized by numerous bacteria, fungi, viruses and archaea that make a rich microbial community known as the oral microbiome. This microbial community is one of the most diverse sites of microbial growth within the human body; only the colon houses a more diverse consortia of microbes (Huttenhower et al., 2012). To date over 1000 different bacterial species have been found to colonize the oral cavity (Dewhirst et al., 2010) on various surfaces including the tongue, teeth, cheek, and gingivae (Huttenhower et al., 2012). These communities of microbes are responsible for various functions that can both maintain and deplete oral health. For example, the presence of biofilms containing bacterial species such as *Streptococcus mutans* and other aciduric bacteria can damage hard dental surfaces and lead to dental caries (Takahashi & Nyvad, 2010; Wade, 2013). Furthermore, the oral microbiome is known to play a role in a myriad of other oral diseases including oral cancer (Karpiński, 2019), periodontitis (Kumar et al., 2003; Socransky, Haffajee, Cugini, Smith, & Kent Jr., 1998), and gingivitis (Kolenbrander et al., 2006; Murray, Prakobphol, Lee, Hoover, & Fisher, 1992). In addition to well-established associations between oral and cardiac health (Shungin et al., 2019), recent work has also begun to show that the oral microbiome may play a role in the health of other distal sites within the human body. For example, the enrichment of both *Porphyromonas gingivalis* and *Aggregatibacter actinomycetemcomitans* has been associated with a higher risk of pancreatic cancer (X. Fan, Alekseyenko, et al., 2018). Furthermore, several oral bacteria including *Streptococcus* and *Prevotella* species have been found to be in higher abundance among individuals with colorectal cancer (Flemer et al., 2017). Other than these two cancers a number of other distal diseases have been associated with oral microbiome composition including, prostate cancer (Porter, Shrestha, Peiffer, & Sfanos, 2018) and inflammatory bowel disease (Said et al., 2013).

Due to the associations between these diseases and the oral microbiome, its composition has been proposed as a useful biomarker for human health and disease. With this in mind, various studies have attempted to identify core members of the "healthy" oral microbiome (De Filippis et al., 2014; Huttenhower et al., 2012; Nasidze, Li, Quinque, Tang, & Stoneking, 2009; Takeshita et al., 2016; Zaura, Keijser, Huse, & Crielaard, 2009) to help aid in disease detection. These studies have uncovered that, at the genus level, the oral microbiome remains relatively stable between individuals (Huttenhower et al., 2012; Zaura et al., 2009) and across multiple geographic locations (J. Li et al., 2014; Nasidze et al., 2009), but at deeper taxonomic resolutions, it can be variable. This may indicate that other factors such as dietary, anthropometric, or sociodemographic factors could play a role in shaping the oral microbiome (Belstrøm et al., 2014; De Filippis et al., 2014; Mason et al., 2013; Peters et al., 2018; Renson et al., 2019; Takeshita et al., 2016). Various studies have focused on individual factors that may cause shifts in the oral microbiome such as ethnicity (Huttenhower et al., 2012; Mason et al., 2013), alcohol consumption (X. Fan, Peters, et al., 2018), smoking (J. Wu et al., 2016), obesity (Yujia Wu, Chi, Zhang, Chen, & Deng, 2018; Y. Yang, Cai, Zheng, et al., 2019), and dietary patterns (Hansen et al., 2018). However, to date only a small number of studies have looked at the relative contributions of each of these factors to oral microbiome variability in a single cohort. Takeshita et al., examined the oral microbiome of 2,343 adults living in Japan using 16S rRNA gene sequencing and identified that higher abundances of *Prevotella*, and *Veillonella* species were associated with old age, higher body mass index (BMI), and poor overall oral health (Takeshita et al., 2016). Another study by Renson et al., in adults living in New York City also found that variation in taxonomic abundances could be linked to marital status, ethnicity, education, and age (Renson et al., 2019). Further, work by Belstrøm et al., examined the oral microbiome of 292 Danish individuals with low levels of dental caries and periodontitis using microarrays and found that while socioeconomic status impacted oral microbiome profiles, diet, BMI, age, and sex had no statistical impact on microbial abundances (Belstrøm et al., 2014). This study, however, was only able to identify the abundances of taxa that had a corresponding probe which, could explain its disagreement with other work. Overall, these studies have indicated that both biological differences such as sex and BMI as well as lifestyle and sociodemographic differences can impact oral microbiome composition.

While these studies have shed light on the variation of the oral microbiome, it is currently unclear to what extent these factors play a role in shaping the oral microbiome of an individual. Without identifying the effect size of each of these factors relative to one another, it is difficult to identify the correct variables that should be controlled for in case-control studies of the oral microbiome. Furthermore, each of these studies have identified different taxa that are impacted by various factors such as sex, BMI, and age. This could be due to many factors, including systemic bias introduced via the use of different sequencing or bioinformatic protocols/tools (Pollock et al., 2018) or differences in the studied cohorts. Therefore, the identification of microbes that are impacted by factors such as sex, BMI, or diet could help identify potential interactions between the oral microbiome, health, and disease.

Herein, we report the variation within the healthy oral microbiome by examining 741 samples from non-smoking healthy individuals living within the Atlantic Provinces of Canada. We then validated our results on a smaller subset of individuals (n=308) from the same cohort (Figure 4.1). The bacterial oral microbiome composition of these individuals was investigated through 16S rRNA gene sequencing from saliva samples provided by each participant. Compositions were then compared with 41 different variables including anthropometric, dietary, and sociodemographic factors (Table 4.1). In this investigation, we determined which of these factors play a role in shaping the oral microbiome and to what extent these factors can explain the overall oral microbiome composition.



Atlantic PATH cohort
N=: 35577
Participants that provided saliva samples
N ~ 9000
Healthy saliva samples sequenced
N: 1214
Saliva samples with ≥ 5000 reads
N: 1049
Core Analysis
(Figure 2)
Validation Analysis
(Sup. Figures 4, 5, & 6)
Samples with missing metadata
N: 308
Samples with complete metadata
N: 741
Initial Analysis
(Figures 3, 4, & 5)

Figure 4.1: ***Flowchart of sample selection from the Atlantic Partnership for Tomorrow's Health cohort.*** *A total of 35,577 individuals participated in the Atlantic Partnership for Tomorrow's Health cohort, and ~9,000 individuals provided saliva samples. Of those, a subset of 1,214 saliva samples from healthy individuals underwent*

16S rRNA gene sequencing. Samples below 5,000 reads were filtered out, and two data sets were created for discovery and validation analysis.

Table 4.1: ***Cohort characteristic and variables analyzed for oral microbiome composition***

	Overall
Number of participants	1214
Rural/Urban (%)	
Urban	1050 (86.5)
Rural	126 (10.4)
NA	38 (3.1)
Province (%)	
New Brunswick	124 (10.2)
Nova Scotia	1070 (88.1)
Prince Edward Island	16 (1.3)
NA	Data repressed
Economic Region	
Annapolis Valley	52
Cape Breton	142
Edmundston – Woodstock	Data repressed
Fredericton – Oromocto	44
Halifax	773
Moncton – Richibucto	32
North Shore	41
Prince Edward Island	16
Saint John – St., Stephen	45
Southern Shore	28
Sex (%)	
Female	846 (69.7)
Male	368 (30.3)
Body mass index (mean (SD))	
Male	27.83 (3.62)
Female	27.06 (4.89)
Waist size (cm) (mean (SD))	
Male	98.10 (10.51)
Female	87.86 (12.45)
Hip size (cm) (mean (SD))	
Male	103.84 (7.15)

Female	104.50 (10.29)
Waist-hip ratio (mean (SD))	
Male	0.94 (0.06)
Female	0.84 (0.07)
Height (cm) (mean (SD))	
Male	176.64 (6.72)
Female	162.90 (6.07)
Weight (kg) (mean (SD))	
Male	86.92 (12.84)
Female	71.81 (13.48)
Age (years) (mean (SD))	55.39 (7.80)
Fat mass (kg) (mean (SD))	
Male	21.65 (7.16)
Female	26.79 (10.01)
Fat free mass (kg) (mean (SD))	
Male	65.21 (7.42)
Female	45.00 (4.77)
Body fat percentage (mean (SD))	
Male	24.46 (5.32)
Female	36.20 (7.24)
Vegetable servings (mean (SD))	2.56 (1.98)
Fruit servings (mean (SD))	2.00 (1.45)
Juice servings (mean (SD))	0.69 (0.95)
Whole grain servings (mean (SD))	2.11 (1.43)
Refined grain servings (mean (SD))	0.67 (0.86)
Milk product servings (mean (SD))	2.04 (1.29)
Egg servings per week (mean (SD))	3.25 (2.68)
Meat/poultry servings (mean (SD))	1.53 (1.35)
Fish servings (mean (SD))	0.51 (0.67)
Tofu servings (mean (SD))	0.04 (0.18)
Bean servings (mean (SD))	0.36 (0.55)
Nut/seed servings (mean (SD))	0.69 (0.68)
Dessert Frequency (%)	
Never	109 (9.0)
Less than once a month	153 (12.6)
About once a month	228 (18.8)
2 to 3 times a month	173 (14.3)
Once a week	85 (7.0)
2 to 3 times a week	115 (9.5)
4 to 5 times a week	58 (4.8)
6 to 7 times a week	169 (13.9)
NA	124 (10.2)
Avoidance of particular foods (%)	

Never	853 (70.3)
Often	11 (0.9)
Prefer not to answer	15 (1.2)
Rarely	163 (13.4)
Sometimes	52 (4.3)
NA	120 (9.9)
Oil on bread (%)	
Butter	371 (30.6)
Low fat margarine	272 (22.4)
Full fat margarine	300 (24.7)
None	109 (9.0)
Olive oil	36 (3.0)
NA	126 (10.4)
Artificial sweeteners (%)	
Almost never	976 (80.4)
About 1/4 of the time	24 (2.0)
About 1/2 of the time	16 (1.3)
About 3/4 of the time	12 (1.0)
Almost always or always	53 (4.4)
NA	133 (11.0)
Non-diet soda frequency (%)	
Zero days a week	432 (35.6)
One to three days per month	459 (37.8)
One to five days a week	167 (13.8)
Six to seven days a week	27 (2.2)
NA	129 (10.6)
Diet sugar drink frequency (%)	
Zero days a week	513 (42.3)
One to three days per month	356 (29.3)
One to five days a week	156 (12.9)
Six to seven days a week	57 (4.7)
NA	132 (10.9)
Soy/fish usage (%)	
Never at the table	424 (34.9)
Rarely at the table	441 (36.3)
Sometimes at the table	217 (17.9)
At most meals of eating occasions	9 (0.7)
NA	123 (10.1)
Salt seasoning (%)	
Never	368 (30.3)

Rarely	347 (28.6)
Sometimes	219 (18.0)
Most meals	157 (12.9)
NA	123 (10.1)
Fast food frequency (%)	
Never	149 (12.3)
Less than once per month	384 (31.6)
One - three times per month	366 (30.1)
One - six per week	191 (15.7)
One or more times per day	Data Repressed
NA	122 (10.0)
Alcohol Frequency (%)	
Never	61 (5.0)
Less than once a month	192 (15.8)
About once a month	70 (5.8)
2 to 3 times a month	171 (14.1)
Once a week	170 (14.0)
2 to 3 times a week	259 (21.3)
4 to 5 times a week	127 (10.5)
6 to 7 times a week	112 (9.2)
NA	52 (4.3)
Education level (%)	
Highschool or below	208 (17.1)
Non-bachelors post secondary	425 (35.0)
Bachelors	334 (27.5)
Graduate	242 (19.9)
NA	Data Repressed
Income (%)	
Below $25 000 CAD	41 (3.4)
$25 000 - $49 999 CAD	157 (12.9)
$50 000 - $74 999 CAD	244 (20.1)
$75 000 - $99 999 CAD	244 (20.1)
$100 000 - $149 999 CAD	291 (24.0)
Greater than $150 000 CAD	179 (14.7)
NA	58 (4.8)
Sleeping trouble frequency (%)	
None	104 (8.6)
A little of the time	411 (33.9)
Some of the time	507 (41.8)
Most of the time	161 (13.3)

All the time	25 (2.1)
NA	Data Repressed
Last dental visit (%)	
Less than 6 months ago	851 (70.1)
6 months to less than 1 year ago	221 (18.2)
1 year to less than 2 years ago	56 (4.6)
2 years to less than 3 years ago	17 (1.4)
3 or more years ago	24 (2.0)
NA	45 (3.7)
Sleeping light exposure (%)	
Virtually no light	561 (46.2)
Some light	613 (50.5)
A lot of light	36 (3.0)
NA	Data Repressed
DNA extraction batch (%)	
Extraction.1	85 (7.0)
Extraction.10	66 (5.4)
Extraction.11	80 (6.6)
Extraction.12	78 (6.4)
Extraction.13	85 (7.0)
Extraction.14	57 (4.7)
Extraction.15	79 (6.5)
Extraction.16	0 (0.0)
Extraction.17	67 (5.5)
Extraction.2	85 (7.0)
Extraction.3	81 (6.7)
Extraction.4	68 (5.6)
Extraction.5	85 (7.0)
Extraction.6	92 (7.6)
Extraction.7	85 (7.0)
Extraction.8	60 (4.9)
Extraction.9	61 (5.0)

4.3 – Materials and Methods

4.3.1 - Study Design and Population

The current study includes the analysis of saliva samples from the Atlantic Partnership for Tomorrow's Health (PATH) study. Atlantic PATH is part of the Canadian Partnership for Tomorrow's Health (CanPath) project, a pan-Canadian prospective cohort study examining the influence of environmental, genetic, and lifestyle factors on the development of chronic disease (Sweeney et al., 2017). The applicable provincial and regional ethics boards approved the study protocol and all participants provided written informed consent prior to participation. The primary inclusion criteria were that participants were aged 30-74 years at time of recruitment and a resident in one of the Atlantic Canadian provinces (Nova Scotia, New Brunswick, Prince Edward Island, and Newfoundland and Labrador). Recruitment and baseline data for all participating regions was collected between 2000 and 2019. Details on participant recruitment and a descriptive cohort profile have been published elsewhere (Sweeney et al., 2017). The questionnaire included sociodemographic information, health information, behaviours, environmental factors, and self-reported anthropometric information. Participants also had anthropometric measures (height, weight, waist and hip circumferences, body composition, blood pressure, grip strength, and resting heart rate) and biological samples (blood, urine, saliva, and toenails) collected. Approximately 9000 participants in the Atlantic PATH cohort provided a saliva sample. Participants were instructed to refrain from eating, smoking, or chew gum for at least 30 minutes prior to oral specimen collection. Oral saliva specimens were collected during normal clinic hours 9:00 am – 7:00 pm after completion of the approximately one-hour interview and registration process. Oral samples (3 ml) were collected in sterile 50 ml conical tubes after rinsing with water. Samples were stored at 4°C and batch shipped on ice to the central processing facility at the QEII Health Sciences Centre in Halifax, Nova Scotia. Samples were processed within 24 hours of collection, aliquoted into cryovials and stored at -80ºC until analysis.

The current analysis includes a total of 1214 saliva samples from healthy Atlantic Canadians living within the provinces of Nova Scotia, New Brunswick, and Prince Edward Island. Based on self-reported data, participants were defined as healthy if they had not been diagnosed with any of the following conditions: hypertension, myocardial infarction, stroke, asthma, chronic obstructive pulmonary disease, major depression, diabetes, inflammatory bowel disease, irritable bowel syndrome, chronic bronchitis, emphysema, liver cirrhosis, chronic hepatitis, dermatologic disease (psoriasis and eczema), multiple sclerosis, arthritis, lupus, osteoporosis, and cancer. A total of 165 of these samples were removed due to insufficient sequencing depth and of the remaining 1049, an additional 308 were removed due to incomplete answering of the 41 variables examined in this study. These 308 samples that were removed were then used in validation analysis (details below) to confirm findings within the larger 741 participant cohort.

4.3.2 - Socio-Demographic, Lifestyle and Anthropometric Variables

Questionnaires were used to collect socio-demographic and lifestyle variables. Self-reported variables included age, sex, education level, household income, rural/urban, province, dental visits, sleep patterns, alcohol consumption, smoking status, and dietary variables such as food avoidance, the use of specific types of fat/oil, artificial sweetener usage, the frequency of dessert, soda drinks, soy/fish sauce, salt seasoning, and, fast food, as well as servings of vegetables, fruit, juice, whole grains, refined grains, dairy products, eggs, fish, tofu, beans, and nuts/seeds. Anthropometric measures were collected by trained personnel in assessment centres. Waist and hip circumferences were measured using Lufin steel tape. Height was measured by a Seca stadiometer. Height and weight measures were used to calculate body mass index (BMI; weight in kilograms divided by height in meters squared; kg/m^2). Body weight, fat mass, and fat-free mass were measured using the Tanita bioelectrical impedance device (Tanita BC-418, Tanita Corporation of America Inc., Arlington Heights, Illinois). Table 4.1 lists all variables that were used for analysis.

4.3.3 - Oral Microbiome 16S rRNA Sequencing

Frozen saliva samples were thawed at room temperature and aliquoted into 96 well plates. DNA from samples were then extracted using a QIAamp 96 PowerFecal QIAcube HT Kit following the manufacturer's instructions using a TissueLyser II and the addition of Proteinase K. Sequencing of the 16S rRNA gene was performed by the Integrated Microbiome Resource at Dalhousie University. The V4-V5 region was amplified from extracted DNA in a PCR using 16S rRNA gene V4-V5 fusion primers (515FB – 926R) (Comeau et al., 2017) and high-fidelity Phusion polymerase. Amplified DNA concentrations were then normalised and pooled together to be sequenced on an Illumina MiSeq. Sequencing of samples was conducted over 6 Illumina MiSeq runs producing 300 base pair paired-end reads.

4.3.4 - 16S rRNA Gene Sequence Processing

Primers were removed from paired-end 300 base pair sequences using cut adapt (M. Martin, 2011). Primer free reads were then stitched together using the QIIME2 (v. QIIME2-2018.8) (Bolyen et al., 2019) VSEARCH (Rognes et al., 2016) join-pairs plugin. Stitched reads were then filtered using the QIIME2 plugin q-score-joined using the default parameters. Quality filtered reads were then input into the QIIME2 plugin Deblur (Amir et al., 2017) to produce amplicon sequence variants (ASV). A trim length of 360 base pairs and a minimum number of reads required to pass filtering was set to 1. Amplicon sequence variants that were found in an abundance of less than 0.1% of the mean sample depth (18) were then removed from analysis. This is to keep inline with the approximate bleed-through rate on an Illumina MiSeq sequencer. After filtering a total of 13248 ASVs were recovered. Representative sequences were then placed into the Greengenes 13_8 99%(Andersen, DeSantis, Liu, & Knight, 2008) reference 16S rRNA tree using the QIIME2 (2019.7) fragment-insertion SEPP (Janssen et al., 2018; MIRARAB et al., 2011) plugin. Rarefaction curves were then generated using the QIIME2 alpha-rarefaction plugin and a suitable rarefaction depth of 5000 was chosen for diversity analysis based on when the number of newly discovered ASVs came to a plateau (Supplemental Figure 1). Representative sequences were then assigned taxonomy using a custom trained V4-V5 16S rRNA naive Bayesian QIIME2 classifier (Bokulich et al., 2018) trained on the 99% Silva V132 database (Pruesse et al., 2007).

4.3.5 - Oral Microbiome Composition Analysis

Taxonomic composition tables were generated using the QIIME2 taxa plugin and collapsed at the genus level. All samples over 5000 reads in depth (1049) were subsampled to a depth of 5000 reads each and taxa that contributed less than a mean relative abundance of 1% were grouped together under an “Other” category. The composition stacked bar chart was then generated in R using ggplot2 (Wickham, 2009) and the x-axis was ordered based on the PC1 weighted UniFrac coordinates of each sample.

4.3.6 - Core Oral Microbiome Analysis

Taxonomic tables subsampled previously at 5000 reads were collapsed at the genus and ASV level using QIIME2. To examine the mean relative abundance explained by genera/ASVs at different prevalence levels we remove genera/ASVs that were not present in a varying number of samples (5-99%). After removal of these genera/ASVs the remaining total mean relative abundance of all genera/ASVs that passed the filtering parameter was calculated.

4.3.7 - Oral Microbiome Alpha Diversity Analysis

Alpha diversity metrics were generated using QIIME2 (v2019.7) and the previously generated tree containing both representative sequences and reference sequences. All samples were subsampled to a depth of 5000 reads. Association between four different alpha diversity metrics (Faith’s Phylogenetic Diversity, Shannon, Evenness, Number of ASVs) were then tested using general linear models while controlling for DNA extraction. A base model containing only DNA extraction as a covariate and a testing modelling containing DNA extraction and the covariate of interest were then compared using an ANOVA and p-values were recorded. Recorded p-values were then corrected for false discovery (Benjamini and Hochberg (Benjamini & Hochberg, 1995) with a chosen alpha of $q < 0.1$.

4.3.8 - Oral Microbiome Beta Diversity Analysis

Beta diversity metrics were generated using QIIME2 and the previously generated phylogeny. All sequences were subsampled to a depth of 5000 reads based on the plateauing stage of rarefaction plots (Supplemental Figure 1). Association between two different beta diversity metrics (weighted UniFrac distance, Bray Curtis dissimilarity) were then tested using a PERMANOVA (adonis2 function in Vegan (Dixon, 2003)) while controlling for DNA extraction. Marginal p values were then corrected for false discovery (Benjamini and Hochberg) and an alpha value of $q < 0.1$ was chosen. Significant features from univariate analysis were then included in a single multivariate model that underwent backwards covariate selection, where each co-variation with the highest p-value was removed from the model until all features were found to be significant. Additional testing using adonis2 on fat free mass and height were done while controlling for both sex and DNA extraction. Finally, overall relationships between taxa, metadata, and samples were visualized with a redundancy analysis triplot. This plot was constructed using the rda function within the vegan R package. Within this function non-rarified center-log-ratio genera count tables were filtered for features with at least 10% prevalence and then used as the response variable within the RDA model. Each feature previously associated with either weighted UniFrac or Bray-Curtis dissimilarity profiles were input as explanatory variables within the RDA model. The significance of the RDA model was checked using the function anova.cca within the vegan R package. Finally, visualization of the resulting RDA model was done with the R package ggord (Beck, 2017) using symmetrical species and site scaling.

4.3.9 - Differential Abundance Analysis

Differential abundance analysis was conducted using the Corncob (B. D. Martin et al., 2020) (v 0.1.0) and Phyloseq (McMurdie & Holmes, 2013) R packages. A genus level taxonomic table was generated using QIIME2 (2019.7) and genera that were not found in at least 10% of samples were removed. The fifteen covariates that were found to be significantly associated to either weighted UniFrac or Bray Curtis dissimilarities were chosen for testing. Testing of each covariate was done using the "differentialtest" function in the Corncob package while controlling for differences in DNA extraction and differential variability across DNA extraction and the covariate of interest. Heatmaps were then constructed containing any genera/ASV that were significantly associated to at least one of the covariates that were tested.

4.3.10 - Prediction of Microbial Pathway Abundances Using Picrust2

Amplicon sequence variant abundance tables were rarified at a depth of 5000 reads and input into the picrust2_pipeline.py script to generate predicted microbial pathway abundances. MetaCyc pathway identifiers were then mapped to their respective pathway names using the picrust2 add_descripition.py script. Differential abundance analysis of predicted pathway abundances using the R package Corncob was done in the same manner as previously explained for taxonomic data. Only features that were found to be significantly associated with weighted UniFrac or Bray-Curtis dissimilarities were tested. DNA extraction round and differential variability within the tested feature were controlled for as previously described and p-values were corrected using Benjamini Hochberg false discovery correction (Benjamini & Hochberg, 1995). An alpha value of

0.05 was chosen for corrected p values and pathways with an effect size lower than |0.05| log odds were filtered out.

4.3.11 - Validation Analysis

A total of 308 samples had not completely answered all 41 metadata variables of interest and therefore were removed from the original analysis. This smaller cohort was used to test our previous results by removing samples during testing of each covariate that had not answered that question on the questionnaire. Both beta diversity analysis and differential abundance analysis on taxa and pathways were carried out in the same manner as previously explained. Both beta diversity metrics using PERMANOVA tests and differential abundance analysis using corncob were done in a univariate fashion while also controlling for DNA extraction batch. Furthermore, only features/taxa that were originally identified as being significantly associated to oral microbiome composition in our initial cohort were tested. As there was previous evidence that these features were associated with that covariate/metric, p-values were not corrected for false discovery but an alpha value of 0.05 was chosen. Furthermore, to keep with the original pathway analysis only pathways that had an effect size of |0.05| log odds in the discovery cohort were tested for differential abundance in the validation cohort.

4.3.12 - Random Forest Model Training and Validation

Non-rarified ASV abundances were converted in relative abundances and used to train random forest classification and regression models for each feature that was significantly associated with either weighted UniFrac or Bray-Curtis dissimilarities. An optimal mtry parameter was chosen using three fold repeated cross validation within the caret R package (Kuhn, 2008). Trained models for each feature were then validated on the hold-out validation cohort to determine model performance. Model performance for classification was visualized using the PRROC R package (Grau, Grosse, & Keilwagen, 2015) and r^2 performance of regression models was determined using the postResample function within the caret R package.

4.4 – Results

4.4.1 - The Healthy Oral Microbiome is Stable at the Genus Level but Variable at Higher Resolutions

We examined the oral microbiome composition of the overall cohort containing 1049 healthy individuals (Figure 4.1) from Atlantic Canada to understand how anthropometric, socio-demographic, and dietary choices could alter oral microbiome composition. We found that 16 genera were found to have a mean relative abundance greater than 1% (Fig 4.2A) with *Veillonella* having the largest mean contribution (21.49% +- 0.38%) followed by *Neisseria* (13.04% +- 0.40%), *Streptococcus* (11.86% +- 0.26%), and *Prevotella 7* (11.55% +- 0.24%).

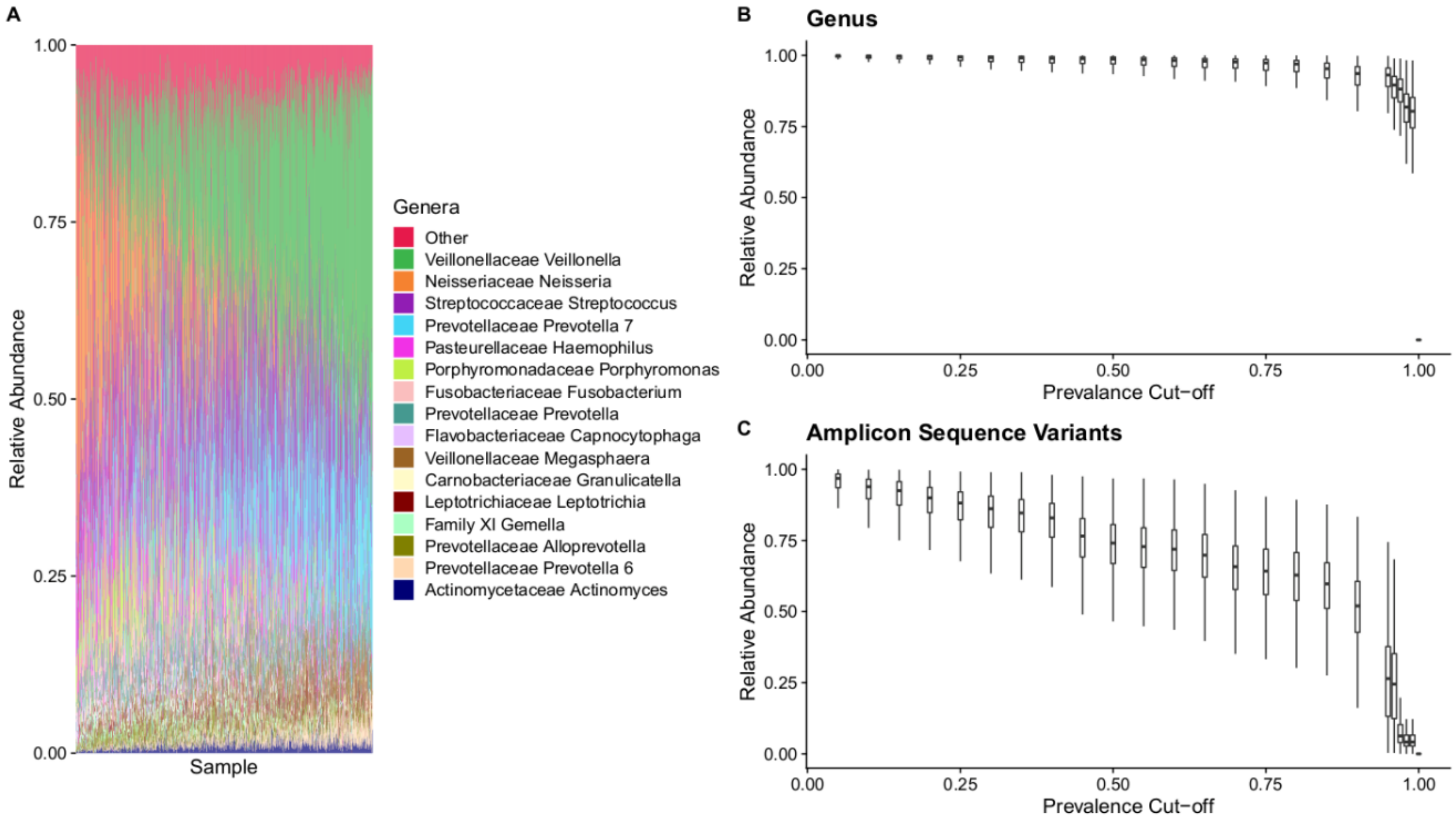

A
Relative Abundance
1.00
0.75
0.50
0.25
0.00
Sample
Genera
🟥 Other
🟩 Veillonellaceae Veillonella
🟧 Neisseriaceae Neisseria
🟪 Streptococcaceae Streptococcus
🟦 Prevotellaceae Prevotella 7
🟪 Pasteurellaceae Haemophilus
🟩 Porphyromonadaceae Porphyromonas
🟥 Fusobacteriaceae Fusobacterium
🟩 Prevotellaceae Prevotella
🟪 Flavobacteriaceae Capnocytophaga
🟫 Veillonellaceae Megasphaera
🟨 Carnobacteriaceae Granulicatella
🟥 Leptotrichiaceae Leptotrichia
🟩 Family XI Gemella
🟩 Prevotellaceae Alloprevotella
🟧 Prevotellaceae Prevotella 6
🟦 Actinomycetaceae Actinomyces
B
Genus
Relative Abundance
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Prevalance Cut−off
C
Amplicon Sequence Variants
Relative Abundance
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Prevalence Cut−off

Figure 4.2: ***Atlantic Canadian oral microbiome composition is dominated by the genus Veillonella and is relatively similar at the genus level but highly variable at the ASV level****. Samples were from the Atlantic Partnership for Tomorrow's Health project (n = 1,049). Samples were subsampled to a depth of 5,000 reads. (A) Genera that had a mean relative abundance of less than 1% were grouped into "Other." (B) Genera were removed at different sample presence cutoffs, and the remaining total mean relative abundance of nonfiltered genera was then calculated. (C) ASVs were removed at different sample presence cutoffs, and the remaining total mean relative abundance of nonfiltered ASVs was then calculated.*

To characterise the core relative abundance of core genera and ASVs within the oral microbiome of these samples the mean relative abundance of genera/ASVs that were present in greater than a specific percentage of samples was analysed. Interestingly, we found that at the genus level the oral microbiome is relatively stable with 11 genera (Supplemental Figure 2A, Supplemental Table 1) present in greater than 99% of all individuals making up on average a total relative abundance of 77.82% (Figure 4.2B). However, this was not the case when we examined composition at a higher taxonomic resolution. We then found that only 5.17% on average of the total relative abundance of the oral microbiome was made up of 3 ASVs (Supplemental Figure 2B) shared between 99% of all participants in the study (Figure 4.2C). These ASVs were classified as being in the *Granulicatella*, *Streptococcus*, and *Gemelli* genera but could not confidently be assigned to a specific species.

4.4.2 - Demographic, Anthropometric, and Lifestyle Choices Have Small but Significant Impacts on Oral Microbiome Composition

We examined the relationship of both alpha and beta diversity of the oral microbiome between 41 different variables that described various demographic, lifestyle, and anthropometric measures (Table 4.1). Samples were split into two different cohorts based on whether they had answered all 41 variables of interest. A total of 741 individuals answered all 41 variables and were included in the exploratory cohort. From this cohort we did not find any significant associations between any of the 41 variables tested and four different alpha diversity metrics (Faith's PD, number of ASVs, Shannon, Evenness) after correction for multiple testing using linear models that were adjusted for DNA extraction batch (Supplemental Data 1). We did, however, find ten variables that were associated with differences in beta diversity as measured by both weighted UniFrac (Figure 4.3A) and Bray Curtis dissimilarity (Figure 4.3B) (PERMANOVA, $q < 0.1$) (Supplemental Data 1). We found two additional variables that were only associated with weighted UniFrac distances, and three additional variables only associated with Bray Curtis dissimilarity (PERMANOVA, $q < 0.1$). Redundancy analysis (ANOVA, $p=0.001$) revealed that multiple anthropometric measures such as height, fat free mass, refined grain servings, sleeping light exposure, and waist to hip ratio were associated in similar manners. Furthermore, as expected increases in all these features were associated in an opposite direction to being female (Figure 4.3C). As sex plays an important role in determining the height, fat free mass, and waist-hip ratio of an individual, we attempted to determine whether sex was confounding our results from these variables. A separate analysis on weighted UniFrac distances controlling for sex indicated that fat free mass

(p=0.02, r^2=0.0039) and waist-hip ratio (p=0.03, r^2=0.0039), but not height (p=0.44, r^2=0.0012) was significantly associated to microbial composition despite differences in sex.

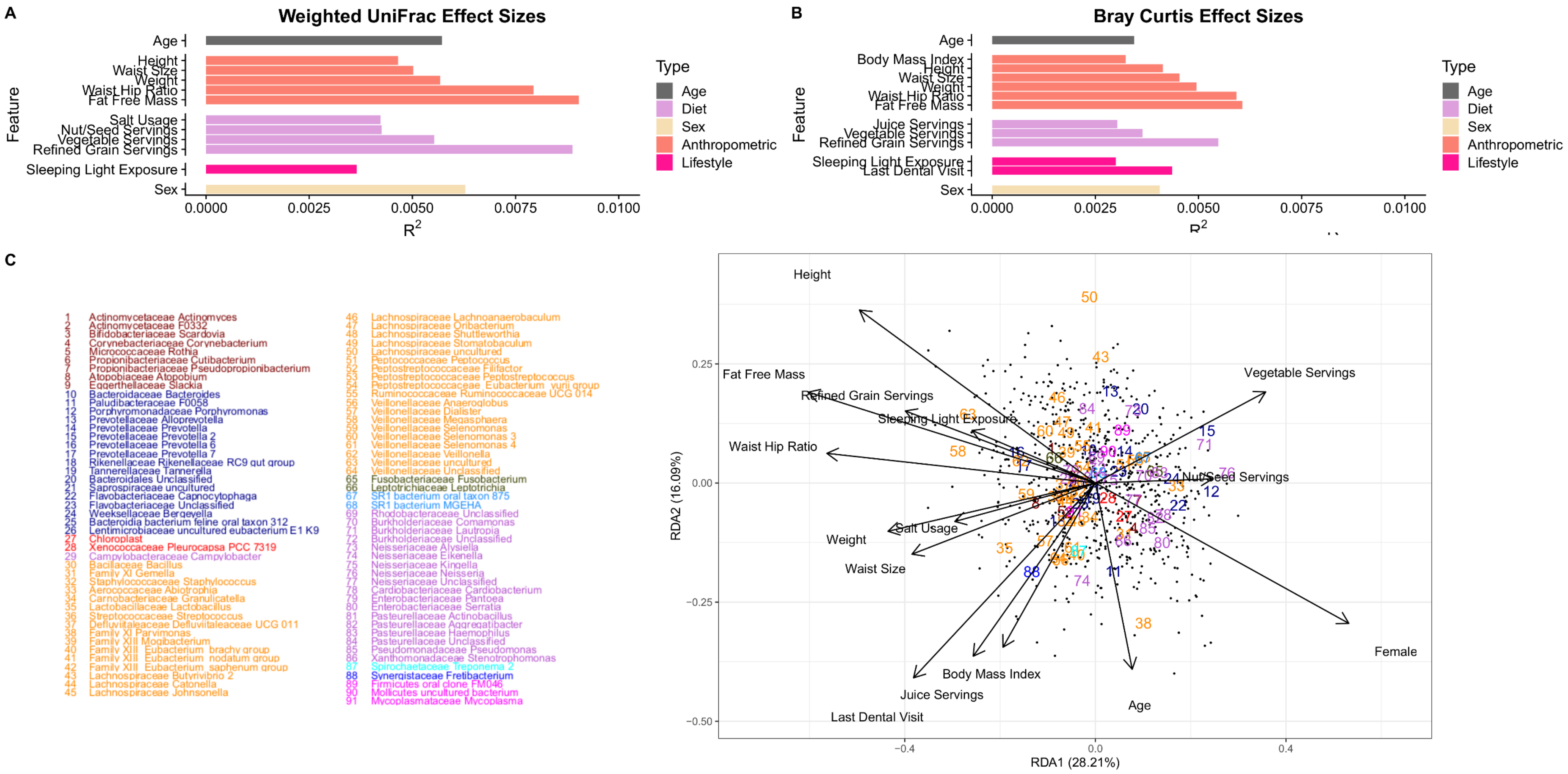

A
Weighted UniFrac Effect Sizes
Feature
Age
Height
Waist Size
Weight
Waist Hip Ratio
Fat Free Mass
Salt Usage
Nut/Seed Servings
Vegetable Servings
Refined Grain Servings
Sleeping Light Exposure
Sex
0.0000
0.0025
0.0050
0.0075
0.0100
R^2
Type
Age
Diet
Sex
Anthropometric
Lifestyle
B
Bray Curtis Effect Sizes
Feature
Age
Body Mass Index
Height
Waist Size
Weight
Waist Hip Ratio
Fat Free Mass
Juice Servings
Vegetable Servings
Refined Grain Servings
Sleeping Light Exposure
Last Dental Visit
Sex
0.0000
0.0025
0.0050
0.0075
0.0100
R^2
Type
Age
Diet
Sex
Anthropometric
Lifestyle
C
1 Actinomycetaceae Actinomyces
2 Actinomycetaceae F0332
3 Bifidobacteriaceae Scardovia
4 Corynebacteriaceae Corynebacterium
5 Micrococcaceae Rothia
6 Propionibacteriaceae Cutibacterium
7 Propionibacteriaceae Pseudopropionibacterium
8 Atopobiaceae Atopobium
9 Eggerthellaceae Slackia
10 Bacteroidaceae Bacteroides
11 Paludibacteraceae F0058
12 Porphyromonadaceae Porphyromonas
13 Prevotellaceae Alloprevotella
14 Prevotellaceae Prevotella
15 Prevotellaceae Prevotella 2
16 Prevotellaceae Prevotella 6
17 Prevotellaceae Prevotella 7
18 Rikenellaceae Rikenellaceae RC9 gut group
19 Tannerellaceae Tannerella
20 Bacteroidales Unclassified
21 Saprospiraceae uncultured
22 Flavobacteriaceae Capnocytophaga
23 Flavobacteriaceae Unclassified
24 Weeksellaceae Bergeyella
25 Bacteroidia bacterium feline oral taxon 312
26 Lentimicrobiaceae uncultured eubacterium E1 K9
27 Chloroplast
28 Xenococcaceae Pleurocapsa PCC 7319
29 Campylobacteraceae Campylobacter
30 Bacillaceae Bacillus
31 Family XI Gemella
32 Staphylococcaceae Staphylococcus
33 Aerococcaceae Abiotrophia
34 Carnobacteriaceae Granulicatella
35 Lactobacillaceae Lactobacillus
36 Streptococcaceae Streptococcus
37 Defluviitaleaceae Defluviitaleaceae UCG 011
38 Family XI Parvimonas
39 Family XIII Mogibacterium
40 Family XIII Eubacterium brachy group
41 Family XIII Eubacterium nodatum group
42 Family XIII Eubacterium saphenum group
43 Lachnospiraceae Butyrivibrio 2
44 Lachnospiraceae Catonella
45 Lachnospiraceae Johnsonella
46 Lachnospiraceae Lachnoanaerobaculum
47 Lachnospiraceae Oribacterium
48 Lachnospiraceae Shuttleworthia
49 Lachnospiraceae Stomatobaculum
50 Lachnospiraceae uncultured
51 Peptococcaceae Peptococcus
52 Peptostreptococcaceae Filifactor
53 Peptostreptococcaceae Peptostreptococcus
54 Peptostreptococcaceae Eubacterium yurii group
55 Ruminococcaceae Ruminococcaceae UCG 014
56 Veillonellaceae Anaeroglobus
57 Veillonellaceae Dialister
58 Veillonellaceae Megasphaera
59 Veillonellaceae Selenomonas
60 Veillonellaceae Selenomonas 3
61 Veillonellaceae Selenomonas 4
62 Veillonellaceae Veillonella
63 Veillonellaceae uncultured
64 Veillonellaceae Unclassified
65 Fusobacteriaceae Fusobacterium
66 Leptotrichiaceae Leptotrichia
67 SR1 bacterium oral taxon 875
68 SR1 bacterium MGEHA
69 Rhodobacteraceae Unclassified
70 Burkholderiaceae Comamonas
71 Burkholderiaceae Lautropia
72 Burkholderiaceae Unclassified
73 Neisseriaceae Alysiella
74 Neisseriaceae Eikenella
75 Neisseriaceae Kingella
76 Neisseriaceae Neisseria
77 Neisseriaceae Unclassified
78 Cardiobacteriaceae Cardiobacterium
79 Enterobacteriaceae Pantoea
80 Enterobacteriaceae Serratia
81 Pasteurellaceae Actinobacillus
82 Pasteurellaceae Aggregatibacter
83 Pasteurellaceae Haemophilus
84 Pasteurellaceae Unclassified
85 Pseudomonadaceae Pseudomonas
86 Xanthomonadaceae Stenotrophomonas
87 Spirochaetaceae Treponema 2
88 Synergistaceae Fretibacterium
89 Firmicutes oral clone FM046
90 Mollicutes uncultured bacterium
91 Mycoplasmataceae Mycoplasma
Height
Fat Free Mass
Refined Grain Servings
Sleeping Light Exposure
Waist Hip Ratio
Salt Usage
Weight
Waist Size
Vegetable Servings
Nut/Seed Servings
Female
Body Mass Index
Juice Servings
Last Dental Visit
Age
0.25
0.00
−0.25
−0.50
RDA2 (16.09%)
−0.4
0.0
0.4
RDA1 (28.21%)

Figure 4.3: **Various anthropometric, dietary, and lifestyle features are significantly associated with oral microbiome composition.** *Saliva samples were from the Atlantic Partnership for Tomorrow's Health cohort (n = 741). Samples were subsampled to a depth of 5,000 reads. Two different metrics measuring beta diversity were tested, weighted Unifrac distances (A) and Bray-Curtis dissimilarity (B), using a PERMANOVA test while controlling for differences in DNA extraction and correction for false discovery (q < 0.1). Relationships between significant features, samples, and genera that were present in at least 10% of samples were then visualized by redundancy analysis (RDA) on center-log-ratio genus count tables. (C) Genera are colored by phylum and labeled numerically.*

Examining the amount of variation explained by each metadata feature by itself after controlling for DNA extraction showed small effect sizes for both weighted UniFrac distances and Bray Curtis dissimilarities (r^2 0.0030 - 0.009) (Figure 4.3A-B). Of the features that were significant, sleeping light exposure explained the least amount of variation in both weighted UniFrac distances ($r^2 = 0.0036$) and Bray-Curtis dissimilarity ($r^2=0.0030$). We also found that fat free mass explained the largest amount of variation in both weighted UniFrac ($r^2=0.009$) and Bray Curtis dissimilarity ($r^2=0.006$). In general, we found that the rankings of effect sizes between these two different metrics agreed (Figure 4.3A-B).

We also examined Random Forest machine learning classification and regression performance for each of these significant features. We found that overall Random Forest models preformed poorly but did show slight associations between some variables (Supplemental Figure 3). For example, the AUROC for sex classification was 0.638

indicating slightly better than random performance. Regression models for features such as height and age showed an r^2 of 0.10 and 0.075 with a RMSE of 8.629 cm and 7.635 years, respectively (Supplemental Figure 3). Interestingly some features, previously identified as being significant, performed extremely poorly such as the number of refined grain servings (r^2=8.22E-6) or vegetable servings (r^2=0.004) (Supplemental Figure 3).

Examining each significant factor in our weighted UniFrac analysis using a backward selected multivariate PERMANOVA, we found that 7.0% of total oral microbiome variation could be explained by a total of 6 significant factors including DNA extraction batch despite using the same protocol, equipment, and personnel for each round (Supplemental Table 2). Interestingly, of these 6 factors DNA extraction number explained a considerable amount of the variation alone (4.18%) (Supplemental Table 2). We found similar results examining beta diversity variation using Bray Curtis dissimilarity with a slightly higher number of significant features and lower total variation explained (5.87%) (Supplemental Table 3). It should be noted that many features were highly correlated with one another (R > 0.7) and as such model selection for these multivariate PERMANOVAs could have suffered due to collinearity of these features. However, a model containing all features that were significantly associated with either weighted UniFrac or Bray-Curtis dissimilarity during univariate testing explained a similar level of variation for both weighted UniFrac and Bray-Curtis dissimilarity profiles (8.09%, 6.81%).

Redundancy analysis revealed several potential taxonomic associations with various features (ANOVA, p=0.001). For example, the genus *Megasphaera* [58] is in the same direction as increasing fat free mass, height, waist-hip ratio, and daily refined grain servings while also opposite to being female (Figure 4.3C). Another uncultured genus in Veillonellaceae family [63] was similarly grouped. The genus *Parvimonas* [38] is in a similar direction as increasing age and being female. Both *Lautropia* [71] and *Prevotella 2* are associated with increasing vegetable intake and *Neisseria* [76] is associated with increasing nut/seed servings and decreasing refined grain servings (**Fig 3C**). The only genus in the phylum Synergistetes that passed the 10% prevalence filtering was found to be associated with increasing juice servings, BMI, and time since last dental appointment. Overall, we found that phyla tended to cluster together, with Firmicutes and Proteobacteria clustering in opposite directions (Figure 4.3C).

To help validate the associations we found between features and weighted UniFrac and Bray-Curtis dissimilarities we analyzed an additional 308 samples from a smaller subset of the Atlantic PATH cohort that had not completely answered all 41 variables of interest. We found that associations between both beta diversity metrics (weighted UniFrac, Bray-Curtis dissimilarity) and anthropometric features such as height, weight, waist-hip ratio, and fat free mass were recoverable within our smaller cohort (Table 4.2, Supplemental Figure 4). We were unable to recover any significant taxonomic dietary associations within this smaller validation cohort. We also were unable to recover taxonomic associations between lifestyle variables such as sleeping light exposure or the time since an individual's last dental visit. The inability to recover these differences could have been due to the highly reduced sample size within this validation cohort.

Table 4.2: ***Validation of beta diversity results***

Metric	**Feature**	**P-value**	**r^2**
Weighted UniFrac	Waist-hip Ratio	0.0190	0.0116
	Height	0.001	0.0117
	Weight	0.010	0.0102
	Fat Free Mass	0.002	0.0172
	Sex	0.0390	0.0080
	Age	0.0120	0.0105
Bray-Curtis	Waist-hip Ratio	0.0140	0.0072
	Height	0.0030	0.0118
	Weight	0.0020	0.0096
	Fat Free Mass	0.0040	0.0110
	Waist Size	0.0210	0.0065
	Age	0.0020	0.0106
	Sex	0.0380	0.0059

4.4.3 - The Abundance of Various Oral Bacterial Genera and ASVs are Associated With Anthropometric Measurements, and Dietary Choices in Healthy Individuals

We next decided to identify genera that were associated with the fifteen features previously identified as being associated with beta diversity in either the weighted UniFrac or Bray Curtis dissimilarity analysis. We found 42 genera (Figure 4.4A) and 42 ASVs (Figure 4.4B) that had abundance profiles that were significantly associated with at least one of these features after controlling for DNA extraction. We found that sex, height, and fat free mass shared similar genera and ASV associations. To control for the possibility of sex confounding our height and fat free mass associations we reanalysed the data controlling for sex. We found that no ASVs or genera were significantly associated to fat free mass after controlling for sex and only 3 genera Chloroplast, Burkholderiaceae unclassified, and *Treponema 2* were significantly associated to height. Interestingly two of these three genera were not previously associated to height in our initial analysis. These results suggest that many of these features associated to height or fat free mass may be driven by differences in sex. To test this, we also tested for differences in sex while controlling for both fat free mass and height. Interestingly, we did not find any significantly associated ASVs and only three significantly associated genera Defluvittaleaceae *UCG-011*, *Leptotrichia*, and *Treponema 2*.

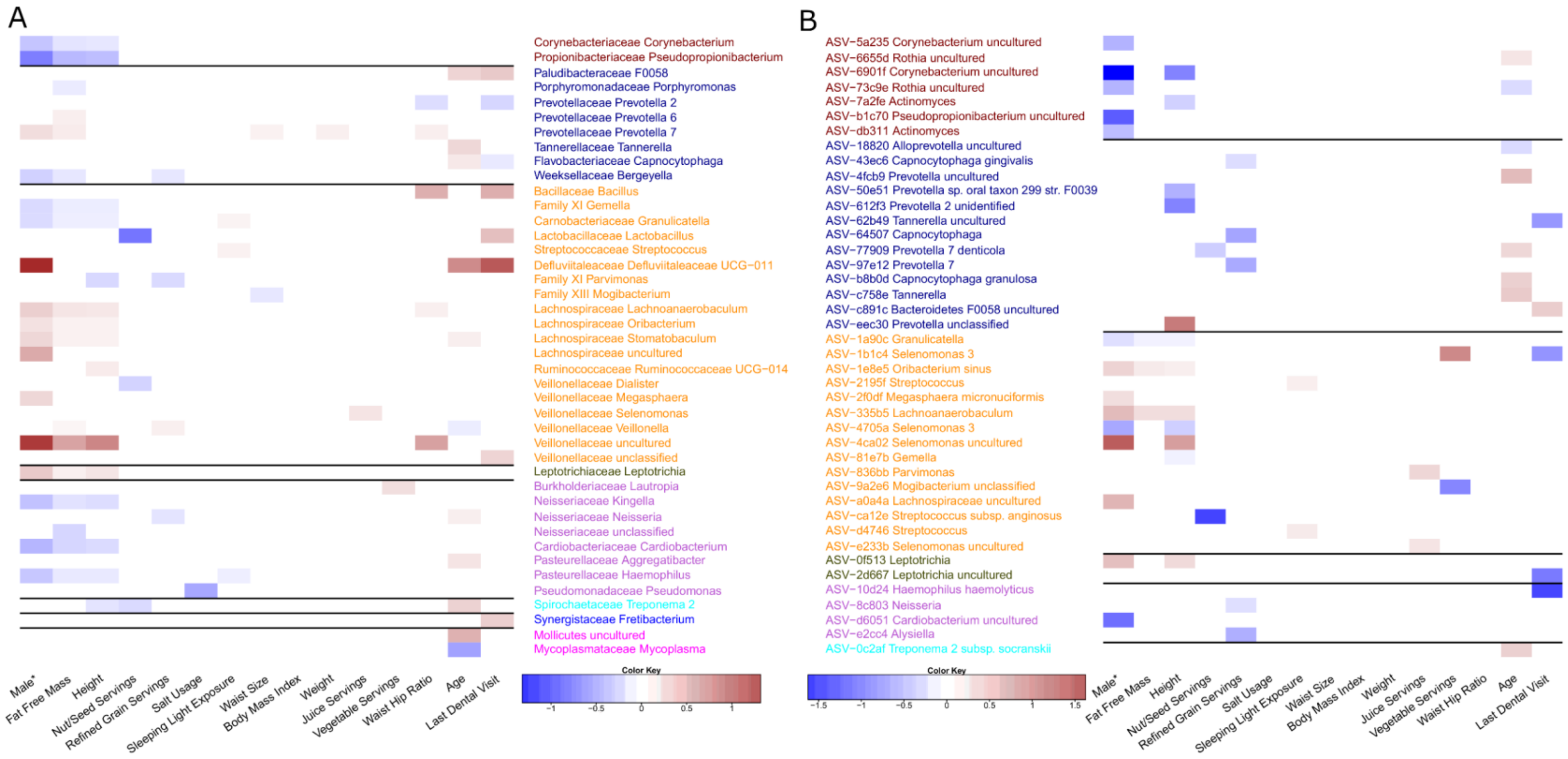

A
Corynebacteriaceae Corynebacterium
Propionibacteriaceae Pseudopropionibacterium
Paludibacteraceae F0058
Porphyromonadaceae Porphyromonas
Prevotellaceae Prevotella 2
Prevotellaceae Prevotella 6
Prevotellaceae Prevotella 7
Tannerellaceae Tannerella
Flavobacteriaceae Capnocytophaga
Weeksellaceae Bergeyella
Bacillaceae Bacillus
Family XI Gemella
Carnobacteriaceae Granulicatella
Lactobacillaceae Lactobacillus
Streptococcaceae Streptococcus
Defluviitaleaceae Defluviitaleaceae UCG−011
Family XI Parvimonas
Family XIII Mogibacterium
Lachnospiraceae Lachnoanaerobaculum
Lachnospiraceae Oribacterium
Lachnospiraceae Stomatobaculum
Lachnospiraceae uncultured
Ruminococcaceae Ruminococcaceae UCG−014
Veillonellaceae Dialister
Veillonellaceae Megasphaera
Veillonellaceae Selenomonas
Veillonellaceae Veillonella
Veillonellaceae uncultured
Veillonellaceae unclassified
Leptotrichiaceae Leptotrichia
Burkholderiaceae Lautropia
Neisseriaceae Kingella
Neisseriaceae Neisseria
Neisseriaceae unclassified
Cardiobacteriaceae Cardiobacterium
Pasteurellaceae Aggregatibacter
Pasteurellaceae Haemophilus
Pseudomonadaceae Pseudomonas
Spirochaetaceae Treponema 2
Synergistaceae Fretibacterium
Mollicutes uncultured
Mycoplasmataceae Mycoplasma
Male*
Fat Free Mass
Height
Nut/Seed Servings
Refined Grain Servings
Salt Usage
Sleeping Light Exposure
Waist Size
Body Mass Index
Weight
Juice Servings
Vegetable Servings
Waist Hip Ratio
Age
Last Dental Visit
Color Key
−1
−0.5
0
0.5
1
B
ASV−5a235 Corynebacterium uncultured
ASV−6655d Rothia uncultured
ASV−6901f Corynebacterium uncultured
ASV−73c9e Rothia uncultured
ASV−7a2fe Actinomyces
ASV−b1c70 Pseudopropionibacterium uncultured
ASV−db311 Actinomyces
ASV−18820 Alloprevotella uncultured
ASV−43ec6 Capnocytophaga gingivalis
ASV−4fcb9 Prevotella uncultured
ASV−50e51 Prevotella sp. oral taxon 299 str. F0039
ASV−612f3 Prevotella 2 unidentified
ASV−62b49 Tannerella uncultured
ASV−64507 Capnocytophaga
ASV−77909 Prevotella 7 denticola
ASV−97e12 Prevotella 7
ASV−b8b0d Capnocytophaga granulosa
ASV−c758e Tannerella
ASV−c891c Bacteroidetes F0058 uncultured
ASV−eec30 Prevotella unclassified
ASV−1a90c Granulicatella
ASV−1b1c4 Selenomonas 3
ASV−1e8e5 Oribacterium sinus
ASV−2195f Streptococcus
ASV−2f0df Megasphaera micronuciformis
ASV−335b5 Lachnoanaerobaculum
ASV−4705a Selenomonas 3
ASV−4ca02 Selenomonas uncultured
ASV−81e7b Gemella
ASV−836bb Parvimonas
ASV−9a2e6 Mogibacterium unclassified
ASV−a0a4a Lachnospiraceae uncultured
ASV−ca12e Streptococcus subsp. anginosus
ASV−d4746 Streptococcus
ASV−e233b Selenomonas uncultured
ASV−0f513 Leptotrichia
ASV−2d667 Leptotrichia uncultured
ASV−10d24 Haemophilus haemolyticus
ASV−8c803 Neisseria
ASV−d6051 Cardiobacterium uncultured
ASV−e2cc4 Alysiella
ASV−0c2af Treponema 2 subsp. socranskii
Color Key
−1.5
−1
−0.5
0
0.5
1
1.5
Male*
Fat Free Mass
Height
Nut/Seed Servings
Refined Grain Servings
Salt Usage
Sleeping Light Exposure
Waist Size
Body Mass Index
Weight
Juice Servings
Vegetable Servings
Waist Hip Ratio
Age
Last Dental Visit

*Figure 4.4***: *Differentially abundant genera and ASVs whose abundance profiles are associated with features found to influence oral microbiome composition.*** *Genera (A) and ASVs (B) meeting a false discovery rate of q < 0.1 using the Corncob R package, which uses beta-binomial regressions. Each feature's false discovery rate was corrected separately, and each was tested to control for differences in DNA extraction and differential variability within that feature. Ordinal variables were converted into a ranked scale for testing, and all features except for sex were scaled. The asterisk indicates that sex was treated as a categorical value; therefore, the magnitude is not directly comparable to other log odd ratios.*

We did not find any other features that shared similar patterns of taxonomic associations but there were multiple genera with multiple feature associations. The genus *Prevotella 7* had the highest number of features (5) associated with its relative abundance including four anthropometric measurements (height, fat free mass, waist size, waist-hip ratio, and weight) and sex. Interestingly, BMI did not have any genera or ASVs significantly associated despite many other anthropometric measures showing strong taxonomic signals. We were unable to identify any single ASVs associated to waist size and weight but were able to identify a small number of genera including *Prevotella 7*, which was related to both and *Mogibacterium* with waist size. We also found that for some phyla, all taxa with significant associations had the same effect size direction. For example, genera in the Actinobacteria or Proteobacteria phyla tended to be negatively associated with fat free mass, height, and being male. We also found several genera in the Proteobacteria phylum that were significantly associated with the amount of time since an individuals last dental appointment.

In contrast, examining the ASVs associated with each feature we found that in a small number of cases ASVs in the same genera had opposite directions of association to the same features. For example, two ASVs classified as *Rothia* uncultured were both significantly associated to age but in opposite directions suggesting that lower taxonomic resolution is required to identify some associations. Furthermore, we also identified cases were ASVs that were associated to a feature were classified in a genus that was found not to be related to that feature. For example, ASV-4ca02 *Selenomonas* uncultured was strongly associated with being male even though this entire collective genus was not (Figure 4.4). Further examples include ASV-e2cc4 which was classified in the genus *Alysiella*, and significantly associated with reduced refined grain servings. Examples of the opposite occurrence are also present with genera such as *Mycoplasma* being associated with age but no single ASV for this associated could be identified.

We further validated our differential abundance analysis using this cohort and found 8/17 genera associated with sex, 8/16 genera associated with fat free mass, 5/15 genera associated with height, and 3/11 genera associated with age were recoverable within this smaller cohort (Supplemental Figure 5A). Additionally, the negative association between *Prevotella 2* and waist-hip ratio was also verified within this cohort. Furthermore, several associations between ASVs and features such as sex (5/14), height (4/12), fat free mass (2/3), and sleeping light exposure (1/2) were also found within this smaller validation cohort (Supplemental Figure 5B). All significant effect sizes that were recovered in the validation cohort except for one, between sleeping light exposure and ASV-d4746 *Streptococcus*, remained in the same direction as the original cohort indicating relationships that were robust to sample choice.

4.4.4 - Predicted Microbial Pathway Abundances Reveal Multiple Pathways Associated With Anthropometric, Dietary and Age and Sex Features

Microbial pathway abundances were predicted using PICRUSt2 (Douglas et al., 2020) to determine potential associations between pathway abundances and features previously identified to be significantly associated with differences in beta diversity. Differential analysis between features and predicted pathway abundances were done using Corncob with Benjamini-Hochberg corrected p-values at an alpha of 0.05 and associations with effect sizes under |0.05| log odds filtered out. We found 9/15 features originally associated with beta diversity metrics to have at least one predicted pathway association (Figure 4.5). Of these features we found that refined grain servings had the largest number (N=33) of predicted pathway associations. Of these associations many were negatively associated with increasing refined grain intake including various TCA cycle derivatives, glucose, and xylose degradation, 2-methylcitrate cycle, and heme biosynthesis. Furthermore, only a smaller number of pathways were associated with increasing refined grain intake such as phylloquinol biosynthesis, and CMP-legionanimate biosynthesis (Figure 4.5).

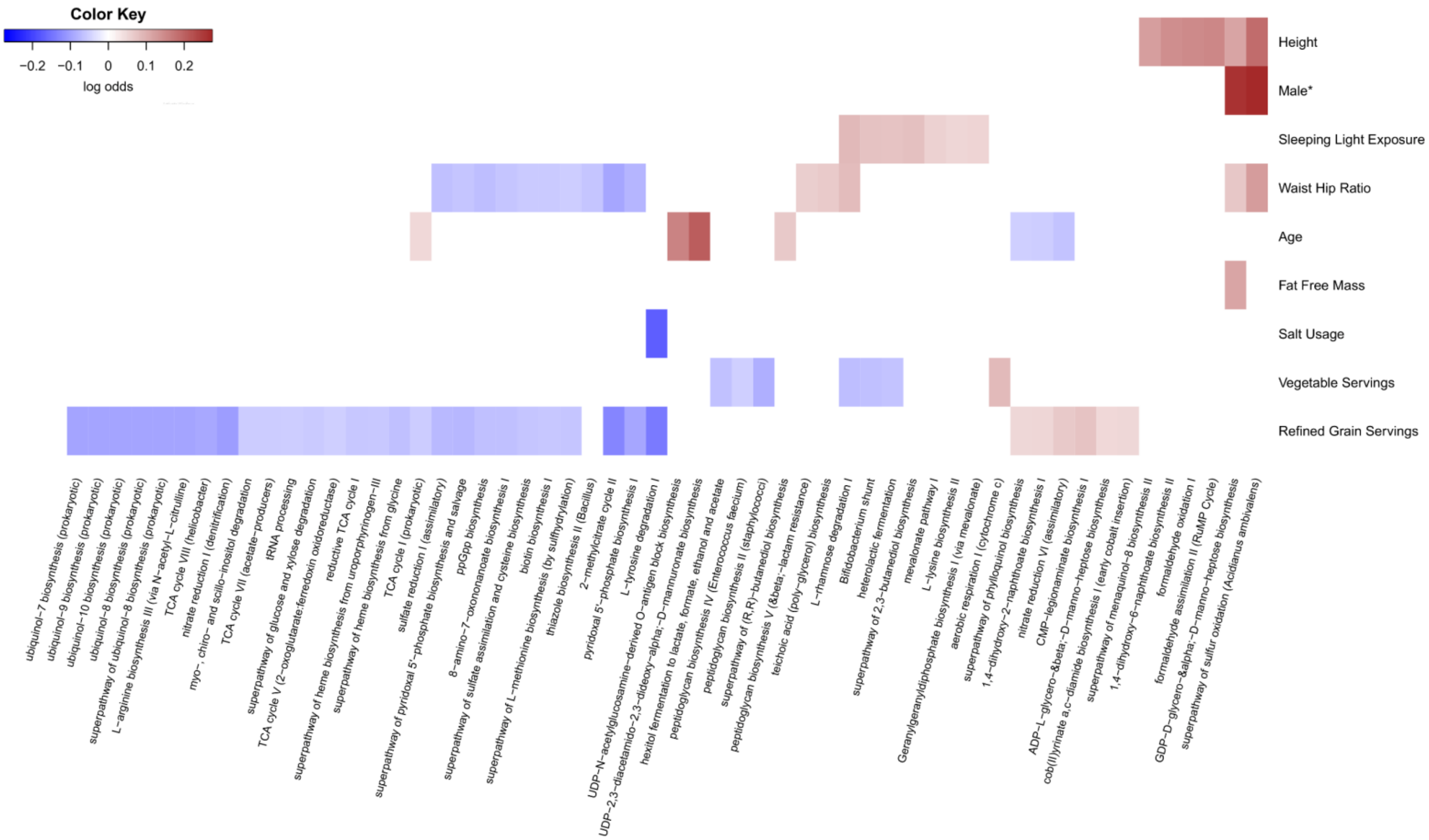


*Figure 4.5***: Various predicted pathway abundances are associated with features significantly associated with overall microbiome composition.** *Pathway abundances were predicted from 16S rRNA gene sequencing data using PICRUSt2. Predicted pathway abundances meeting an FDR of <0.05 and an effect size of |0.05| log odds were considered significant associations using the Corncob R package. Each feature's false discovery rate was corrected separately, and each was tested to control for differences in DNA extraction and differential variability within that feature. Ordinal variables were converted into a ranked scale for testing, and all features except for sex were scaled. The asterisk indicates that sex was treated as a categorical value; therefore, the magnitude is not directly comparable to other log odd ratios.*

Only one pathway, aerobic respiration I (cytochrome c) was predicted to be associated with increasing vegetable serving while six pathways were associated in the opposite direction. These pathways included fermentation of carbohydrates into lactate, lactic acid, ethanol, acetate, and formate as well as the biosynthesis of peptidoglycan. We found only one association with salt usage (L-tryosine degradation) and did not find any predicted associations with juice serving intake.

A number of predicted pathway abundances were also associated with various anthropometric features with waist-hip ratio having the highest number of predicted associations (N=15) and fat free mass having only one predicted association (GDP-D-glycero-α-D-mannose -heptose biosynthesis). We also found a small number of predicted pathway abundances associated with Age (N=7) and Sex (N=2).

Validation analysis on the second smaller dataset was only able to validate a minority of predicted pathway associations many of which were associated with an individual's waist to hip ratio (8/15) (Supplemental Figure 6). Only four out of the original 33 predicted pathway associations with refined grain intake were verified within this cohort. These pathways included L-tyrosine degradation, ADP-L-glycero-α-D-manno-heptose biosynthesis, phylloquinol biosynthesis, and 1,4-dihydroxy-2-naphthoate biosynthesis. Three out of six pathways associated with height and four out of seven pathways associated with age were also found to significant within the smaller validation dataset (Supplemental Figure 6).

4.5 – Discussion

Our analysis of 1049 healthy (Figure 4.1) individuals from Atlantic Canada revealed that much of the oral microbiome of Atlantic Canadians was made up of eleven "core" genera that belong to six different phyla (*Actinobacteria*, *Fusobacteria*, *Proteobacteria*, *Firmicutes*, *Bacteroidetes*, and *Fusobacteria*). Interestingly some of these core genera found in 99% of all samples were found in relatively low abundance (<2% mean abundance) indicating that bacteria within the oral microbiome can be consistently observed with minor contributions (Supplemental Table 1). In contrast, the composition at the ASV level had only 3 ASVs being present in 99% of samples and only contributing 5.17% of the total oral microbiome composition on average. Overall, these results indicate that individuals tend to share similar genera within the oral cavity, but the species/strains shared between individuals can be highly variable. These findings are inline with previous work from the Human Microbiome project that found the oral microbiome to be relatively similar between individuals at the genus level (Huttenhower et al., 2012).

We found that various anthropometric and lifestyle features were significantly associated with overall oral microbiome composition, however, they explained only a small amount of total oral microbiome variance while controlling for DNA extraction batch (5.87-7.00%) (Supplemental Table 2-3). We found that fat free mass explained the highest amount of variance (0.6-0.9%) (Figure 4.3A-B) of all biological features. While this feature had many differential abundant genera and ASVs associated with it, we were unable to recover any of them after controlling for differences in sex. This could indicate that these associations could be driven by sex and not underlying fat mass, however, we were also unable to recover many relationships between sex and taxonomic abundance while controlling for fat free mass indicating that both of these factors significantly confound the other. However, despite this issue there is previous evidence to suggest that some bacteria are related to differences in body size. A study in children found reduced abundance of *Veillonella*, *Prevotella*, *Selenomonas,* and *Streptococcus* in obese children (Raju et al., 2019). Interestingly, in our adult population we found similar trends with members of the *Veillonella* family being positively associated with increasing fat free mass, and members of the *Provetella* genus also being linked with higher fat free mass. Another publication on the Southern Community Cohort Study found that both *Granulicatella* and *Gemella* were associated with obesity (Y. Yang, Cai, Zheng, et al., 2019) which we also found within our cohort at both the genus and ASV level. One interesting result from our study was our inability to identify any genera or ASVs linked to BMI, despite numerous relationships between anthropometric measurements being identified. These results indicate that future studies should be advised to include sex and other measurements of body composition, such as lean body mass, when looking at relationships between the microbiome and obesity.

We found two genera *Defluviitaleaceae* UCG-011 and an uncultured genus from *Veillonellaceae* which were strongly associated with being male (Figure 4.4A). However, neither of these associations were recovered in our validation cohort indicating that they could either be false positives or require a larger sample size to recover due to their low mean relative abundance 0.0042% and 0.063%, respectively. Despite this we were still able to recover eight genus level associations in our validation cohort (Supplemental Figure 5A), however, only a few of these associations match those that were previously reported. Renson et al. found two genera *Lactobacillus* and *Actinobacillus* to be higher in males, which we did not find within our study (Renson et al., 2019). This could have been due to multiple differences including sampling procedures, systemic protocol bias or the compositional nature of microbiome data (Morton et al., 2019). Raju et al. found that there was a high relative abundance of *Haemophilus* in females which we also found in our study, however they also found *Oribacterium* to be increased in females which was opposite from what was found in this study (Raju et al., 2019). Differences between these studies and ours can in part be attributed to differences in sample collection procedure and sequencing primers used highlighting the technical biases within the field (Sinha et al., 2017).

We were unable to recover any taxonomic relationships between dietary features within our validation cohort, however, refined grain servings per day had the largest impact on overall oral microbiome composition and microbial pathway potential in our initial analysis. During this initial analysis, we found that bacteria from four genera *Bergeyella*, *Parvimonas, Veillonella,* and *Neisseria* decreased in relative abundance with increasing refined grain intake (Figure 4.4). Interestingly, refined grain intake had a very strong association with inflammatory bowel disease in a previous analysis of this cohort (DeClercq, Langille, & Van Limbergen, 2018), and alterations in the oral microbiome have been linked to inflammatory bowel disease in the past (Xun, Zhang, Xu, Chen, & Chen, 2018). Work by Said et al., found multiple genera in differential abundance between individuals with and without IBD including the increased presence of *Veillonella* in IBD patients (Said et al., 2013), which we found to be linked positively with refined grain intake.

We found that of all features significantly associated with overall oral microbiome composition refined grain intake had the largest number of predicted pathway associations (Figure 4.5). Many of these pathways were related to metabolic functions and the biosynthesis of various cofactors and metabolic building blocks indicating a shift in metabolic potential within the microbial community. This shift is not surprising given that differing levels of refined grain intake could impact the availability of various carbohydrates to oral microbiota. However, it should be noted that only a small number of these pathway associations were verified within our small validation dataset.

Other dietary factors we found linked to overall oral microbiome composition in our original analysis include both juice servings and vegetable servings. However, we were only able to find a small number of genera, ASVs, and predicted pathway abundances linked to vegetable serving intake. We found a number of fermentation pathways were predicted to be associated with reduced vegetable intake indicating a possible shift in anaerobic activity. While we found a small number of taxonomic associations with juice serving intake we found no predicted pathway associations. Furthermore, we were unable to recover any taxonomic or pathway associations for both vegetable intake or juice serving intake in our validation cohort indicating the possibility for a false positive or the requirement of a large sample size to see these effects. Previous work within the field has found conflicting evidence on the role of diet impacting oral microbiome composition and may be reflective of different dietary assessment methods. observed results from a study.

Looking at all features that were significantly associated with oral microbiome composition together in a single model we were only able to explain a small portion of the total variance between samples (5.87-7.00%). This indicates that while many of these features are significantly related to microbial composition each one by themselves tends to only cause small shifts in overall microbial composition. Furthermore, a majority of the variance accounted for was due to differences in DNA extraction date. This shows that while slight technical variations such as the time when DNA extraction was done can have larger impacts on sample composition emphasizing the need to control for these technical variations during large population-based studies.

One large limitation to our study was our lack of detailed dental history information from participants. While we did record how recently each individual last visited the dentist, we were unable to retrieve detailed information on dental health, which has been found to have dramatic impacts on oral microbiome composition (Takeshita et al., 2016). Furthermore, our study was also unable to capture potential variance that could have been attributed to the time of sampling. Various studies have shown that oral microbiome composition can vary with regard to collection time due to events such as teeth brushing and eating throughout the day (Tomás, Diz, Tobías, Scully, & Donos, 2012). These could explain some of the missing variation that was not accounted for in our study, however, it is unlikely to explain all 93.00% indicating we are still missing a suitable amount of information on what determines an individual's oral microbiome composition. In addition to this we would also like to note that we were unable to account for the length of time samples were on ice during shipping which could have resulted in altered microbial compositions.

In conclusion, our study indicates that the healthy oral microbiome is relatively similar between individuals at the genus level and is impacted very little by any one factor. Future studies that attempt to identify oral microbial biomarkers associated with disease may be encouraged by the lack of major confounding variables and may be justified in controlling only for sex, body composition, oral health, and basic dietary information.

Chapter 5 – Investigating the Oral Microbiome as a Biomarker for Retrospective and Prospective Cases of Prostate, Colon, and Breast Cancer

This chapter is a near reproduction of the same paper currently submitted for publication and reported as a pre-print on bioRxiv. I was the first author on this work and was the primary contributor toward data processing, analysis, figure creation and manuscript writing. I also participated in study design along with my co-authors Vanessa DeClercq and Morgan G.I. Langille.

The pre-print of this publication was produced under a CC-BY 4.0 international license which allows the sharing and adaption of the work if appropriate attribution is given. A copy of this license can be found here: https://creativecommons.org/licenses/by/4.0/. All additional files and supplemental information referred to in this chapter are freely available as part of the original pre-print on the bioRxiv website and can be found at the following link: https://doi.org/10.1101/2022.10.11.511800.

5.1 – Abstract

The human microbiome has been proposed as a useful biomarker for several different human diseases including various cancers. To answer this question, we examined salivary samples from two Canadian population cohorts, the Atlantic Partnership for Tomorrow's Health project (PATH) and Alberta's Tomorrow Project (ATP). Sample selection was then divided into both a retrospective and prospective case control design examining individuals with prostate, breast, or colon cancer. In total 89 retrospective and 260 prospective cancer cases were matched to non-cancer controls and saliva samples were sequenced using 16S rRNA gene sequencing to compare bacterial diversity, and taxonomic composition. We found no significant differences in alpha or beta diversity across any of the three cancer types and two study designs. Although retrospective colon cancer samples did show evidence on visual clustering in weighted beta diversity metrics. Differential abundance analysis of individual taxon showed several taxa that were associated with previous cancer diagnosis in all three groupings within the retrospective study design. However, only one genus (*Ruminococcaceae UCG-014*) in breast cancer and one ASV (*Fusobacterium periodonticum*) in colon cancer was identified by more than one differential abundance (DA) tool. In prospective cases of disease three ASVs were associated with colon cancer, one ASV with breast cancer, and one ASV with prostate cancer. None overlapped between the two different study cohorts. Attempting to identify microbial signals using Random Forest classification showed relatively low levels of signal in both prospective and retrospective cases of breast and prostate cancer (AUC range: 0.394-0.665). Contrastingly, colon cancer did show signal in our retrospective analysis (AUC: 0.745) and in one of two prospective cohorts (AUC: 0.717). Overall, our results indicate that it is unlikely that reliable oral microbial biomarkers of

disease exist in the context of both breast and prostate cancer. However, they do suggest that further research into the relationship between the oral microbiome and colon cancer could be fruitful. Particularly in the context of early disease progression and risk of cancer development.

5.2 – Introduction

The oral microbiome is a highly diverse microbial community that is shaped by several different dietary, anthropometric and lifestyle choices (Belstrøm et al., 2014; Nearing et al., 2020). Recent works have shown that this community of microbes plays important roles in both oral and systemic health (Wade, 2013). For example, the composition of the oral microbiome has been associated with oral diseases such as periodontitis (Lundmark et al., 2019) or more distal diseases such as colorectal cancer (CRC) (Flemer et al., 2017). Recent work has proposed that the oral microbiome may possess biomarkers for the identification of various diseases. However, research on some of the most common cancers such as prostate, colon, and breast cancer are limited. In the year 2022, it is estimated that 28,900 cases of breast cancer, 24,600 cases of prostate and 24,300 cases of colon cancer will be diagnosed in the Canadian population (Brenner et al., 2022). Numerous works have previously examined the associations of these three cancers and the human microbiome; however, most studies have focused on their relationship with the gut microbiota or microbiota associated with the organ of interest.

Of the cancers outlined above, prostate cancer has arguable received the least amount of attention within the microbiome field. Studies on the human microbiome and this disease have been mostly focused on prostate tissue, and the gastrointestinal tract (Javier-DesLoges et al., 2022). Investigation into the microbial inhabitants of prostate tissue have given mixed results due to the low biomass of these samples and contamination issues. Indeed, studies on healthy prostate tissue have given conflicting results, with some indicating the presence of bacteria and others finding no evidence of bacterial inhabitants (Porter et al., 2018). However, in prostate cancer multiple works have demonstrated evidence for bacterial communities in tumors and benign tissue

(Cavarretta et al., 2017; Feng et al., 2019; Yow et al., 2017). This led to interest in determining whether specific microbes might be associated with tumor and non-tumor tissue. To date these studies have given inconsistent results about specific bacteria but ultimately point to similar overall microbial community structures between tumor and benign prostate tissue (Cavarretta et al., 2017; Feng et al., 2019; Yow et al., 2017).

In contrast to the prostate microbiome, it is well known that the gastrointestinal tract contains a rich population of microbes including bacteria, archaea, and fungi (Huttenhower et al., 2012; Wilson & Blitchington, 1996). Work on these communities have shown a myriad of associations with various diseases, metabolites, and immunological states (Petrosino, 2018). Several smaller studies have examined the relationship between the gut microbiome and prostate cancer with mixed results. Two studies by Liss et al., and Golombos et al., found higher levels of Bacteroidetes in patients with prostate cancer indicating potential linkages between the disease and this broad taxonomic group (Golombos et al., 2018; Liss et al., 2018). Furthermore, work by Matsushita et al., in Japanese men with high-Gleason prostate cancer linked an enrichment of short chain fatty acid producing bacteria in the gut and cancer status (Matsushita et al., 2021). However, other works by Alanee et al. and Katz et al., have found no significant differences between individual’s with and without prostate cancer (Alanee et al., 2019; Katz et al., 2022). Leading to mixed results on whether the microbial composition on the gut is truly linked to the development of prostate cancer.

Considering these investigations, work on the oral microbiome and prostate cancer is rather lacking. To the best of our knowledge only one study has examined this potential linkage despite previous works linking periodontitis with the likelihood of prostate cancer development (Lee et al., 2017; Wei, Zhong, Wang, & Huang, 2021). In 2017, work by Estemalik et al., found in a group of 24 patients with chronic prostatitis or benign prostatic hyperplasia that 70.8% had one or more bacteria found within their oral cavity to also be residing within prostatic secretions (Estemalik et al., 2017). These results may indicate a linkage between prostate inflammation and the oral cavity warranting further investigation into the linkage between the oral microbiome and prostate health.

Breast cancer is one of the most commonly diagnosed cancers in females and is one of three cancers examined within this report (Brenner et al., 2022). Studies on breast cancer and the human microbiome have shown variable results with some studies suggesting associations between disease status and microbial community composition within the gut, breast tissue, and urine (Fernández et al., 2018). Indeed, early work on the gut microbiota and breast cancer suggested that microbes within these environments had the capacity to metabolize estrogen and estrogen related products. (Järvenpää, Kosunen, Fotsis, & Adlercreutz, 1980; Plottel & Blaser, 2011). Evidence for these microbial functions termed broadly as the "estrobolome" was first uncovered during the 20th century by several studies that identified fluctuations in estrogen and estrogen related metabolites within the plasma, urine, and feces of pregnant women using antibiotics (Järvenpää et al., 1980; F. Martin, Peltonen, Laatikainen, Pulkkinen, & Adlercreutz, 1975; Tikkanen, Adlercreutz, & Pulkkinen, 1973; Tikkanen, Pulkkinen, & Adlercreutz, 1973; Willman & Pulkkinen, 1971). Since these initial discoveries further work has characterized bacterial enzymes within the gut such as β-glucuronidase that can process conjugated estrogen metabolites (Ervin et al., 2019; Järvenpää et al., 1980). From these findings it has been suggested that the estrobolome may play a role in the risk of breast cancer development through the control of recirculating estrogen levels (Adams, 2016; K. L. Chen & Madak-Erdogan, 2016; Plottel & Blaser, 2011).

Investigation into this hypothesis through examining the relationship between the gut microbiome and breast cancer status has shown varying results. For example, work by Goedert et al, found reduced microbial diversity within the gut of post-menopausal breast cancer patients, however, work in 2018 by Zhu et al., reported significant findings in the opposite direction (Goedert et al., 2015; J. Zhu et al., 2018). Similarly, several gut microbes including those that possess β-glucuronidases have been associated with breast cancer status depending on study and menopausal status (Byrd et al., 2021; Goedert et al., 2018, 2015; Hou et al., 2021; J. Zhu et al., 2018). However, due to the high variability of results between studies and the lack of non-sequenced based validation, no strong conclusions have been made on the role of any specific gut taxon and breast cancer risk.

In addition to the gut, several studies have examined breast cancer status and microbial communities within and on breast tissue and urine. Although, these studies just like the gut, have also shown variable results indicating the need for further work within the field (Fernández et al., 2018). Interestingly, in the case of the oral microbiome, despite work showing that breast cancer is associated with periodontal disease to the best of our knowledge only two published studies have examined its relation with cancer status (Chung, Tsai, Huang, Kao, & Chen, 2016). Early work by Wang et al., examining oral rinses from individuals in the United States found no differences in community composition or specific taxa (H. Wang et al., 2017). However, recent work by Wu et al., did show differences in both microbial diversity and specific taxon within salivary samples from the Ghana Breast Health Study (Z. Wu et al., 2022). These inconsistent results highlight the need for further investigation into the potential relationship of the oral microbiome and breast cancer in different populations.

Finally, one of the most well studied cancers in the context of the human microbiome is CRC. Numerous studies have been conducted on the relationship between CRC and the gut microbiome showing significant differences in community composition between cancer and non-cancer individuals (Zhao, Cho, & Nicolls, 2021). Highlighted within these studies have been the relationships between various bacteria and CRC development or progression; including *Fusobacterium nucleatum*, enterotoxigenic *Bacteroides fragilis*, and (pks+) *Escherichia coli* (Garrett, 2019). However, to date it is still not entirely clear how influential these species are on the development or progression of the disease. For example, work by Dziubańska-Kusibab et al., has shown that colibactin produced by pks+ *E. coli* can cause DNA damage within the colon and is associated with mutational signatures often found in colon cancer patients (Dziubańska-Kusibab et al., 2020). However, (pks+) *E. coli* has yet to be formally recognized as a causative agent of CRC by the World Health Organization.

Barring from these investigations has been research on the relationship between the oral microbiome and CRC. Work in this area has shown that shifts in community composition within the oral microbiome has been associated with disease and in some cases may be predictable. Work in 2018 by Flemer et al., found that at diagnosis the oral microbiome could classify individuals with or without colon cancer with an area under the receiver operator curve (AUROC) of 0.91 (Flemer et al., 2017). Furthermore, their work along with others has also shown that many taxa commonly attributed to the oral cavity are enriched in the gut microbiome of CRC patients (Flemer et al., 2017; Thomas et al., 2019). Additionally, three further studies on the oral microbiome and colorectal cancer within various group settings reported distinct bacterial signatures associated with disease diagnosis (Komiya et al., 2019; Yao Wang et al., 2021; Y. Yang, Cai, Shu, et al., 2019). These results suggests that the oral microbiome may be a useful tool in CRC risk assessment.

Overall, despite the work presented above in prostate, breast, and colon cancer there remains large knowledge gaps in our understanding of the oral microbiome's applicability to population screening for these diseases. For this reason, we were interested in investigating these three cancers at a population level in both a case-control retrospective and prospective study design. To do this we leverage two different population cohorts within the Canadian Partnership Tomorrow’s Health project, the Atlantic Partnership for Tomorrow’s Health (PATH), and Alberta’s Tomorrow Project (ATP). From these two cohorts we selected saliva samples from both retrospective cases and prospective cases of prostate, breast, and colon cancer. This unique study design allowed us to investigate the relationship of these cancers with the oral microbiome both before and after diagnosis. Critically, this has allowed us to assess whether compositional changes within the oral cavity of breast, prostate, and colon cancer individuals exist both before and after diagnosis. The former being crucial for our understanding of whether the oral microbiome may be useful for the detection of individuals at risk for disease development.

5.3 - Materials and Methods

5.3.1 - Study Design

This report includes the analysis of saliva samples from individuals who had previously been enrolled in two regional cohorts within the Canadian Partnership for Tomorrow's Health project, a pan-Canadian prospective cohort study focused on examining the influence of genetics, the environment, and lifestyle factors on Canadian's health. The regional cohorts of interest for this study include Atlantic PATH (which includes participants from the 4 Atlantic provinces: (Nova Scotia, New Brunswick, Prince Edward Island, and Newfoundland and Labrador) and ATP (participants from the western province of Alberta). For this study, both retrospective (cases diagnosed prior to baseline data and sample collection) and prospective (cases diagnosed after baseline data and sample collection) nested case-control designs were employed. This study was granted ethics approval from Dalhousie University Health Sciences Research Ethics Board (REB #2018-4420).

5.3.2 - Atlantic Partnership for Tomorrow's Health Cohort Characteristics

At baseline, demographics, lifestyle, personal and family medical history were self-reported on questionnaires, and a subset attended assessment centers where physical measurements and biospecimens such as saliva were collected. For more details on baseline characteristics of the Atlantic PATH cohort, an in depth descriptive cohort profile has been previously published (Sweeney et al., 2017). Follow-up questionnaire data was collected between 2016-2019.

For the purposes of this study sample selection within the Atlantic cohort was divided into either a retrospective or prospective nested case-control design as previously described. Prior cancer diagnosis was determined through baseline questionnaires filled out by each participant. All available breast, prostate, and colon cancer case saliva samples were included in this study, and control samples (non-cancer) were selected to match cases (1:5) by sex, age (+/-3 years), BMI (+/-3), and smoking status (current vs never/former). The retrospective design included 588 saliva samples from the Atlantic PATH biospecimen repository based on case and non-cancer control matches to individuals that had been diagnosed with breast (n=61), prostate (n= 23), or colon cancer (n= 14) prior to baseline saliva collection.

For the prospective design, new incident cases of cancer were determined through follow up questionnaire surveys filled out by each participant. A one-to-one case control design was used with non-cancer control samples being matched to case samples based on age (+/- 3 years), sex, and BMI (+/- 3). A minor number of current smokers ranging from 0% - 3.70% depending on cancer status were included in this analysis (Table 2). The prospective design included 230 samples from the Atlantic PATH cohort who had breast (n=67), prostate (n=35), or colon cancer (n=13).

The median length of time between sample collection and cancer diagnosis for each cancer can be found in Table 5.1 (retrospective) and Table 5.2 (prospective) along with other sample characteristics broken down by cancer type, study design, and case control status.

5.3.3 - Alberta's Tomorrow Project Cohort Characteristics

Recruitment and baseline data collection took place between 2000 and 2015 with biospecimen collection beginning in 2009. Details on cohort characteristics, recruitment, and design have been previously published (Ye et al., 2017).

ATP collected self-reported baseline and follow-up questionnaire data on demographics and health risks. New incident cases of cancer were confirmed through linkage to Alberta Cancer Register (ACR). Case and control samples were matched in 1:1 design based on age (+/- 2 years), sex, and smoking status (current, former, never). In total 414 saliva samples were identified from ATP's biospecimen repository based on non-cancer controls and cases of breast (n=102), prostate (n=76), or colon cancer (n=29) that were diagnosed after saliva sample collection. The median length of time between saliva collection and cancer diagnosis along with other relative metadata can be found in Table 3.

5.3.4 - Oral Microbiome 16S rRNA Gene Sequencing

Samples from the Atlantic PATH cohort were processed as previously described (Nearing et al., 2020). Frozen saliva samples were stored at -80C and then thawed at room temperature and aliquoted into 96 well plates. In a biosafety cabinet using standard sterile techniques DNA was extracted using a QIAamp 96 PowerFecal QIAcube HT kit following the manufacturer's instructions using a TissueLyser II and the addition of Proteinase K at the indicated optional step. Sequencing was done at the Integrated Microbiome Resource at Dalhousie University. PCR amplification of the V4-V5 16S rRNA gene region was done using V4-V5 fusion primers (515FB - 926R) and a high-fidelity Phusion polymerase. A total of 25 cycles of PCR were done: denaturing at 98°C, annealing at 55°C, and elongating at 72°C. Sequencing was then conducted using an Illumina MiSeq to produce 300-bp demultiplexed paired-end reads.

Samples from the Alberta's Tomorrow Project cohort were collected using an Oragene® DNA OG-250 kit manufactured by DNA Genotek. Samples were collected either in person at local study centers or saliva sample kits were sent to participants by regular postal mail with a return envelope included. Participants were instructed not to eat, chew gum, or smoke 30 minutes prior to providing a saliva sample. They were asked to spit into the container until the saliva reached the indicated level, screw the cap on, shake for 10 seconds and send the sample back through the mail. DNA from samples were then extracted using a DNA Genotek PrepIT PT-LP2 kit. After extraction samples were sequenced at the Integrated Microbiome Resource in the same manner as samples from the Atlantic PATH cohort.

5.3.5 - 16S rRNA Gene Sequence Processing

Processing of 16S sequencing data was conducted as previously described (Nearing et al., 2020). Primers were removed using cutadapt with default settings (M. Martin, 2011). Primer-free reads were stitched using the QIIME2 VSEARCH join-pairs plugin (Bolyen et al., 2019; Rognes et al., 2016). Stitched reads were then filtered with default settings using the QIIME2 plugin q-score-joined. Reads were then corrected into amplicon sequence variants using the QIIME2 plugin Deblur with a trim length of 360 bp, and one read set as the minimum number required to pass filtering (Amir et al., 2017). For each dataset examined, 0.1% of the mean sample depth was calculated and ASVs below this abundance across all samples within that dataset were removed. This is to keep in line with the previously described Illumina MiSeq bleed-through rate. ASVs were placed into the Greengenes 13_8 99% reference 16S rRNA tree using the QIIME2 fragment-insertion SEPP plugin (DeSantis et al., 2006; Janssen et al., 2018; MIRARAB et al., 2011). Rarefaction curves were generated for each dataset separately and a suitable rarefaction depth of 5,000 was chosen for the Atlantic PATH retrospective cohort and 3,000 for the Atlantic PATH prospective and ATP prospective cohorts. Rarefied data was used to generate both alpha and beta diversity metrics. Samples below these sequencing depths along with those that had no remaining case or control samples were removed from further analysis. Additionally, a single sample in the ATP prospective dataset was removed due to significant contamination during sample preparation. Final case-control sample numbers for each dataset that pass all quality filtering are presented in Tables 1-3. ASVs were then assigned taxonomy using a naive Bayesian QIIME2 classifier trained on the 99% Silva V138 16S rRNA database (Bokulich et al., 2018; Quast et al., 2013).

5.3.6 - Microbial Diversity Analysis

Alpha and beta diversity metrics were generated using the QIIME2 command "core-metrics-phylogenetic" with the previously described rarefaction values and phylogenetic tree. Diversity matrices were then exported into R and analyzed between case and control samples for each separate cohort and study design. Alpha diversity between case/control samples were examined using linear models with the inclusion of an "extraction_run" covariate for the Atlantic PATH retrospective samples due to the large number of batch extractions and amount of time passed between sample extractions. In total four different alpha diversity metrics were investigated: Faith's phylogenetic diversity, Shannon diversity, evenness, and richness. An alpha value of 0.05 was chosen as our significance threshold before conducting any statistical analysis. Violin boxplots were generated using ggplot2 while jitter points were added using the R package ggbeeswarm (Wickham, 2009).

Beta diversity metrics were compared using a PERMANOVA test between case samples and case matched control samples for each cancer type within each cohort and study design using the 'adonis2' function within the vegan R package (Dixon, 2003). In the case of the Atlantic PATH retrospective data we included the covariate extraction number due to the large number of different extraction runs and time taken between sample extractions for this dataset. An alpha value of 0.05 was chosen before any statistical testing was conducted. A secondary PERMANOVA analysis was also conducted between all cancer types within each cohort and study design followed by pairwise tests between cancer type and all controls within that cohort and study design. In total three different beta diversity metrics were examined: weighted UniFrac, unweighted UniFrac and Bray-Curtis dissimilarity. These three beta diversity metrics were visualized using principal coordinate analysis using the function cmdscale within an R programming environment and ggplot2. Ellipses were added to each sample type using the function 'stat_ellipse()'.

5.3.7 - Microbial Differential Abundance Analysis

Differential abundance analysis was conducted using four different tools developed to analyze microbiome data: ALDEx2 (Fernandes et al., 2014a), ANCOM-II (Kaul et al., 2017; Mandal et al., 2015), corncob (B. D. Martin et al., 2020), and MaAsLin2 (Mallick et al., 2021). These tools range in their consistency and power to detect differences between groupings and should give a broad range on the ability to detect differentially abundant taxa (Nearing et al., 2022). Each tool was run at both the ASV and genus taxonomic levels. All tools were run comparing taxonomic abundance against case versus control status. Each tool was run separately for each cancer type, cohort, and study design. During the examination of the Atlantic PATH retrospective dataset we also included DNA extraction as a covariate due to not all samples being extracted at the same time. For all tools, taxa that were not found in at least 5% of samples were removed from consideration. Filtered p-values were then corrected for false discovery using Benjamini–Hochberg correction (Benjamini & Hochberg, 1995).

ALDeX2 analysis was run using default settings and general linear models. This includes using a center-log-ratio transformation, and 128 Monte Carlo samplings to generate probability distributions from the observed count data.

ANCOM-II was run using scripts available at: https://github.com/FrederickHuangLin/ANCOM-Code-Archive . Genus and ASV abundance tables were first processed using the function “feature_table_pre_process”. The main grouping variable of interest during pre-processing and the determination of structural zeros was case versus controls. A value of 0.05 was used to determine outlier zeros and outlier values. Pre-processed tables were then passed into the main ANCOM function with the inclusion of DNA extraction batch as a covariate when examining retrospective PATH data. Significance was determined using a percentage cutoff of 70% for the w statistic.

Corncob was run by first importing taxonomic abundance tables and their corresponding metadata into phyloseq objects (McMurdie & Holmes, 2013). The function differentialTest was then run using the above phyloseq object with the “wald” test option. The phi formula was set to match the phi-null formula to control for differences in variability across sample groupings.

MaAsLin2 was run using default settings and an arcsine transformation. Case versus control was used as a fixed effect. In the case of the PATH retrospective dataset an additional fixed effect of the DNA extraction batch was included.

5.3.8 - Random Forest Model Training

Random Forest models were trained and used to classify case and control samples from each dataset and study design. Training and classification were done using 100-repeat-5-fold cross validation. In the retrospective dataset control samples were randomly downsampled within each fold training session to avoid unbalanced model training and biasing data within the hold-out fold. In all datasets taxon found in less than 5% prevalence were filtered out prior to model training. After training and cross validation, the mean number of votes on each hold-out set across all repeats was then calculated. Receiver operator curves were constructed using pROC and confidence intervals were estimated using 2000 bootstrap replicates (Robin et al., 2011). Variable importance was calculated from models trained on the entire dataset by determining the difference in the out-of-bag prediction error rate after the variable of interest was permuted.

5.4 – Results

5.4.1 - Investigation of the Oral Microbiome in Retrospective Cases of Breast, Prostate, and Colon Cancer

First, oral microbiome diversity trends between case and control samples were examined within retrospective cases of breast, prostate, and colon cancer within the Atlantic PATH cohort (Table 5.1). In total we examined four different alpha diversity metrics: richness, Shannon diversity, Evenness, and Faith's Phylogenetic Diversity while controlling for DNA extraction batch. Investigation into these four metrics did not show any differences in alpha diversity between case and non-cancer controls in breast, prostate, or colon cancer ($p > 0.05$) (Figure 5.1, Supplemental Figure 1). Subsequently, we also compared three different beta diversity metrics, two that consider weighted abundances; weighted UniFrac and Bray-Curtis dissimilarity and one that considers presence/absence; unweighted UniFrac. When comparing cases of each cancer type to matched non-cancer controls, we found no significant differences in any weighted beta diversity metrics (PERMANOVA, $p > 0.05$) (Figure 5.1, Supplemental Figure 2). However, we did see evidence of visual clustering in PCoA's generated from weighted UniFrac and Bray-Curtis dissimilarity when comparing colon cancer cases to non-cancer controls (Figure 5.1, Supplemental Figure 2C) (PERMANOVA: $p=0.107$, $p=0.124$; respectively). While comparing unweighted Unifrac distances we did find a significant difference between breast cancer cases and controls ($r^2=0.007$, $p=0.008$), (Supplemental Figure 2A).

Table 5.1: ***Atlantic PATH cohort characteristics for retrospective cases of breast, prostate, and colon cancer.***

Cancer Type	**Breast Cancer**		**Prostate Cancer**		**Colon Cancer**	
Case vs. Control	**Case**	**Control**	**Case**	**Control**	**Case**	**Control**
Number of samples	54	218	24	92	11	47
Sex (% female)	100%	100%	0%	0%	53%	53%
Mean Age	57.5 (8.15)	55.6 (8.16)	60.6 (5.92)	57.7 (6.36)	59.5 (10.2)	56.9 (9.21)
Mean BMI	29.0 (5.01)	28.3 (4.72)	29.6 (3.27)	28.8 (3.49)	28.9 (4.16)	28.5 (4.13)
% Current Smoker	0	0	0	0	0	0
Median Time Since Diagnosis	6	N/A	4	N/A	5	N/A

To investigate whether we were missing an effect of cancer status due to the passage of time since diagnosis, we correlated each alpha diversity metric to this variable (Supplemental Figure 3**)**. We found no significant relationships except for CRC which showed a positive association between time passed since diagnosis and alpha diversity (rho=0.62, p=0.04). We also re-examined samples that were within six years of diagnosis to see whether more significant microbiome effects were present closer to cancer diagnosis. Examining these samples showed no major differences to our original analysis apart from a significant decrease in richness in CRC samples compared to matched controls (p=0.011) (Supplemental Figure 4).

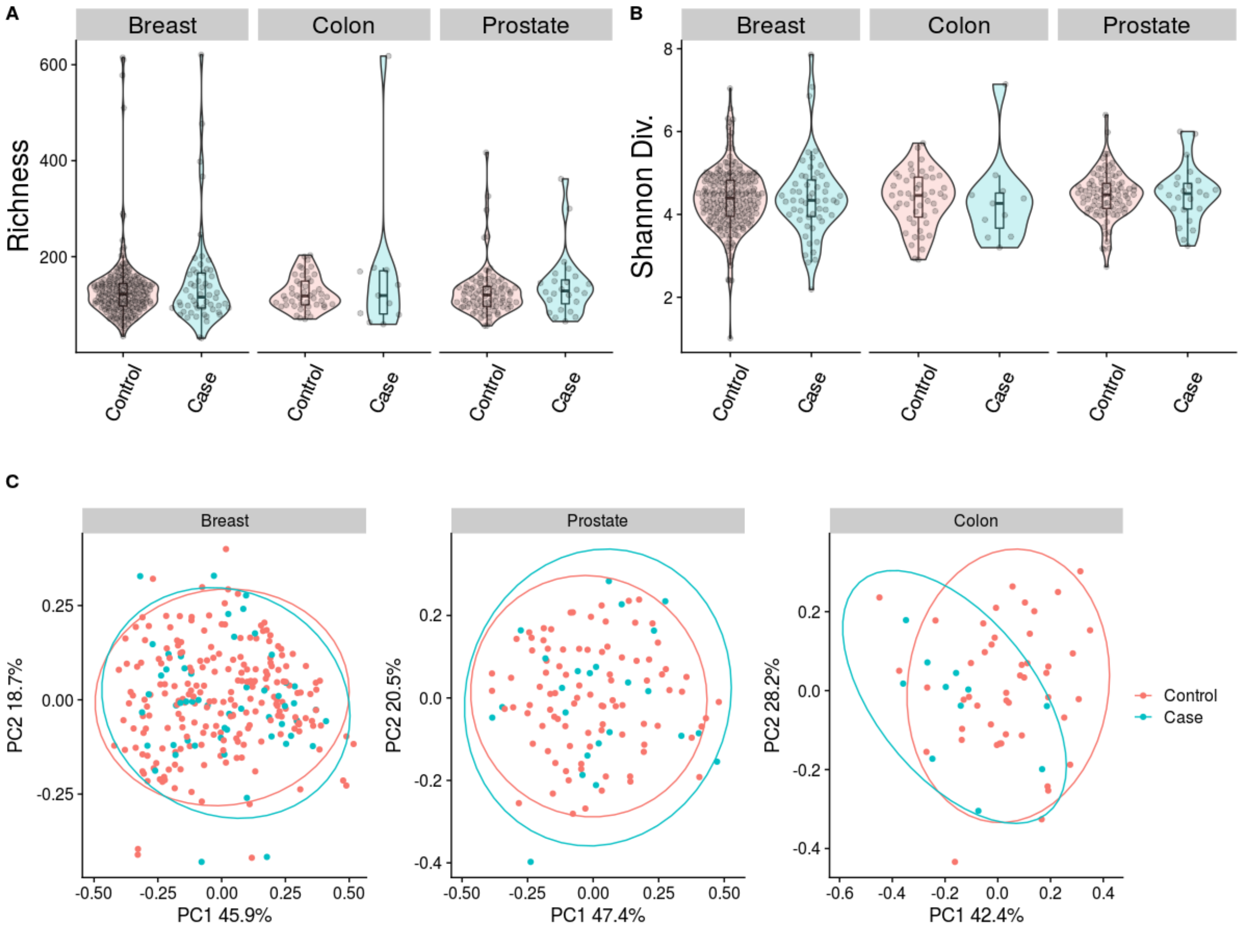


Figure 5.1: **Oral microbiome diversity metrics of retrospective cases of breast, colon, and prostate cancer in the retrospective Atlantic PATH cohort.** *Comparing microbial diversity of non-cancer matched controls to case samples of retrospective prostate, colon*

and breast cancer showed no significant differences in alpha diversity as measured by richness (a), and Shannon diversity (b). Further examination into beta diversity metrics showed no significant differences in any cancer type (c) although visual clustering was identified in colon cancer (PERMANOVA; p=0.107).

After comprehensively examining samples for differences in diversity we decided to conduct DA analysis to identify genera or ASVs that might be associated with having previously been diagnosed with cancer. Across all cancers examined we found a total of 25 genera and 30 ASV's associated with one or more cancer diagnoses (Figure 5.2, Supplemental Figure 5). In breast cancer we found one genus Rumminococcaecae UCG-014 that was detected as being significantly lower in relative abundance in breast cancer samples by two separate DA tools (Figure 5.2). We also identified an additional 4 ASVs one of which belonged within the Rumminococcaecae UCG-014 genera (Supplemental Figure 5). The other three ASVs, two of which were classified within the *Capnocytophaga* genera and one within *Bergeyella* were all detected to be enriched in breast cancer cases (Supplemental Figure 5).

In prostate cancer, corncob identified 19 genera and 24 ASVs that are potentially differentially abundant between case and control samples (Figure 5.2, Supplemental Figure 5). However, no other tools detected these taxa, and the overall effect size of these differences were minor ranging from 1.001 - 1.180 in log10 relative abundance mean fold changes. Furthermore, we found inconsistencies between corncob's coefficient directionalities and the observed differences between mean relative abundances between case and control samples (Figure 5.2, Supplemental Figure 5).

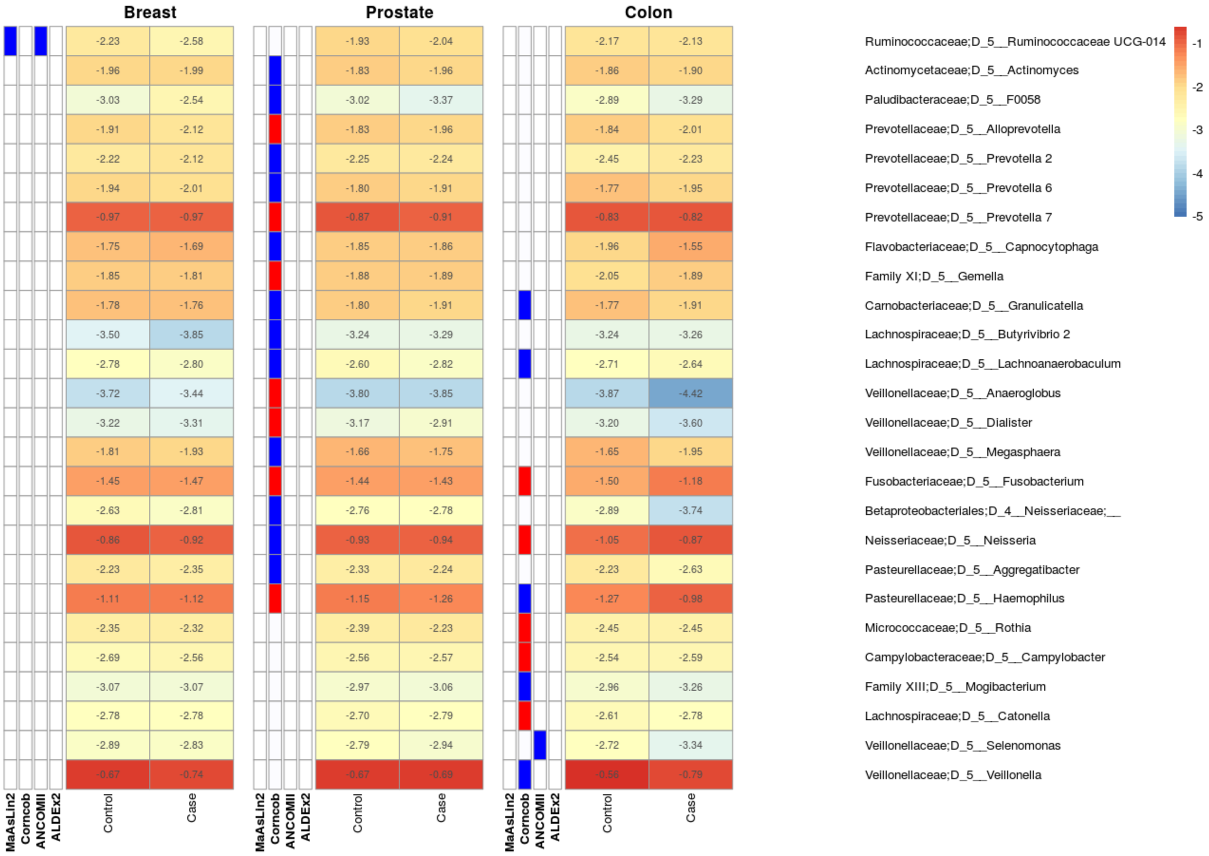


Figure 5.2: **Several genera are detected as differentially abundant in the oral microbiome of retrospective cases of prostate, colon, and breast cancer in the Atlantic PATH cohort.** *The heatmap is divided by cancer type where the first four columns represent the detection of significant associations by one of four tools: MaAsLin2, Corncob, ANCOM-II, and ALDEx2. Blue bars in the first four columns of each subgroup represent a detected increase in control samples while red bars represent a detected increase in case samples. The final two columns within each cancer sub grouping represent the log10 mean relative abundance of each genus with red representing higher abundance values and blue representing lower abundance values. Overall, corncob found the largest number of associated genera with a single genus in breast cancer also being detected by ANCOM-II.*

In colon cancer we identified 10 genera and 6 ASVs as being differentially abundant between retrospective case and non-cancer matched control samples. All these features were identified by corncob and a single ASV classified as Fusobacterium was additionally detected by ANCOM-II to be increased in colon cancer cases (Supplemental Figure 6). Inspection into the identity of this ASV at lower taxonomic levels using sortmeRNA (Kopylova, Noé, & Touzet, 2012) identified this ASV as potentially coming from the species *Fusobacterium periodonticum*.

Despite the relatively small taxonomic differences we found between case and control samples we were still interested in determining whether Random Forest classification models could pick up differences between case and control samples. Examining model performance on hold-out sets during cross validation showed that both breast cancer and colon cancer models performed best. Contrastingly, prostate cancer models performed at or below an AUC of 0.5 (Figure 5.3A-C). Although breast cancer models were only modestly better with AUCs ranging from 0.575 - 0.613 (Figure 5.3A).

Colon cancer models performed the best although large confidence intervals were observed due to low sample size. Furthermore, colon cancer models showed highly variable performance depending on feature normalization. Models built using center-log-ratio normalizations performed only slightly better than random expectation with ASVs having an AUC of 0.588 (0.370 - 0.806 95% CI), and genera having an AUC of 0.567 (0.377 - 0.756 95% CI). However, models built using relative abundances showed stronger results with models having AUCs ranging from 0.729 - 0.745 (Figure 5.3). Examining the top 10 most important features of each of these models showed relatively small decreases in accuracy for any single ASV/genera (Supplemental Figure 7-8).

Indeed, further inspection also showed that of these ASV/genera only 1 ASV overlapped with the previously identified differentially abundant taxon (Supplemental Figures 6-7). This ASV was classified into the genera *Veillonella* and upon further inspection, best aligned to *Veillonella atypica* (100% identity) within the SILVA V138 database.

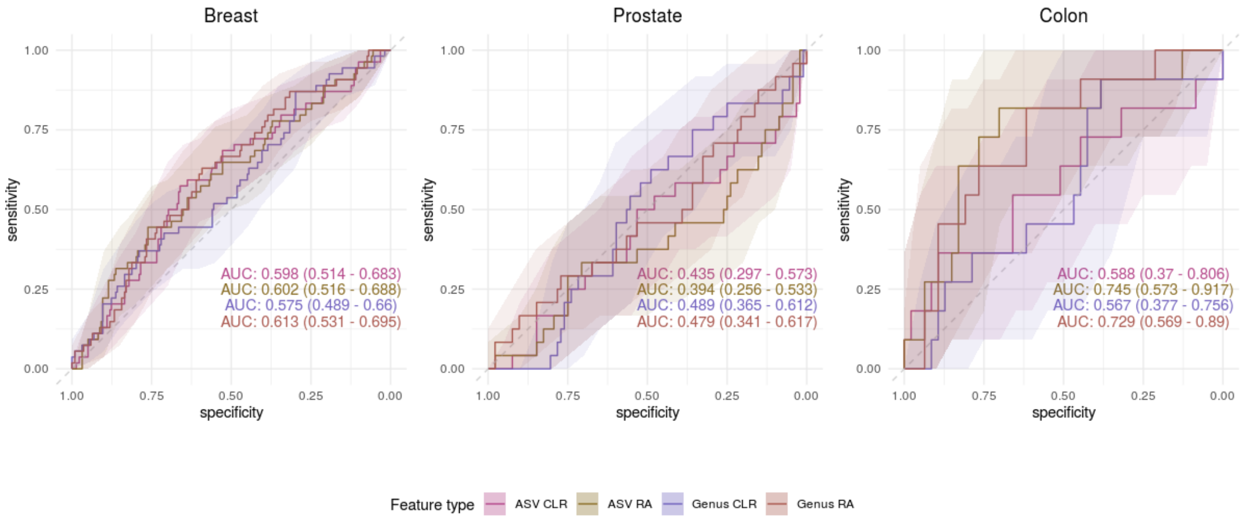


Figure 5.3: **Random Forest classification of retrospective cases of breast, prostate, and colon cancer based on microbial taxonomic composition in the Atlantic PATH cohort.** *Receiver operator curves (ROC) showing the specificity and sensitivity of the classification of non-cancer matched controls and retrospective cases of breast, prostate or colon cancer. Models were constructed using 100-repeat 5-fold cross validation and hold-out performance was determined through taking the mean number of votes for each hold-out sample across all 100 repeats. Within each plot four different ROCs are*

represented showing the classification accuracy using ASVs or genera normalized with either total-sum-scaling or center-log-ratio abundance. Shaded areas represent 95% confidence intervals determined through 2000 bootstrap samplings.

5.4.2 - Investigation of the Oral Microbiome in Prospective Cases of Breast, Prostate and Colon Cancer

We next decided to investigate if compositional changes within the oral microbiome are present before the diagnosis of breast, prostate, and colon cancer. Like our retrospective analysis we first examined changes in overall microbial community structure by looking for differences in alpha and beta diversity between cancer cases and matched non-cancer controls (Tables 5.2-5.3). We found no significant differences in alpha diversity in either cohort using linear models comparing case vs. control in any of the four metrics (Figure 5.4, Supplemental Figures 9-10). Correspondingly we did not find any significant differences in beta diversity between breast, colon, or prostate cancer cases and matched non-cancer controls in weighted UniFrac, unweighted UniFrac, or Bray-Curtis dissimilarity (Figure 5.4. Supplemental Figures 11-12) (PERMANOVA p > 0.05).

Table 5.2: **Atlantic PATH cohort characteristics for the investigation of the oral microbiome in prospective cases of breast, colon, and prostate cancer.**

Cancer Type	Breast Cancer		Prostate Cancer		Colon Cancer	
Case vs. Control	Case	Control	Case	Control	Case	Control
Number of samples	54	54	28	28	10	10
Sex	100%		0%		70%	

Cancer Type	Breast Cancer		Prostate Cancer		Colon Cancer	
Case vs. Control	Case	Control	Case	Control	Case	Control
Mean Age	56.6 (7.74)	56.9 (8.38)	60.6 (4.43)	61.0 (4.93)	60.4 (7.88)	60.7 (8.25)
Mean BMI	26.9 (5.25)	26.9 (5.45)	28.3 (4.46)	28.1 (4.72)	27.9 (6.08)	28.3 (5.80)
% Current Smoker	1.85%	3.70%	0%	3.57%	0%	0%
Median Time Before Diagnosis	4	N/A	3.5	N/A	3	N/A

Table 5.3: ***ATP cohort characteristics for investigation of the oral microbiome in prospective cases of breast, colon, and prostate cancer.***

Cancer Type	Breast Cancer		Prostate Cancer		Colon Cancer	
Case vs. Control	Case	Control	Case	Control	Case	Control
Sex (% Female)	100%		0%		50%	
Number of samples	82	82	64	64	22	22

Cancer Type	Breast Cancer		Prostate Cancer		Colon Cancer	
Case vs. Control	Case	Control	Case	Control	Case	Control
Mean Age	57.7 (8.74)	57.7 (8.70)	63.2 (6.92)	63.2 (6.88)	60.0 (10.0)	60.0 (9.99)
Mean BMI	28.3 (6.00) 11 N/a	27.1 (5.70) 9 N/a	27.3 (4.3) 7 N/a	28.2 (4.93) 5 N/a	30.6 (6.94) 1 N/a	27.1 (4.49) 3 N/a
Current Smoker	1.22%	1.22%	15.6%	15.6%	9.09%	9.09%
Median Time Before Diagnosis	4.28	N/A	3.10	N/A	2.98	N/A

Like our previous retrospective analysis, we also examined whether the time between sample collection and diagnosis had a major impact on signal within the case samples. Spearman correlations showed no significant relationships in any of the alpha diversity metrics in both cohorts (Supplemental Figures 13-14) ($p > 0.05$). Furthermore, examining case samples that were collected within four years of diagnosis showed similar results except for prostate cancer in the Atlantic PATH cohort which showed a significant increase in Faith's phylogenetic diversity ($p=0.025$) and richness ($p=0.041$) in case participants (Supplemental Figures 15-16).

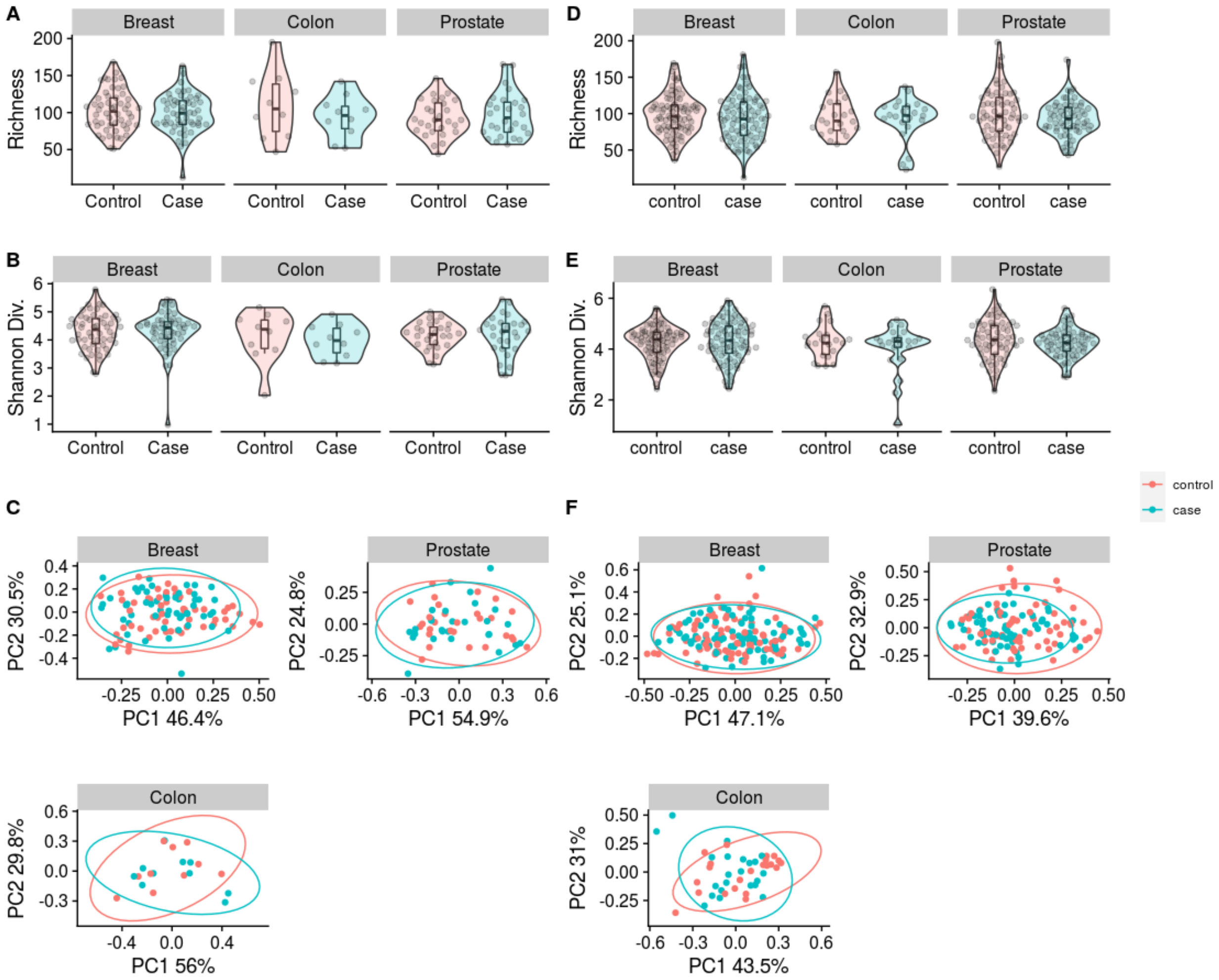


Figure 5.4: **Oral microbiome diversity metrics of prospective cases of breast, prostate and colon cancer in Atlantic PATH and ATP cohorts.**

Oral microbiome diversity analysis comparing non-cancer matched controls to prospective cases of colon breast and prostate cancer. Alpha diversity analysis as measured by richness (A) and Shannon diversity (B) as well as beta diversity measured as weighted UniFrac (C) showed no significant differences in the prospective PATH cohort. Similarly no significant differences in richness (D), Shannon diversity (E) or weighted UniFrac (F) in any cancer type in the ATP dataset.

After examining overall oral microbial community structure through various diversity metrics we were interested in determining whether there was any evidence of specific ASVs or genera being associated with disease status. In both prospective cohorts (Atlantic PATH, ATP) we found no genera being associated with disease status, however, we did find a small number of ASVs associated with disease status in both cohorts. In the Atlantic PATH cohort, we found an increase in the relative abundance of an ASV classified as *Alloprevotella rava* in prostate cancer (Supplemental Figures 17-18). We additionally found a decrease in the relative abundance of an ASV classified as *Streptococcus* in colon cancer (Supplemental Figures 17-18). Interestingly, none of these ASVs overlapped with those identified in the ATP dataset. Within the ATP cohort we detected two ASVs being decreased in relative abundance in colon cancer, although the significance of these taxa was mostly driven by outliers within control samples (Supplemental Figures 19-20). Within this cohort ANCOM-II also detected that the relative abundance of an ASV classified to an uncultured Stomatobaculum to be decreased in prospective breast cancer samples (Supplemental Figures 19-20).

As with our previous analysis we were interested in applying Random Forest models to each prospective cancer type to help identify whether disease signatures exist within the oral microbiome. Separate models were generated for each cancer type and cohort. Overall, models for breast cancer performed poorly in both cohorts with accuracies similar to random classification in both Atlantic PATH and ATP cohorts (Supplemental Figures 21-22). Interestingly, in the case of prostate cancer three of four models in the Atlantic PATH cohort performed slightly above random classification with AUCs ranging from 0.602 - 0.665 (Supplemental Figure 22). However, in the ATP cohort all models performed at or below an AUC of 0.5 (Supplemental Figure 23) although it should be noted that 95% confidence intervals on these AUCs were large due to small sample sizes (Table 5.2-5.3).

Finally, models of prospective cases of colon cancer showed variable results between the two cohorts of interest. With models in the Atlantic PATH cohort showing low performance AUCs ranging from 0.380 - 0.620 (Figure 5.5, Supplemental Figure 21). Contrastingly, in the ATP dataset stronger classification accuracies were found when using center-log-ratio normalizations with the genera level model performing the best (AUC 0.717; 95% CI: 0.549 - 0.884) (Figure 5.5, Supplemental Figure 22).

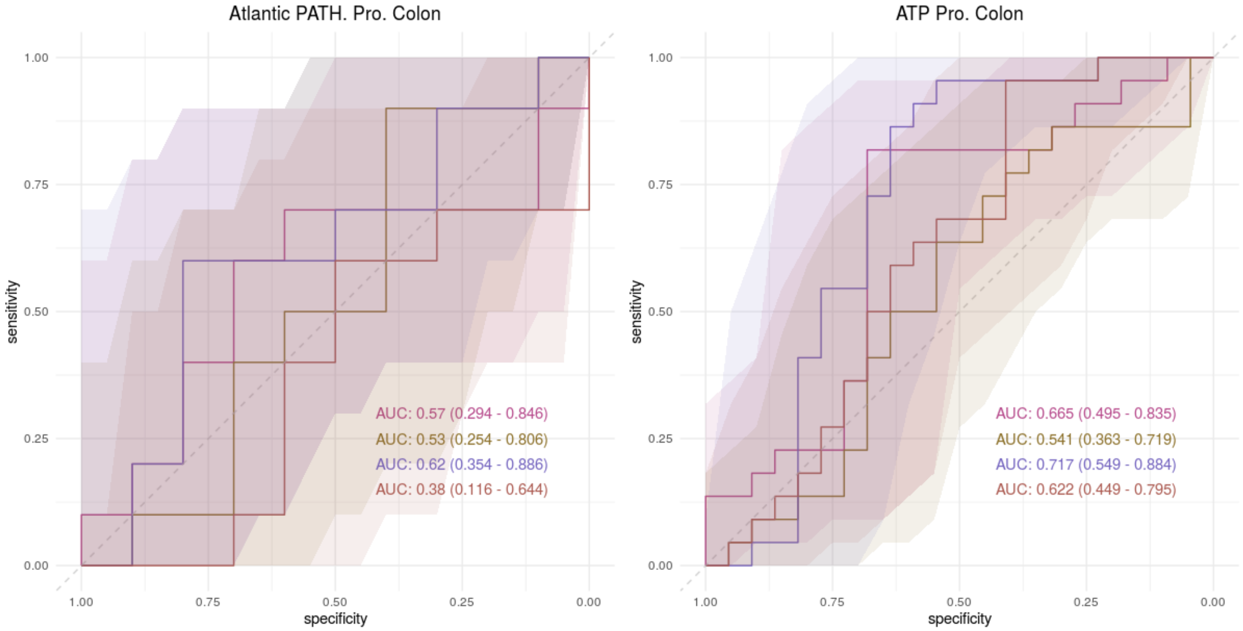


Figure 5.5: ***Random Forest Classification performance of prospective cases of colon cancer in the Atlantic PATH and ATP cohorts.*** *Receiver operator curves (ROC) showing the specificity and sensitivity of the classification of non-cancer matched controls and prospective cases of colon cancer in the PATH and ATP datasets. Models were constructed using 100-repeat 5-fold cross validation and hold-out performance was determined through taking the mean number of votes for each hold-out sample across all 100 repeats. Within each plot four different ROCs are represented, showing the classification accuracy using ASVs or genus normalized with either total-sum-scaling or center-log-ratio abundance. Shaded areas represent 95% confidence intervals determined through 2000 bootstrap samplings.*

Inspecting the top ten most important genera through out-of-bag permutation analysis within our CLR normalized colon cancer model showed no single genera as being particularly important to classification accuracy (accuracy decrease ranging from: 0.003 - 0.016). The most important genera within the model only decreased out-of-bag accuracy by 0.016 (SD; 0.002) although inspection of its CLR abundance did show a notable increase in case samples when compared to non-cancer controls. Inspection of other important genera within this model showed interesting CLR abundance patterns although as mentioned previously none were identified in our previous differential abundance analysis (Figure 5.6).

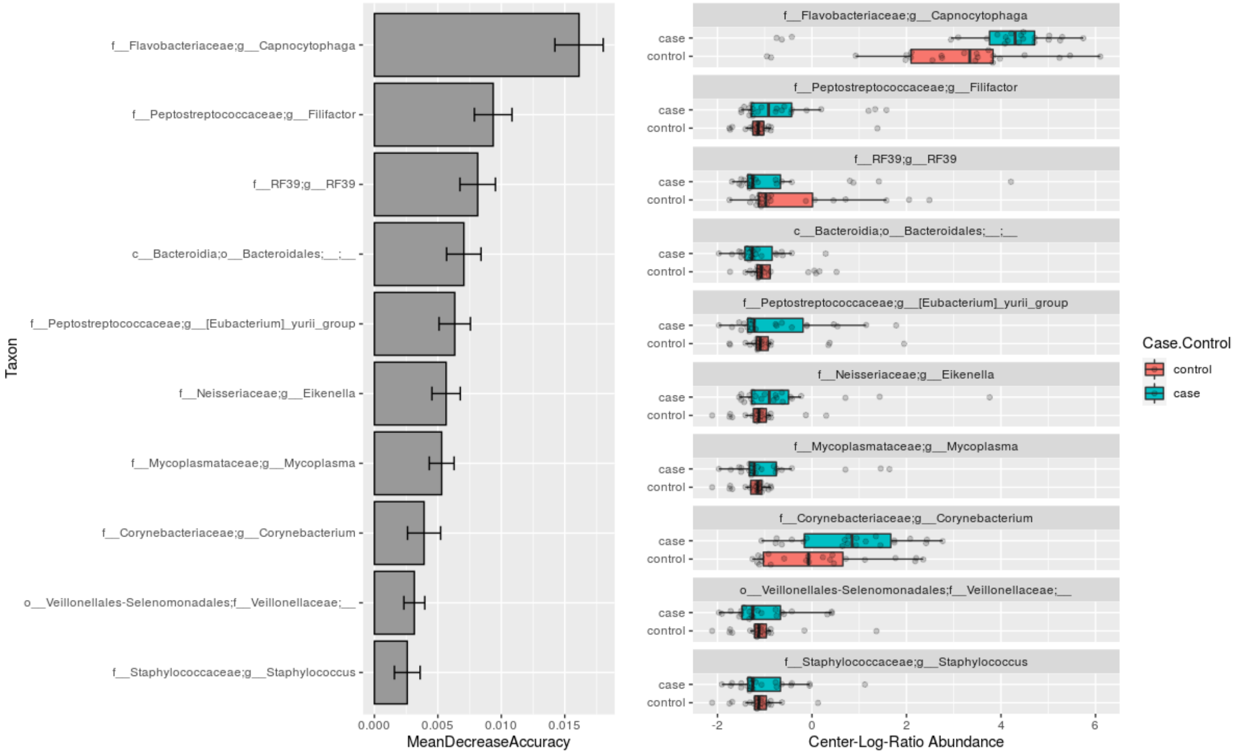


Figure 5.6: ***Feature importance of genus level center-log-ratio normalized Random Forest classification of prospective cases of colon cancer in the ATP dataset.*** *Feature*

importance was determined using out-of-bag permutation feature analysis. MeanDecreaseAccuracy represents the mean out-of-bag accuracy loss when that feature was randomly permuted across samples (A). Feature center-log-ratio abundance patterns are shown in panel (B) and show possible examples of interesting genera to further investigate in future studies.

5.5 – Discussion

Herein we examined the oral microbiome in the context of both retrospective and prospective cases of prostate, colon, and breast cancer in a population setting. Our analysis showed no significant changes in oral microbiome diversity in either retrospective or prospective cases of these cancers. Although we did find evidence of visual clustering of retrospective colon cancer cases when examining weighted beta diversity metrics. Investigating the relationships of individual taxon and cancer status showed some evidence of potential associations, although the majority were only detected by one of four DA tools indicating a low level of evidence. Accordingly, Random Forest classification of case samples and non-cancer matched controls showed relatively low classification accuracies with colon cancer showing the strongest signal in both retrospective and prospective analysis. Overall, our findings suggest that no large community changes exist in the oral microbiome of individuals with retrospective or prospective cases of prostate and breast cancer. Although a minor amount of evidence in our report does suggest there may be potential individual taxon relationships within these diseases. Contrastingly, through Random Forest modeling we did find some signal in retrospective cases of colon cancer and in one of our prospective colon cancer cohorts. Highlighting that future studies on prospective colon cancer cases are warranted.

Examining our results in breast cancer more closely showed strong concordance with previous work by Wang et al., who also found no changes in overall oral microbiome composition in United States individuals with breast cancer (H. Wang et al., 2017). These results contrast with a recent study by Wu et al., who identified differences in microbial diversity and the abundance of Porphyromonas and Fusobacterium (Z. Wu et al., 2022). This could be due to several reasons including the fact that these studies were conducting under highly different populations, as geographic differences have been shown to impact oral microbiome composition (J. Li et al., 2014). Unlike either of these studies, we did find evidence for a modest decrease in the relative abundance of Ruminococcaceae UCG-014 an uncultured genus that we previously linked to differences in height within healthy individuals of the same cohort (Nearing et al., 2020). Whether this taxon plays any role in disease status is unclear, however, due to this association not being recovered in either of our prospective cohorts it's likely that its association is limited to during or soon after cancer development. Similarly, we also identified an increase in two ASVs classified within the genus *Capnocytophaga* which were not detected in our prospective cohorts. Interestingly, within this genus, *C. gingivalis* has previously been associated with oral squamous cell carcinomas (Healy & Moran, 2019). Moreover, recent work has shown that supernatant from this species has the potential to enhance cellular invasion and migration of tumor cells (W. Zhu et al., 2022). Highlighting that this species may play a role in disease development or progression.

Investigating signal within the oral microbiome of breast cancer individuals using Random Forest modeling showed relatively little signal with accuracies in retrospective cases only slightly better than random assignment. Based on these results, we believe it is unlikely that the oral microbiome could be used as a biomarker to detect the risk of breast cancer development within a population setting.

To the best of our knowledge this report is the first to examine the relationship of the oral microbiome and prostate cancer diagnosis. Similar, to breast cancer we found no large shifts in oral microbiome diversity in prospective or retrospective cases of prostate cancer. Although, we did see a possible time dependent effect in the Atlantic PATH cohort which was not found in our second ATP cohort. Whether these differing results are due to DNA extraction, regional differences, or simply a false positive discovery would require further investigation in future studies.

Despite not identifying any consistent differences in diversity, multiple ASVs and genera were identified by corncob to be associated with retrospective cases of disease. Unsurprisingly, comparing these results to those previously identified within the gut showed little overlap (Golombos et al., 2018; Liss et al., 2018; Matsushita et al., 2021). Additionally, none of these retrospective taxonomic relationships were recovered in our prospective datasets. These results suggest that these retrospective associations may be related to other broad lifestyle changes that occur after prostate cancer diagnosis. Indeed, several of these retrospective taxa associations were previously associated with various lifestyle, dietary and anthropometric measurements within healthy PATH participants (Nearing et al., 2020). This highlights that many microbes, even those potentially associated with disease, are often affected by multiple daily life factors some of which could be associated with disease diagnosis.

Accompanying these results, we saw little to no signal in our Random Forest prostate cancer classification models. Some signal was recovered from the best models trained on the Atlantic PATH cohort (AUROC 0.665); however, due to small sample sies (N=28) confidence intervals remained large. Moreover, this signal was not recovered in our additional ATP cohort. These results suggest that the oral microbiome is unlikely to be a strong biomarker of prostate cancer risk.

Colon cancer showed the strongest evidence for differences in oral microbiome diversity, though, none of these differences were detected as being significant. Despite this we did see evidence of visual clustering in weighted UniFrac profiles of retrospective cases of disease (p=0.102, p=0.124). This matches with previous work by both Flemer et al., and Wang et al., who found significant differences in oral microbiome beta diversity between healthy controls and individuals diagnosed with colon cancer (Flemer et al., 2017; Yao Wang et al., 2021). Further investigation of diversity metrics also showed a reduction in oral microbiome richness in individuals that were diagnosed within 6 years of sample collection. Interestingly, this conflicts with previous reports by Wang et al., who found increases in oral microbiome diversity within CRC patients (Yao Wang et al., 2021). However, it should be noted that this could be due to several different factors including those associated with differing treatment regimens, sample timing, or environmental exposures.

In retrospective cases of colon cancer, we found evidence of an increase in an ASV classified as *Fusobacterium peridonticum* by two different microbiome DA frameworks (corncob, ANCOM-II). The relative abundance of *Fusobacterium peridonticum* a close relative to *Fusobacterium nucleatum* has previously been identified as being increased in the oral microbiome of oral small cell carcinoma (C.-Y. Yang et al., 2018), head and neck cancer (Mougeot et al., 2021), and pancreatic cancer patients (H. Sun et al., 2020). Accordingly, whether the relative abundance of this oral microbe represents a specific connection to colon cancer is still in question as it may represent a broader connection to other events associated with cancer diagnosis. Finally, of the three cancers examined, colon cancer showed the most consistent signal in our Random Forest modeling, although substantial differences in classification accuracies were noted between our two prospective cohorts. This could have been due to a few factors including sample size differences, collection method, or possibly differences in health risk factors between Atlantic Canada and Alberta. Unfortunately, attempts to train Random Forest models on combined datasets were not successful due to different collection and DNA extraction approaches causing significant bias between studies (Nearing et al., 2021; Pollock et al., 2018).

One puzzling result we noticed from our colon cancer Random Forest models, was that the type of normalization used had large impacts on model accuracies and was not consistent between retrospective and prospective datasets. We found total-sum-scaling to perform the best in our retrospective cohort but found center-log-ratio transformation to perform better in our prospective ATP cohort. Whether this has biological significance is unclear and suggests that future models may be interested in testing several different data normalizations during model training.

Within our analysis we have also identified several limitations that should be noted when reviewing our results. The first is that in the case of both prostate and colon cancer our sample size numbers are small in all three datasets we examined (Table 1-3). These smaller samples most likely interfered with our ability to detect small differences in both community composition and individual taxa abundances. A second limitation of our study is the different extraction methods used for samples that came from the two different population cohorts, Atlantic PATH and ATP. This along with other technical variations such as differences in sample collection led to the need to conduct stratified analysis reducing our ability to valid signal between cohorts. Finally, we would like to acknowledge that our datasets were relatively homogenous, with our dataset being predominantly from white Canadians, with income and education levels above average

Canadian census data (Sweeney et al., 2017; Ye et al., 2017). As such this significantly limits our ability to identify distinct oral microbial signatures in groups that are disproportionately affected by cancer development.

Taken together our results indicate that the oral microbiome is unlikely to be an effective population-based risk marker for cases of prostate or breast cancer, although changes in specific bacterial abundances within these diseases may still exist. Contrastingly, in the case of colon cancer our work indicates that disease status is related to changes in the oral microbiome and may be useful as a risk marker for colon cancer development. Future studies should aim to evaluate when oral microbiome changes occur in prospective colon cancer cases to determine its suitability for risk stratification.

Chapter 6 – Discussion

The above chapters have highlighted and explored two major themes within my thesis. The first being the importance of computational tool choice during microbiome experiments and how they may impact biological conclusions. The second major theme examined the use of marker gene sequencing in large population cohorts to understand oral microbiome variation and the potential for microbial biomarkers of cancer. Both themes and works within have led to important insights into the identification of robust microbial biomarkers in sequencing data. For example, work in Chapters 2 and 3 showed that choice of computational methodology, can have large impacts on biological conclusions and therefore potential insights into biomarker identification. Highlighting the need to understand the consequences of choice when performing microbiome experimentation. Furthermore, in Chapters 4 and 5 we used salivary microbiome data to identify not only patterns of cancer association but also those associated with daily life factors. Emphasizing the importance of examining biomarkers in a multifaceted context and not in the isolation of a specific disease. Within this final chapter an overview of these topics will be discussed with the goal of identifying future steps and avenues toward more reproducible and thus robust conclusions derived from microbiome sequencing data. Following this I will also discuss the potential future directions of oral microbiome science in the context of understanding compositional variation and biomarkers within breast, prostate, and colon cancer.

6.1 – Moving Toward Robust Sequenced Based Microbiome Biomarkers

Exemplified throughout Chapters 2 and 3 is the importance of making choices in microbiome sequencing experiments. Indeed, often these choices are made arbitrarily or with little information on how they might impact the downstream conclusions that are drawn. While this may stem from the fact that no gold standard exists within the field, it could also be argued that no single choice could possibly cover all potential questions that could be answered within a microbiome study (McLaren, Willis, & Callahan, 2019). For example, some sampling devices may be well suited for sequence-based analysis but perform poorly in metabolomics or culturomics, two equally important techniques for the analysis of microbial communities. Decisions such as these, lead to difficult choices during study design necessarily leading to various trade offs.

There is also a question of what the study at hand is aiming to address. Some researchers may be interested in building strong supervised classification models where the true biological underpinnings are not as important as getting consistent results (Marcos-Zambrano et al., 2021). In this case perhaps biases that lead to perturbed observations may not be as important if they do not impact model classification and consistency. On the other hand, if the goal of the study is to examine the underlying biology these biases could be of significant importance to help address the potential of being led astray (Pollock et al., 2018). For example, the correct classification of 16S rRNA gene sequences is of vital importance to answering biological questions as it tells us what potential taxa to further validate. However, if the goal is simply to classify samples as cancer versus non-cancer this misclassification may not be as important if its consistent between samples. This points us toward the vital need to understand how the choices we make as researchers impact underlying observations. By doing so we can begin to understand how suitable they may be toward answering the question at hand.

Unfortunately, examining how choices in microbiome research impact biological conclusions can be difficult, cumbersome, and in some cases impossible due to the numerous combinations available. For these reasons, we cannot reasonably expect researchers to test every single potential combination of choices within a study, especially those that are non-computational, due to limited sample biomass. However, we can extrapolate data from other studies to understand how actions such as protocol choice can vary between studies. Indeed, study to study bias stemming from computational methodology was one of the major focuses of this thesis. For example, in Chapter 2 we showed that weighted beta diversity metrics were similar among sequence denoisers while unweighted metrics tended to vary. While this result did not lead to novel biological information it critically showed the expectations we can make when reviewing literature or making decisions about or own studies. Likewise, in Chapter 3 we showed that DA method choice, can have profound effects on the number of significant taxa uncovered within a study. However, within this chapter we showed that despite tools showing variable performance, we were able to generalize certain characteristics such as ALDEx2 and ANCOM-II being conservative but consistent among DA choices. This will not only be important for assessing previous and future literature using these methods, but also in the identification of potential solutions.

In Chapter 3 we suggested the use of a consensus approach due to our analysis as well as previous literature failing to clearly identify a gold standard microbiome DA method (Calgaro et al., 2020; Knight et al., 2018; Pollock et al., 2018; S. Weiss et al., 2017).

Moreover, due to the computational nature of this problem we can make use of parallel computing and the fact that sequencing data can easily be reused (unlike biological samples) to run multiple tools at once. The suggested use of ensemble methods in differential analysis is not a new concept. In fact, multiple packages designed to analysis RNA sequencing data use the consensus between various differential expression pipelines to identify potential gene associations (Costa-Silva, Domingues, & Lopes, 2017; H.-S. Li, Ou-Yang, Zhu, Yan, & Zhang, 2022; Waardenberg & Field, 2019; Wolf, Epping, Andreotti, Reinert, & Semmler, 2021). The benefits of which are three-fold; the first being that researchers can easily determine how robust identified features are to methodology choice. The second being that genes identified can easily be compared to previous literature that may have used a myriad of different protocols. The final benefit is that these methods have consistently shown to outperform the usage of a single differential expression pipeline.

Nevertheless, consensus approaches are not without their own faults. The need to run multiple analysis on the same data can become cumbersome especially when various tools differ in input and output. Furthermore, consensus approaches need to be carefully designed and interpreted as the usage of similar models, that given equally wrong results, can led to false confidence of importance. Indeed, an example of this would be the testing of multiple methods of similar origin, such as the two limma voom methods used in Chapter 3. This highlights the need to characterize and identify methods of high quality and varying characteristics before apply consensus strategies. Fortunately, in the case of microbiome DA analysis this step has already been completed within this thesis. Going forward the development of an easily interpreted consensus pipeline could have profound effects on the reproducibility and interpretability of microbiome DA results.

Another important consideration that needs to be made when using consensus approaches is the importance of understanding why tools may potentially vary in the results they give. For example, in Chapter 3 we identified tools that examine microbiome abundances in fundamentally different ways. For example, tools such as corncob are focussed on examining relative abundances in proportions, while others such as ALDEx2 examine relative abundances in comparison to the sequence geometric mean abundance (Fernandes et al., 2014a; B. D. Martin et al., 2020). While on the surface these tools ask similar questions, the normalizations and transformations made by them may favor specific scenarios or biological questions. Highlighting the fact that associations identified be a single DA method may not necessarily represent a false positive result but could be due to a slight difference in normalization. Nonetheless consensus approaches can help identify these cases and future work will need to be done to identify the biological scenarios that these situations represent.

The use of consensus approaches may not just stop at DA analysis and could also be applied to various other computational settings including those presented in Chapter 2. Indeed, consensus approaches in dealing with sequencing error may also prove to be a fruitful approach. While its clear based on the results of Chapter 2 that highly weighted sequences are consistent between major methods, error within sequences of low abundance remain difficult to assess. One way to potentially address this issue would be the usage of multiple approaches to assess the likelihood of those sequences being of true biological origin. Doing so may help address the biases introduced by different tools and improve the observed estimates of unweighted diversity and richness between different studies. Moreover, this approach would also give researchers better context before attempting to validate these sequences through other experimental procedures such quantitative PCR or culturomics.

Unfortunately, consensus approaches do not easily apply to non-computational steps within microbiome sequencing experiments. Indeed, choices such as sampling device, sample type, and DNA extraction method are all subject to specific resource requirements and the biomass available within a specimen. Due to this, various works have attempted to either standardize protocol choice or compare protocols across studies to interpret potential differences and extrapolate them to their own studies (Costea et al., 2017; Panek et al., 2018; Pollock et al., 2018; W.-K. Wu et al., 2019). For example, work by Costea et al., examined 21 different DNA extraction protocols on the same fecal samples to identify how they differed from one another. In their work, they not only identified bias in the ratio of Gram-positives to Gram-negatives but also in community diversity (Costea et al., 2017). Highlighting the potential for study-to-study bias due to differences in DNA extraction choice. In fact various studies examining different steps of microbiome sequencing experiments have been done to identify the depth of bias they introduce into a study (Nearing et al., 2021).

Overall, it has become clear that a significant amount of study-to-study bias exists within microbiome literature that cannot be fixed after DNA sequencing. This bias is often referred to as batch effects and hinders the ability to directly compare microbiome profiles from one study to another (Yiwen Wang & LêCao, 2020). In fact, batch effects can even create issues within the same study as exemplified in Chapter 5 of this thesis.

Indeed, the use of two different device collection kits and DNA extraction procedures introduced significant cohort bias leading to the inability to train and test Random Forest models across the Atlantic PATH and ATP prospective cohorts. This difficulty emphasizes the importance of consistency within microbiome studies and the need to deal with batch effects when sequencing data is not generated from the same protocols.

Removal of batch effects has become common practice in many RNAseq applications using various tools including ComBat, SVASeq and RUVSeq (Leek, 2014; Risso, Ngai, Speed, & Dudoit, 2014; Yuqing Zhang, Parmigiani, & Johnson, 2020). However, applying these methods to microbiome data has remained challenging due to its multivariate nature, high sparsity, and compositionality (Gloor et al., 2017). Despite this several promising methods developed specifically for microbiome data have recently been proposed. These methods may serve as important stepping stones toward the inclusion of multiple datasets in future studies and classification models (Dai, Wong, Yu, & Wei, 2019; Gibbons, Duvallet, & Alm, 2018; Ling et al., 2022). Future work examining salivary microbiome biomarkers using data from this study in conjunction with other cohorts may be interested in their application.

Without regard for study to study bias, sequenced based observations are still distorted from the ground truth, as microbes within samples can be detected at different efficiencies (McLaren et al., 2022, 2019). For example, DNA is preferentially extracted from Gram-negative bacteria due to their lack of a thick peptidoglycan layer. This is highlighted by the fact that including a mechanical lyses step during DNA extraction increases the ratio between Gram-positive and Gram-negative bacteria (Costea et al., 2017). Another example of differing detection efficiencies between taxa specific to 16S rRNA gene sequencing, is the fact that bacteria often have differing 16S rRNA gene copy numbers. This can lead to a biased detection toward those with larger copy numbers if not corrected through copy number estimation (Kembel, Wu, Eisen, & Green, 2012).

While these non-computational biases have not been directly addressed in this thesis, I believe they also contribute to high degree of variance in results between microbiome studies, as suggested by McLaren et al (McLaren et al., 2019). Unfortunately, none of the DA methods that were tested in Chapter 3 address these biases specifically but may interact with them in differing ways. Indeed, methods such as ALDEx2 and ANCOM-II attempt to address issues of compositionality in microbiome data but fail to address potential differences in the detection efficiency of different bacteria. Potentially these detection efficiencies are just one piece of the puzzle leading to the large variation in results between differing DA tools. Fortunately, since the publication of Chapter 3 highlighting the variance in microbiome DA methods, I have begun to contribute to work that attempts to improve the robustness of DA analysis in the face of differing taxon efficiencies (McLaren et al., 2022). Indeed, these methods based on the use of mock community calibrations and spike-in controls may potentially lead to more robust microbial biomarker findings.

Finally, I would like to highlight that all biological conclusions generated from sequencing data should eventually be validated through other experimental techniques. While it may not always be necessary if the end goal of an application is diagnostic, or classification based in nature, understanding the true underlying biology requires further steps after DNA sequencing. Indeed, validation using quantitative PCR, flow cytometry or culturing represents key steps toward understanding how robust results are from the various biases introduced during a typical microbiome sequencing experiment. Moreover, validation of key community associations through a reductionist approach in model organisms can help identify key mechanistic relations. However, the use of model organisms does come with a significant cost in terms of resources and time. Highlighting the need to produce robust biological conclusions from sequencing data. A feat that can be difficult when the impact of our choices during sequencing experiments, are not clearly understood.

6.2 – Biomarker Detection in the Context of Microbial Community Variation

The human microbiome is a multifaceted community that is shaped by health status, environment, and various other aspects of human life. This multifaceted nature often makes it difficult to pin down the exact reason why specific microbial abundances and compositions exist. Indeed to this day, what can be defined as a health microbiome is often open to interpretation (Hill, 2020). For example, within the gut, it is generally assumed that a healthy gut microbiome is linked with higher microbial diversity (K. Johnson & Burnet, 2016). Yet, this general thought is not always true with some works indicating increased microbial richness in association with various health conditions (Castaño-Rodríguez, Goh, Fock, Mitchell, & Kaakoush, 2017; Sanapareddy et al., 2012; Wan et al., 2022). Moreover, a range of microbial ecotypes exist within various areas of the human microbiome, regardless of health status (Lloyd-Price et al., 2016). These findings highlight that while broad patterns of community composition exist there remains a significant amount of interpersonal variation, especially at the bacterial species and strain level (Huttenhower et al., 2012). A result highlighted and discussed within Chapter 4 of this thesis.

In fact, within Chapter 4 myself and colleagues identified several anthropometric, dietary, and demographic features that are significantly associated with variation within the salivary microbiome of Atlantic Canadians. These findings along with others discussed within Chapter 4, highlight that microbial community variation is associated with countless variables in daily human life. Yet, when examining literature, often microbiome cohort analysis is done with the sole goal of identifying changes associated with a disease of interest. Indeed, regularly within these studies non-disease related variables are either controlled for in some fashion or completely ignored. Giving us a constrained picture that can potentially lead to difficulties in reproducibility and interpretation.

For example, recent work on the gut microbiome and autism has shown several convincing studies indicating large shifts in general community structure and a specific taxon. This work has led to various proposals that the gut microbiome may play a role in autism development or progression (Cryan, O'Riordan, Sandhu, Peterson, & Dinan, 2020). However, Yap et al., has recently argued that these shifts are due to the dietary constraints that many individuals with autism face (Yap et al., 2021), despite work within mouse models recapitulating autism like behaviors from fecal microbiota transplant studies (X. Lu et al., 2021; Sharon et al., 2019). These results leave us with an unclear picture of whether the gut microbiome does truly change or play any type of role in autism, a condition that has been under intensive investigation over the past decades (Cryan et al., 2020). However, more generally they highlight the difficulty in untwining the relationship between the human microbiome and disease in the face of complex daily life. Especially when we lack detailed knowledge on how daily life factors such as diet or body mass interact with the microbial communities that we harbor.

For these reasons, its clear that we require a better understanding of how various factors interact with microbial communities and to what degree they can cause variation. These were the main motivations behind Chapter 4, with the main goal of identifying how much salivary microbiome variation could be explained through the measurement of dietary, anthropometric, and demographic variables. Yet, when we examined the 41 different variables captured within our study, we were only able to explain 6-7% of the total variation found within the cohort. In fact, the total amount of variation we could explain from all significant biological variables was less than what was attributed to technical variation between DNA extraction batches. Again, not only reinforcing the importance of technical bias within microbiome studies but also showing that a large proportion of community fluctuation between individuals remains to be explained. Future work could improve our understanding of these contributing factors by including microbiome analysis as a primary outcome rather than a secondary analysis. This would allow for measures such as oral health and technical biases such as sample processing time to be included and highlighted within the study.

As previously mentioned within the discussion of Chapter 4, some of this variation can be attributed to multiple variables of daily life that were not captured by the Atlantic PATH study. Despite, collecting over 200 data points, the only measurement of oral health captured by questionnaire data was the amount of time passed since the participant's last dental visit. Indeed, even though saliva samples were collected, they were not initially anticipated to be used for microbial community surveys and thus missed potentially important questionnaire data. This was a clear limitation in Chapter 4 of our study and highlights the difficulty in designing cohorts that can adequately address all possible measures that may be related to human microbiome variation. In fact, the required sample size and cost of measuring all these potential factors are infeasible for most studies. However, as we move forward in microbiome research and continue to uncover large factors of variation, we can begin to better incorporate them within studies and understand how they interact with various human phenotypes.

Finally, I would like to acknowledge that the DA analysis done in Chapter 4 was conducted with a single tool corncob. This section of my thesis was completed before the in depth DA analysis conducted in Chapter 3. As such while corncob is an appropriate tool for identifying taxonomic associations and performed relatively well in our own analysis, it was found at times to identify false positives. Future improvements to the DA analysis conducted in Chapter 4 could have included the use of a consensus approach as suggested in Chapter 3 of this thesis. Doing so would have not only allowed us to better identify which taxonomic associations are robust to tool choice. Additionally, the use of a consensus approach with the inclusion of CoDa tools could improve our ability to identify taxa that are changing in abundance across groups when compared to the average microbe or across all potential microbe ratios. By taking this approach we could have improved the ability to compare these findings to currently published literature and may have found greater overlap with findings that have been reported previously.

6.3 – Using the Salivary Microbiome to Detect Breast, Prostate, and Colon Cancer

One of the main goals of this thesis was to identify potential oral microbiome biomarkers of prostate, colon, and breast cancer. In Chapter 5 we explored this potential using both a retrospective and prospective case control design from two separate regional Canadian cohorts. Overall, we found relatively little signal within the oral microbiome from individuals with breast and prostate cancer. Contrastingly, we did find evidence to support previous work suggesting a potential connection between the oral microbiome and colon cancer. However, our sample size was low and future work will need to be to confirm this. In this final section of Chapter 6 we will expand the discussion on the potential for future work within the salivary microbiome and these three cancers. I will highlight the difficulty in uncovering potential relations between microbes and cancer and discuss the various difficulties associated with developing microbiome-based diagnostics.

One of the primary motivators behind Chapter 5's work was to identify oral microbiome profiles or individual microbes associated with the risk of cancer development. Within this work we failed to identify any significant community shifts in both retrospective and prospective cases of breast and prostate cancer. However, we did identify several taxonomic associations in retrospective cases of disease and a handful in prospective cases. Although most of these significant associations were only identified by a single DA tool, they could potentially represent interesting biological phenomena to follow up on. Yet, whether they truly represent a hypothesises that is worth the resources required to follow up on remains a key question. Especially since retrospective associations could not only present as relations with disease but also with lifestyle changes directly caused by disease diagnosis.

Defining causal relationships or even strong associations between microbes and cancer can be extremely difficult. For example, as highlighted within the introduction of Chapter 5, despite (pks+) *E. coli* being strongly associated with colon cancer it has yet to make any significant clinical impact in colon cancer treatment or detection. Highlighting the significant resource costs required to investigate microbe cancer relations and the cost of following an unpromising result. For these reasons it will be important to validate findings within Chapter 5 before further attempting more costly experimental designs.

One potential way of further validating sequenced based results is with quantitative PCR. Although, in the face of numerous potential associations from multiple tools it can become difficult to determine which to prioritize. One potential solution is by choosing taxonomic associations that were either detected using multiple DA tools or those associated with large effect sizes. For example, Rumminococcaecae UCG-014 was identified by two separate DA tools in breast cancer and could warrant further investigation. Additionally, the relation between *Alloprevotella rava* and prospective cases of prostate cancer may represent an interesting case to follow up. One difficulty of this approach is the creation of taxa specific primers; however, shotgun metagenomics of high abundance samples or analysis of already assembled genomes could aid in species specific gene detection.

Outside of specific study validation a further difficulty in detecting microbial associations with cancer development and progression, is the inconsistent relation between microbial presence and disease development (Blaser, 2008). Indeed, often microbes associated with cancer can also be found in healthy individuals. For example, work by Hooi et al., estimated that *Helicobacter pylori* a well-known bacterial carcinogen colonizes close to 4.4 billion individuals across the world (Hooi et al., 2017). However, only 3% of those colonized with *H. pylori* will develop gastric cancer during their lifetime (Uemura et al., 2001). This highlights the potential for highly prevalent microbes within the human body to lead to cancer development under some but not all circumstances. Indeed, there are numerous reasons as to why detection of microbe-cancer associations could fail to be identified through sequencing. These include the high prevalence of the microbe but low incidence of the disease, lack of study power due to low sample number, or strain heterogeneity. Underlining while the results in Chapter 5 show that the salivary microbiome is unlikely to possess microbes that are consistently associated with cancer, it does not rule out all possibilities of oral microbes and breast/prostate cancer association.

On the other hand, despite low sampling numbers colon cancer showed several points of evidence for a potential relation between the salivary microbiome and disease development. Our results along with previous literature point toward the possibility of future salivary microbiome diagnostic for colon cancer (Flemer et al., 2017; Komiya et al., 2019; Thomas et al., 2019; Yao Wang et al., 2021; Y. Yang, Cai, Shu, et al., 2019). However, there remains several roadblocks ahead before the true potential of an oral microbiome-based colon cancer diagnostic can be recognized.

First, its critical that diagnostic models trained on microbiome data are applicable across a broad spectrum of individuals and are not overfit toward any specific cohort. Indeed, overfitting models to specific cohorts can led to inflated accuracies and give unclear estimates or real-world performance. A common problem that has been identified in microbiome research (Quinn, 2021). One way of addressing this is by training classification models across multiple datasets. While study-to-study bias can hinder these efforts work within the gut microbiome and colon cancer has shown its applicability. For example, work by Thomas et al., has shown that machine learning models classifying colon cancer that are trained on metagenomic shotgun data can be highly accurate even across differing studies and populations (Thomas et al., 2019); underscoring its future potential as a diagnostic. Perhaps similar approaches could be applied to salivary microbiome data as more datasets are published. However, just as important as cross study validation, will be the ability to identify model performance in underrepresented populations that are not typically represented in larger cohort studies. For example, within my own studies both cohorts that we collected data from were overrepresented by high income white Canadians with above average educations. Whether classification models from this dataset are transferable to other individuals remains to be seen and should be prioritized as we move closer to the use of microbiome data within the clinic.

A second roadblock that exists for microbiome-based diagnostics is the specificity of classification models toward the disease of interest. This may be hindered due to the multivariate nature of microbiome data, especially under circumstances where diseases show similar community shifts. For example, the reduced relative abundance of *Akkermansia muciniphilia* within the gut has been attributed to several health conditions suggesting its general importance to metabolic health (Derrien, Belzer, & de Vos, 2017). These general shifts highlight the importance of ensuring robust model performance in populations where multiple health conditions are present. To address this, future work may be interesting in specifically testing model performance against various disease cohorts. This will not only highlight disease specificity but also highlight the potential for similar community shifts across diseases.

Future work attempting to validate microbiome based diagnostic models should focus on identifying the relation between disease progression and microbial changes. While work within this thesis along with other prospective results suggest oral microbiota shifts before diagnosis, its unclear when and how early these markers may surface with respect to disease development (Y. Yang, Cai, Shu, et al., 2019). For example, our work varied in timing between sampling and diagnosis ranging between 0 and 8 years. Representing a highly variable range, which might differ in oral microbiome signal. However, unfortunately, due to the small size of our prospective cohorts we were unable to realistically examine whether classification improved when time between sampling and diagnoses decreased. Answering this question represents a critical step in the development of novel microbiome based diagnostic tools.

In closing research into the human microbiome has the potential to better human health through a multitude of different avenues. Whether it by through the use of carefully constructed probiotics, fecal transplants, diagnostics, or dietary recommendations. However, its clear that in the infancy of sequence-based microbiome research we have faced several challenges regarding reproducibility and the translation of findings from one study to another. My hope is that by better understanding and considering how the choices we make as researchers impact our biological conclusions, we can begin to derive more generalizable knowledge. Whether that be how daily life impacts the human microbiome composition or the relation between disease and specific taxa.

References

Adams, S. (2016). Estrobolome disparities may lead to developing biomarkers that could mitigate cancer risk. *J Natl Cancer Inst*, *108*(8), djw130.

Aird, D., Ross, M. G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., … Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, *12*(2), R18. https://doi.org/10.1186/gb-2011-12-2-r18

Aitchison, J. (1982). The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, *44*(2), 139–177.

Alanee, S., El-Zawahry, A., Dynda, D., Dabaja, A., McVary, K., Karr, M., & Braundmeier-Fleming, A. (2019). A prospective study to examine the association of the urinary and fecal microbiota with prostate cancer diagnosis after transrectal biopsy of the prostate using 16sRNA gene analysis. *The Prostate*, *79*(1), 81–87.

Alkanani, A. K., Hara, N., Gottlieb, P. A., Ir, D., Robertson, C. E., Wagner, B. D., … Zipris, D. (2015). Alterations in Intestinal Microbiota Correlate With Susceptibility to Type 1 Diabetes. *Diabetes*, *64*(10), 3510 LP – 3520. https://doi.org/10.2337/db14-1847

Allaband, C., McDonald, D., Vázquez-Baeza, Y., Minich, J. J., Tripathi, A., Brenner, D. A., … Knight, R. (2019). Microbiome 101: Studying, Analyzing, and Interpreting Gut Microbiome Data for Clinicians. *Clinical Gastroenterology and Hepatology*, *17*(2), 218–230. https://doi.org/10.1016/j.cgh.2018.09.017

Allali, I., Arnold, J. W., Roach, J., Cadenas, M. B., Butz, N., Hassan, H. M., … Azcarate-Peril, M. A. (2017). A comparison of sequencing platforms and bioinformatics pipelines for compositional analysis of the gut microbiome. *BMC Microbiology*, *17*(1), 194. https://doi.org/10.1186/s12866-017-1101-8

Allen, B., Kon, M., & Bar-Yam, Y. (2009). A new phylogenetic diversity measure generalizing the Shannon index and its application to phyllostomid bats. *The American Naturalist*, *174*(2), 236–243.

Almeida-Santos, A., Martins-Mendes, D., Gayà-Vidal, M., Pérez-Pardal, L., & Beja-Pereira, A. (2021). Characterization of the Oral Microbiome of Medicated Type-2 Diabetes Patients . *Frontiers in Microbiology* , Vol. 12. Retrieved from https://www.frontiersin.org/articles/10.3389/fmicb.2021.610370

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Zech Xu, Z., … Knight, R. (2017). Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *MSystems*, *2*(2). Retrieved from http://msystems.asm.org/content/2/2/e00191-16.abstract

Andersen, G. L., DeSantis, T. Z., Liu, Z., & Knight, R. (2008). Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Research*, *36*(18), e120–e120. https://doi.org/10.1093/nar/gkn491

Bahl, M. I., Bergström, A., & Licht, T. R. (2012). Freezing fecal samples prior to DNA extraction affects the Firmicutes to Bacteroidetes ratio determined by downstream quantitative PCR analysis. *FEMS Microbiology Letters*, *329*(2), 193–197. https://doi.org/10.1111/j.1574-6968.2012.02523.x

Baker, J. L., Morton, J. T., Dinis, M., Alvarez, R., Tran, N. C., Knight, R., & Edlund, A. (2021). Deep metagenomics examines the oral microbiome during dental caries, revealing novel taxa and co-occurrences with host molecules. *Genome Research* , *31*(1), 64–74. https://doi.org/10.1101/gr.265645.120

Baxter, N. T., Ruffin, M. T., Rogers, M. A. M., & Schloss, P. D. (2016). Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Medicine*, *8*(1), 37. https://doi.org/10.1186/s13073-016-0290-3

Beck, M. (2017). *ggord: Ordintation Plots with ggplot2.* https://doi.org/10.5281/zenodo.3828862

Belstrøm, D. (2020). The salivary microbiota in health and disease. *Journal of Oral Microbiology*, *12*(1), 1723975. https://doi.org/10.1080/20002297.2020.1723975

Belstrøm, D., Holmstrup, P., Nielsen, C. H., Kirkby, N., Twetman, S., Heitmann, B. L., … Fiehn, N.-E. (2014). Bacterial profiles of saliva in relation to diet, lifestyle factors, and socioeconomic status. *Journal of Oral Microbiology*, *6*(1), 23609. https://doi.org/10.3402/jom.v6.23609

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B (Methodol)*, *57*.

Bhushan, B., Yadav, A. P., Singh, S. B., & Ganju, L. (2019). Diversity and functional analysis of salivary microflora of Indian Antarctic expeditionaries. *Journal of Oral Microbiology*, *11*(1), 1581513. https://doi.org/10.1080/20002297.2019.1581513

Bik, E. M., Long, C. D., Armitage, G. C., Loomer, P., Emerson, J., Mongodin, E. F., … Relman, D. A. (2010). Bacterial diversity in the oral cavity of 10 healthy individuals. *The ISME Journal*, *4*(8), 962–974. https://doi.org/10.1038/ismej.2010.30

Blaalid, R., Kumar, S., Nilsson, R. H., Abarenkov, K., Kirk, P. M., & Kauserud, H. (2013). ITS1 versus ITS2 as DNA metabarcodes for fungi. *Molecular Ecology Resources*, *13*(2), 218–224. https://doi.org/https://doi.org/10.1111/1755-0998.12065

Blaser, M. J. (2008, June). Understanding microbe-induced cancers. *Cancer Prevention Research*, Vol. 1, pp. 15–20. https://doi.org/10.1158/1940-6207.CAPR-08-0024

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., … Gregory Caporaso, J. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome*, *6*(1), 90. https://doi.org/10.1186/s40168-018-0470-z

Bokulich, N. A., Rideout, J. R., Mercurio, W. G., Shiffer, A., Wolfe, B., Maurice, C. F., … Caporaso, J. G. (2016). mockrobiota: a Public Resource for Microbiome Bioinformatics Benchmarking. *MSystems*, *1*(5). Retrieved from http://msystems.asm.org/content/1/5/e00062-16.abstract

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., … Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, *37*(8), 852–857. https://doi.org/10.1038/s41587-019-0209-9

Bourgon, R., Gentleman, R., & Huber, W. (2010). Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences*, *107*(21), 9546 LP – 9551. https://doi.org/10.1073/pnas.0914005107

Brenner, D. R., Poirier, A., Woods, R. R., Ellison, L. F., Billette, J.-M., Demers, A. A., … Holmes, E. (2022). Projected estimates of cancer in Canada in 2022. *Canadian Medical Association Journal*, *194*(17), E601 LP-E607. https://doi.org/10.1503/cmaj.212097

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

Byrd, D. A., Vogtmann, E., Wu, Z., Han, Y., Wan, Y., Clegg-Lamptey, J., … Awuah, B. (2021). Associations of fecal microbial profiles with breast cancer and nonmalignant breast disease in the Ghana Breast Health Study. *International Journal of Cancer*, *148*(11), 2712–2723.

Calgaro, M., Romualdi, C., Waldron, L., Risso, D., & Vitulo, N. (2020). Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data. *Genome Biology*, *21*(1), 191. https://doi.org/10.1186/s13059-020-02104-1

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal*, *11*(12), 2639–2643. https://doi.org/10.1038/ismej.2017.119

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Cameron, M., T., M. J., A., M. C., R., T. L., Anupriya, T., Rob, K., & Karsten, Z. (2019). A Novel Sparse Compositional Technique Reveals Microbial Perturbations. *MSystems*, *4*(1), e00016-19. https://doi.org/10.1128/mSystems.00016-19

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*, 335. Retrieved from http://dx.doi.org/10.1038/nmeth.f.303

Caporaso, J. G., Lauber, C. L., Costello, E. K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., … Knight, R. (2011). Moving pictures of the human microbiome. *Genome Biology*, *12*(5), R50. https://doi.org/10.1186/gb-2011-12-5-r50

Cappellato, M., Baruzzo, G., & Di Camillo, B. (2022). Investigating differential abundance methods in microbiome data: A benchmark study. *PLOS Computational Biology*, *18*(9), e1010467. Retrieved from https://doi.org/10.1371/journal.pcbi.1010467

Castaño-Rodríguez, N., Goh, K.-L., Fock, K. M., Mitchell, H. M., & Kaakoush, N. O. (2017). Dysbiosis of the microbiome in gastric carcinogenesis. *Scientific Reports*, *7*(1), 15957. https://doi.org/10.1038/s41598-017-16289-2

Cavarretta, I., Ferrarese, R., Cazzaniga, W., Saita, D., Lucianò, R., Ceresola, E. R., … Briganti, A. (2017). The microbiome of the prostate tumor microenvironment. *European Urology*, *72*(4), 625–631.

Chakravorty, S., Helb, D., Burday, M., Connell, N., & Alland, D. (2007). A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *Journal of Microbiological Methods*, *69*(2), 330–339. https://doi.org/https://doi.org/10.1016/j.mimet.2007.02.005

Chase, J., Fouquier, J., Zare, M., Sonderegger, D. L., Knight, R., Kelley, S. T., … Caporaso, J. G. (2016). Geography and Location Are the Primary Drivers of Office Microbiome Composition. *MSystems*, *1*(2), e00022-16. https://doi.org/10.1128/mSystems.00022-16

Chen, B., Wang, Z., Wang, J., Su, X., Yang, J., Zhang, Q., & Zhang, L. (2020). The oral microbiome profile and biomarker in Chinese type 2 diabetes mellitus patients. *Endocrine*, *68*(3), 564–572. https://doi.org/10.1007/s12020-020-02269-6

Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., … Li, H. (2012). Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics (Oxford, England)*, *28*(16), 2106–2113. https://doi.org/10.1093/bioinformatics/bts342

Chen, K. L., & Madak-Erdogan, Z. (2016). Estrogen and microbiota crosstalk: should we pay attention? *Trends in Endocrinology & Metabolism*, *27*(11), 752–755.

Cheng, Y., Ling, Z., & Li, L. (2020). The intestinal microbiota and colorectal cancer. *Frontiers in Immunology*, *11*, 615056.

Chung, S.-D., Tsai, M.-C., Huang, C.-C., Kao, L.-T., & Chen, C.-H. (2016). A population-based study on the associations between chronic periodontitis and the risk of cancer. *International Journal of Clinical Oncology*, *21*(2), 219–223.

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., … Tiedje, J. M. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, *42*(Database issue), D633–D642. https://doi.org/10.1093/nar/gkt1244

Comeau, A. M., Douglas, G. M., & Langille, M. G. I. (2017). Microbiome Helper: a Custom and Streamlined Workflow for Microbiome Research. *MSystems*, *2*(1). Retrieved from http://msystems.asm.org/content/2/1/e00127-16.abstract

Costa-Silva, J., Domingues, D., & Lopes, F. M. (2017). RNA-Seq differential expression analysis: An extended review and a software tool. *PLOS ONE*, *12*(12), e0190152. Retrieved from https://doi.org/10.1371/journal.pone.0190152

Costea, P. I., Zeller, G., Sunagawa, S., Pelletier, E., Alberti, A., Levenez, F., … Bork, P. (2017). Towards standards for human fecal sample processing in metagenomic studies. *Nature Biotechnology*, *35*(11), 1069–1076. https://doi.org/10.1038/nbt.3960

Costello, E. K., Lauber, C. L., Hamady, M., Fierer, N., Gordon, J. I., & Knight, R. (2009). Bacterial Community variation in human body habitats across space and time. *Science*, *326*. https://doi.org/10.1126/science.1177486

Cryan, J. F., O'Riordan, K. J., Sandhu, K., Peterson, V., & Dinan, T. G. (2020). The gut microbiome in neurological disorders. *The Lancet Neurology*, *19*(2), 179–194. https://doi.org/https://doi.org/10.1016/S1474-4422(19)30356-4

Cullin, N., Azevedo Antunes, C., Straussman, R., Stein-Thoeringer, C. K., & Elinav, E. (2021). Microbiome and cancer. *Cancer Cell*, *39*(10), 1317–1341. https://doi.org/https://doi.org/10.1016/j.ccell.2021.08.006

D'Amore, R., Ijaz, U. Z., Schirmer, M., Kenny, J. G., Gregory, R., Darby, A. C., … Hall, N. (2016). A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. *BMC Genomics*, *17*(1), 1–20.

Dai, Z., Wong, S. H., Yu, J., & Wei, Y. (2019). Batch effects correction for microbiome data with Dirichlet-multinomial regression. *Bioinformatics*, *35*(5), 807–814.

Damgaard, C., Danielsen, A. K., Enevold, C., Massarenti, L., Nielsen, C. H., Holmstrup, P., & Belstrøm, D. (2019). Porphyromonas gingivalis in saliva associates with chronic and aggressive periodontitis. *Journal of Oral Microbiology*, *11*(1), 1653123. https://doi.org/10.1080/20002297.2019.1653123

De Filippis, F., Vannini, L., La Storia, A., Laghi, L., Piombino, P., Stellato, G., … Ercolini, D. (2014). The Same Microbiota and a Potentially Discriminant Metabolome in the Saliva of Omnivore, Ovo-Lacto-Vegetarian and Vegan Individuals. *PLOS ONE*, *9*(11), e112373. Retrieved from https://doi.org/10.1371/journal.pone.0112373

De Tender, C. A., Devriese, L. I., Haegeman, A., Maes, S., Ruttink, T., & Dawyndt, P. (2015). Bacterial Community Profiling of Plastic Litter in the Belgian Part of the North Sea. *Environmental Science & Technology*, *49*(16), 9629–9638. https://doi.org/10.1021/acs.est.5b01093

DeClercq, V., Langille, M. G. I., & Van Limbergen, J. (2018). Differences in adiposity and diet quality among individuals with inflammatory bowel disease in Eastern Canada. *PLOS ONE*, *13*(7), e0200580. https://doi.org/10.1371/journal.pone.0200580

Derrien, M., Belzer, C., & de Vos, W. M. (2017). Akkermansia muciniphila and its role in regulating host functions. *Microbial Pathogenesis*, *106*, 171–181. https://doi.org/https://doi.org/10.1016/j.micpath.2016.02.005

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., … Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, *72*(7), 5069–5072. https://doi.org/10.1128/AEM.03006-05

Dewhirst, F. E., Chen, T., Izard, J., Paster, B. J., Tanner, A. C. R., Yu, W.-H., … Wade, W. G. (2010). The Human Oral Microbiome. *Journal of Bacteriology*, *192*(19), 5002–5017. https://doi.org/10.1128/JB.00542-10

Dinh, D. M., Volpe, G. E., Duffalo, C., Bhalchandra, S., Tai, A. K., Kane, A. V, … Ward, H. D. (2015). Intestinal Microbiota, Microbial Translocation, and Systemic Inflammation in Chronic HIV Infection. *The Journal of Infectious Diseases*, *211*(1), 19–27. https://doi.org/10.1093/infdis/jiu409

Dixon, P. (2003). VEGAN, a package of R functions for community ecology. *J Vegetation Sci*, *14*. https://doi.org/10.1111/j.1654-1103.2003.tb02228.x

Dobell, C. (1932). Antony van Leeuwenhoek and his" Little Animals." Being Some Account of the Father of Protozoology and Bacteriology and his Multifarious Discoveries in these Disciplines. Collected, Translated, and Edited, from his Printed Works, Unpublished Manuscripts, a. *Antony van Leeuwenhoek and His" Little Animals." Being Some Account of the Father of Protozoology and Bacteriology and His Multifarious Discoveries in These Disciplines. Collected, Translated, and Edited, from His Printed Works, Unpublished Manuscripts, A*.

Doolittle, W. F., & Papke, R. T. (2006). Genomics and the bacterial species problem. *Genome Biology*, *7*(9), 116. https://doi.org/10.1186/gb-2006-7-9-116

Douglas, G. M., Hansen, R., Jones, C. M. A., Dunn, K. A., Comeau, A. M., Bielawski, J. P., … Van Limbergen, J. (2018). Multi-omics differentially classify disease state and treatment outcome in pediatric Crohn's disease. *Microbiome*, *6*(1), 13. https://doi.org/10.1186/s40168-018-0398-3

Douglas, G. M., & Langille, M. G. I. (2021). A primer and discussion on DNA-based microbiome data and related bioinformatics analyses. *Peer Community Journal*, *1*. https://doi.org/10.24072/pcjournal.2

Douglas, G. M., Maffei, V. J., Zaneveld, J. R., Yurgel, S. N., Brown, J. R., Taylor, C. M., … Langille, M. G. I. (2020). PICRUSt2 for prediction of metagenome functions. *Nature Biotechnology*, *38*(6), 685–688. https://doi.org/10.1038/s41587-020-0548-6

Dranse, H. J., Zheng, A., Comeau, A. M., Langille, M. G. I., Zabel, B. A., & Sinal, C. J. (2018). The impact of chemerin or chemokine-like receptor 1 loss on the mouse gut microbiome. *PeerJ*, *6*, e5494. https://doi.org/10.7717/peerj.5494

Ducarmon, Q. R., Zwittink, R. D., H, H. B. V, W, van S., B, Y. V, & J, K. E. (2019). Gut Microbiota and Colonization Resistance against Bacterial Enteric Infection. *Microbiology and Molecular Biology Reviews*, *83*(3), e00007-19. https://doi.org/10.1128/MMBR.00007-19

Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A., & Alm, E. J. (2017). Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nature Communications*, *8*(1), 1784. https://doi.org/10.1038/s41467-017-01973-8

Dziubańska-Kusibab, P. J., Berger, H., Battistini, F., Bouwman, B. A. M., Iftekhar, A., Katainen, R., … Aaltonen, L. A. (2020). Colibactin DNA-damage signature indicates mutational impact in colorectal cancer. *Nature Medicine*, *26*(7), 1063–1069.

Eckburg, P., Bik, E., Bernstein, C., Purdom, E., Dethlefsen, L., Sargent, M., … Relman, D. (2005). Diversity of the human intestinal microbial flora. *Science*, *308*. https://doi.org/10.1126/science.1110591

Edgar, R. (2016). UCHIME2: improved chimera prediction for amplicon sequencing. *BioRxiv*, https://doi.org/10.1101/074252. https://doi.org/10.1101/074252

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*(19), 2460–2461. Retrieved from http://dx.doi.org/10.1093/bioinformatics/btq461

Edgar, R. C. (2016). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *BioRxiv*. Retrieved from http://biorxiv.org/content/early/2016/10/15/081257.abstract

Edgar, R. C. (2017). Accuracy of microbial community diversity estimated by closed- and open-reference OTUs. *PeerJ*, *5*, e3889. https://doi.org/10.7717/peerj.3889

Edgar, R. C. (2018). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics*, *34*(14), 2371–2375. https://doi.org/10.1093/bioinformatics/bty113

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, *27*(16), 2194–2200. https://doi.org/10.1093/bioinformatics/btr381

Eren, A. M., Borisy, G. G., Huse, S. M., & Mark Welch, J. L. (2014). Oligotyping analysis of the human oral microbiome. *Proceedings of the National Academy of Sciences*, *111*(28), E2875–E2884. https://doi.org/10.1073/pnas.1409644111

Eren, A. M., Maignien, L., Sul, W. J., Murphy, L. G., Grim, S. L., Morrison, H. G., & Sogin, M. L. (2013). Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods in Ecology and Evolution*, *4*(12), 1111–1119. https://doi.org/https://doi.org/10.1111/2041-210X.12114

Ervin, S. M., Li, H., Lim, L., Roberts, L. R., Liang, X., Mani, S., & Redinbo, M. R. (2019). Gut microbial β-glucuronidases reactivate estrogens as components of the estrobolome that reactivate estrogens. *Journal of Biological Chemistry*, *294*(49), 18586–18599.

Estemalik, J., Demko, C., Bissada, N. F., Joshi, N., Bodner, D., Shankar, E., & Gupta, S. (2017). Simultaneous detection of oral pathogens in subgingival plaque and prostatic fluid of men with periodontal and prostatic diseases. *Journal of Periodontology*, *88*(9), 823–829.

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation*, *61*(1), 1–10.

Fan, X., Alekseyenko, A. V, Wu, J., Peters, B. A., Jacobs, E. J., Gapstur, S. M., … Ahn, J. (2018). Human oral microbiome and prospective risk for pancreatic cancer: a population-based nested case-control study. *Gut*, *67*(1), 120 LP – 127. https://doi.org/10.1136/gutjnl-2016-312580

Fan, X., Peters, B. A., Jacobs, E. J., Gapstur, S. M., Purdue, M. P., Freedman, N. D., … Ahn, J. (2018). Drinking alcohol is associated with variation in the human oral microbiome in a large study of American adults. *Microbiome*, *6*(1), 59. https://doi.org/10.1186/s40168-018-0448-x

Fan, Y., & Pedersen, O. (2021). Gut microbiota in human metabolic health and disease. *Nature Reviews Microbiology*, *19*(1), 55–71. https://doi.org/10.1038/s41579-020-0433-9

Feng, Y., Ramnarine, V. R., Bell, R., Volik, S., Davicioni, E., Hayes, V. M., … Collins, C. C. (2019). Metagenomic and metatranscriptomic analysis of human prostate microbiota from patients with prostate cancer. *BMC Genomics*, *20*(1), 1–8.

Fernandes, A. D., Reid, J. N., Macklaim, J. M., McMurrough, T. A., Edgell, D. R., & Gloor, G. B. (2014a). Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, *2*(1), 15. https://doi.org/10.1186/2049-2618-2-15

Fernandes, A. D., Reid, J. N. S., Macklaim, J. M., McMurrough, T. A., Edgell, D. R., & Gloor, G. B. (2014b). Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, *2*(1), 15. https://doi.org/10.1186/2049-2618-2-15

Fernández, M. F., Reina-Pérez, I., Astorga, J. M., Rodríguez-Carrillo, A., Plaza-Díaz, J., & Fontana, L. (2018). Breast Cancer and Its Relationship with the Microbiota. *International Journal of Environmental Research and Public Health*, *15*(8), 1747. https://doi.org/10.3390/ijerph15081747

Fierer, N., & Jackson, R. B. (2006). The diversity and biogeography of soil bacterial communities. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(3), 626–631. https://doi.org/10.1073/pnas.0507535103

Flemer, B., Warren, R. D., Barrett, M. P., Cisek, K., Das, A., Jeffery, I. B., … O'Toole, P. W. (2017). The oral microbiota in colorectal cancer is distinctive and predictive. *Gut*. Retrieved from http://gut.bmj.com/content/early/2017/10/07/gutjnl-2017-314814.abstract

Flores, R., Shi, J., Fuhrman, B., Xu, X., Veenstra, T. D., Gail, M. H., … Goedert, J. J. (2012). Fecal microbial determinants of fecal and systemic estrogens and estrogen metabolites: a cross-sectional study. *Journal of Translational Medicine*, *10*(1), 253. https://doi.org/10.1186/1479-5876-10-253

Frère, L., Maignien, L., Chalopin, M., Huvet, A., Rinnert, E., Morrison, H., … Paul-Pont, I. (2018). Microplastic bacterial communities in the Bay of Brest: Influence of polymer type and size. *Environmental Pollution*, *242*, 614–625. https://doi.org/https://doi.org/10.1016/j.envpol.2018.07.023

Furrie, E., Macfarlane, S., Kennedy, A., Cummings, J. H., Walsh, S. V, O'Neil, D. A., & Macfarlane, G. T. (2005). Synbiotic therapy (&lt;em&gt;Bifidobacterium longum&lt;/em&gt;/Synergy 1) initiates resolution of inflammation in patients with active ulcerative colitis: a randomised controlled pilot trial. *Gut*, *54*(2), 242 LP – 249. https://doi.org/10.1136/gut.2004.044834

Garrett, W. S. (2019). The gut microbiota and colon cancer. *Science*, *364*(6446), 1133–1135. https://doi.org/10.1126/science.aaw2367

Gehrig, J. L., Portik, D. M., Driscoll, M. D., Jackson, E., Chakraborty, S., Gratalo, D., … Valladares, R. (2022). Finding the right fit: evaluation of short-read and long-read sequencing approaches to maximize the utility of clinical microbiome data. *Microbial Genomics*, *8*(3), 000794. https://doi.org/10.1099/mgen.0.000794

Gevers, D., Cohan, F. M., Lawrence, J. G., Spratt, B. G., Coenye, T., Feil, E. J., … Swings, J. (2005). Re-evaluating prokaryotic species. *Nature Reviews Microbiology*, *3*(9), 733–739. https://doi.org/10.1038/nrmicro1236

Gibbons, S. M., Duvallet, C., & Alm, E. J. (2018). Correcting for batch effects in case-control microbiome studies. *PLOS Computational Biology*, *14*(4), e1006102. Retrieved from https://doi.org/10.1371/journal.pcbi.1006102

Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V, & Knight, R. (2018). Current understanding of the human microbiome. *Nature Medicine*, *24*(4), 392–400. https://doi.org/10.1038/nm.4517

Glassner, K. L., Abraham, B. P., & Quigley, E. M. M. (2020). The microbiome and inflammatory bowel disease. *Journal of Allergy and Clinical Immunology*, *145*(1), 16–27. https://doi.org/10.1016/j.jaci.2019.11.003

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology*, *8*, 2224. https://doi.org/10.3389/fmicb.2017.02224

Goedert, J. J., Hua, X., Bielecka, A., Okayasu, I., Milne, G. L., Jones, G. S., … Feigelson, H. S. (2018). Postmenopausal breast cancer and oestrogen associations with the IgA-coated and IgA-noncoated faecal microbiota. *British Journal Of Cancer*, *118*, 471. Retrieved from https://doi.org/10.1038/bjc.2017.435

Goedert, J. J., Jones, G., Hua, X., Xu, X., Yu, G., Flores, R., … Feigelson, H. S. (2015). Investigation of the Association Between the Fecal Microbiota and Breast Cancer in Postmenopausal Women: a Population-Based Case-Control Pilot Study. *JNCI: Journal of the National Cancer Institute*, *107*(8), djv147–djv147. Retrieved from http://dx.doi.org/10.1093/jnci/djv147

Golombos, D. M., Ayangbesan, A., O'Malley, P., Lewicki, P., Barlow, L., Barbieri, C. E., … Scherr, D. S. (2018). The Role of Gut Microbiome in the Pathogenesis of Prostate Cancer: A Prospective, Pilot Study. *Urology*, *111*, 122–128. https://doi.org/https://doi.org/10.1016/j.urology.2017.08.039

Gonzalez, A., Navas-Molina, J. A., Kosciolek, T., McDonald, D., Vázquez-Baeza, Y., Ackermann, G., … Knight, R. (2018). Qiita: rapid, web-enabled microbiome meta-analysis. *Nature Methods*, *15*(10), 796–798. https://doi.org/10.1038/s41592-018-0141-9

Goodrich, J. K., Di Rienzi, S. C., Poole, A. C., Koren, O., Walters, W. A., Caporaso, J. G., … Ley, R. E. (2014). Conducting a Microbiome Study. *Cell*, *158*(2), 250–262. https://doi.org/https://doi.org/10.1016/j.cell.2014.06.037

Goodrich, J. K., Waters, J. L., Poole, A. C., Sutter, J. L., Koren, O., Blekhman, R., … Ley, R. E. (2014). Human genetics shape the gut microbiome. *Cell*, *159*(4), 789–799. https://doi.org/10.1016/j.cell.2014.09.053

Grau, J., Grosse, I., & Keilwagen, J. (2015). PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics (Oxford, England)*, *31*(15), 2595–2597. https://doi.org/10.1093/bioinformatics/btv153

Hall, M. W., Singh, N., Ng, K. F., Lam, D. K., Goldberg, M. B., Tenenbaum, H. C., … Senadheera, D. B. (2017). Inter-personal diversity and temporal dynamics of dental, tongue, and salivary microbiota in the healthy oral cavity. *Npj Biofilms and Microbiomes*, *3*(1), 2. https://doi.org/10.1038/s41522-016-0011-0

Hansen, T. H., Kern, T., Bak, E. G., Kashani, A., Allin, K. H., Nielsen, T., … Pedersen, O. (2018). Impact of a vegan diet on the human salivary microbiota. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-24207-3

Hawinkel, S., Mattiello, F., Bijnens, L., & Thas, O. (2019). A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Briefings in Bioinformatics*, *20*(1), 210–221. https://doi.org/10.1093/bib/bbx104

He, J., Li, Y., Cao, Y., Xue, J., & Zhou, X. (2015). The oral microbiome diversity and its relation to human diseases. *Folia Microbiologica*, *60*(1), 69–80. https://doi.org/10.1007/s12223-014-0342-2

He, Y., Caporaso, J. G., Jiang, X.-T., Sheng, H.-F., Huse, S. M., Rideout, J. R., … Zhou, H.-W. (2015). Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. *Microbiome*, *3*(1), 20. https://doi.org/10.1186/s40168-015-0081-x

Healy, C. M., & Moran, G. P. (2019). The microbiome and oral cancer: More questions than answers. *Oral Oncology*, *89*, 30–33.

Hill, C. (2020). You have the microbiome you deserve. *Gut Microbiome*, *1*, e3. https://doi.org/DOI: 10.1017/gmb.2020.3

Hoellein, T. J., McCormick, A. R., Hittie, J., London, M. G., Scott, J. W., & Kelly, J. J. (2017). Longitudinal patterns of microplastic concentration and bacterial assemblages in surface and benthic habitats of an urban river. *Freshwater Science*, *36*(3), 491–507. https://doi.org/10.1086/693012

Hooi, J. K. Y., Lai, W. Y., Ng, W. K., Suen, M. M. Y., Underwood, F. E., Tanyingoh, D., … Ng, S. C. (2017). Global Prevalence of <em>Helicobacter pylori</em> Infection: Systematic Review and Meta-Analysis. *Gastroenterology*, *153*(2), 420–429. https://doi.org/10.1053/j.gastro.2017.04.022

Hou, M.-F., Ou-Yang, F., Li, C.-L., Chen, F.-M., Chuang, C.-H., Kan, J.-Y., … Chiang, C.-P. (2021). Comprehensive profiles and diagnostic value of menopausal-specific gut microbiota in premenopausal breast cancer. *Experimental & Molecular Medicine*, *53*(10), 1636–1646. https://doi.org/10.1038/s12276-021-00686-9

Hsiao, J.-R., Chang, C.-C., Lee, W.-T., Huang, C.-C., Ou, C.-Y., Tsai, S.-T., … Lai, Y.-H. (2018). The interplay between oral microbiome, lifestyle factors and genetic polymorphisms in the risk of oral squamous cell carcinoma. *Carcinogenesis*, *39*(6), 778–787.

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., … Consortium, T. H. M. P. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, *486*(7402), 207–214. https://doi.org/10.1038/nature11234

Illumina. (n.d.). *Effects of Index Misassignment on Multiplexing and Downstream Analysis*. Retrieved from https://www.illumina.com/content/dam/illumina-marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf

Irfan, M., Delgado, R. Z. R., & Frias-Lopez, J. (2020). The oral microbiome and cancer. *Frontiers in Immunology*, *11*, 591088.

Jangi, S., Gandhi, R., Cox, L. M., Li, N., von Glehn, F., Yan, R., … Weiner, H. L. (2016). Alterations of the human gut microbiome in multiple sclerosis. *Nature Communications*, *7*(1), 12015. https://doi.org/10.1038/ncomms12015

Janssen, S., McDonald, D., Gonzalez, A., Navas-Molina, J. A., Jiang, L., Xu, Z. Z., … Knight, R. (2018). Phylogenetic Placement of Exact Amplicon Sequences Improves Associations with Clinical Information. *MSystems*, *3*(3), e00021-18. https://doi.org/10.1128/mSystems.00021-18

Järvenpää, P., Kosunen, T., Fotsis, T., & Adlercreutz, H. (1980). In vitro metabolism of estrogens by isolated intestinal micro-organisms and by human faecal microflora. *Journal of Steroid Biochemistry*, *13*(3), 345–349. https://doi.org/https://doi.org/10.1016/0022-4731(80)90014-X

Javier-DesLoges, J., McKay, R. R., Swafford, A. D., Sepich-Poore, G. D., Knight, R., & Parsons, J. K. (2022). The microbiome and prostate cancer. *Prostate Cancer and Prostatic Diseases*, *25*(2), 159–164. https://doi.org/10.1038/s41391-021-00413-5

Ji, P., Parks, J., Edwards, M. A., & Pruden, A. (2015). Impact of Water Chemistry, Pipe Material and Stagnation on the Building Plumbing Microbiome. *PLOS ONE*, *10*(10), e0141087. Retrieved from https://doi.org/10.1371/journal.pone.0141087

Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., … Weinstock, G. M. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications*, *10*(1), 5029. https://doi.org/10.1038/s41467-019-13036-1

Johnson, K., & Burnet, P. (2016). Microbiome: should we diversify from diversity? *Gut Microbes*, *7*(6), 455–458.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Karpiński, M. T. (2019). Role of Oral Microbiota in Cancer Development. *Microorganisms* , Vol. 7. https://doi.org/10.3390/microorganisms7010020

Katz, R., Ahmed, M. A., Safadi, A., Abu Nasra, W., Visoki, A., Huckim, M., … Neuman, H. (2022). Characterization of fecal microbiome in biopsy positive prostate cancer patients. *BJUI Compass*, *3*(1), 55–61.

Kaul, A., Mandal, S., Davidov, O., & Peddada, S. D. (2017). Analysis of Microbiome Data in the Presence of Excess Zeros. *Frontiers in Microbiology*, *8*, 2114. https://doi.org/10.3389/fmicb.2017.02114

Kembel, S. W., Wu, M., Eisen, J. A., & Green, J. L. (2012). Incorporating 16S Gene Copy Number Information Improves Estimates of Microbial Diversity and Abundance. *PLOS Computational Biology*, *8*(10), e1002743. Retrieved from https://doi.org/10.1371/journal.pcbi.1002743

Keohavong, P., & Thilly, W. G. (1989). Fidelity of DNA polymerases in DNA amplification. *Proceedings of the National Academy of Sciences*, *86*(23), 9253–9257. https://doi.org/10.1073/pnas.86.23.9253

Kesy, K., Oberbeckmann, S., Kreikemeyer, B., & Labrenz, M. (2019). Spatial Environmental Heterogeneity Determines Young Biofilm Assemblages on Microplastics in Baltic Sea Mesocosms . *Frontiers in Microbiology* , Vol. 10, p. 1665. Retrieved from https://www.frontiersin.org/article/10.3389/fmicb.2019.01665

Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., … Dorrestein, P. C. (2018). Best practices for analysing microbiomes. *Nature Reviews Microbiology*, *16*(7), 410–422. https://doi.org/10.1038/s41579-018-0029-9

Kolde, R. (2012). Pheatmap: pretty heatmaps. *R Package Version*, *1*(2).

Kolenbrander, P. E., Palmer Jr, R. J., Rickard, A. H., Jakubovics, N. S., Chalmers, N. I., & Diaz, P. I. (2006). Bacterial interactions and successions during plaque development. *Periodontology 2000*, *42*(1), 47–79. https://doi.org/10.1111/j.1600-0757.2006.00187.x

Kõljalg, U., Larsson, K., Abarenkov, K., Nilsson, R. H., Alexander, I. J., Eberhardt, U., … Larsson, E. (2005). UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. *New Phytologist*, *166*(3), 1063–1068.

Kõljalg, U., Nilsson, R. H., Abarenkov, K., Tedersoo, L., Taylor, A. F. S., Bahram, M., … Larsson, K.-H. (2013). Towards a unified paradigm for sequence-based identification of fungi. *Molecular Ecology*, *22*(21), 5271–5277. https://doi.org/10.1111/mec.12481

Komiya, Y., Shimomura, Y., Higurashi, T., Sugi, Y., Arimoto, J., Umezawa, S., … Nakajima, A. (2019). Patients with colorectal cancer have identical strains of &lt;em&gt;Fusobacterium nucleatum&lt;/em&gt; in their colorectal cancer and oral cavity. *Gut*, *68*(7), 1335 LP – 1337. https://doi.org/10.1136/gutjnl-2018-316661

Kopylova, E., Noé, L., & Touzet, H. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, *28*(24), 3211–3217. https://doi.org/10.1093/bioinformatics/bts611

Kroes, I., Lepp, P. W., & Relman, D. A. (1999). Bacterial diversity within the human subgingival crevice. *Proceedings of the National Academy of Sciences*, *96*(25), 14547–14552. https://doi.org/10.1073/pnas.96.25.14547

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software; Vol 1, Issue 5 (2008)* . Retrieved from https://www.jstatsoft.org/v028/i05

Kumar, P. S., Griffen, A. L., Barton, J. A., Paster, B. J., Moeschberger, M. L., & Leys, E. J. (2003). New Bacterial Species Associated with Chronic Periodontitis. *Journal of Dental Research*, *82*(5), 338–344. https://doi.org/10.1177/154405910308200503

Laforest-Lapointe, I., & Arrieta, M.-C. (2018). Microbial Eukaryotes: a Missing Link in Gut Microbiome Studies. *MSystems*, *3*(2), e00201-17. https://doi.org/10.1128/mSystems.00201-17

Lamont, R. J., Koo, H., & Hajishengallis, G. (2018). The oral microbiota: dynamic communities and host interactions. *Nature Reviews Microbiology*, *16*(12), 745–759. https://doi.org/10.1038/s41579-018-0089-x

Lamoureux, E. V, Grandy, S. A., & Langille, M. G. I. (2017). Moderate Exercise Has Limited but Distinguishable Effects on the Mouse Microbiome. *MSystems*, *2*(4), e00006-17. https://doi.org/10.1128/mSystems.00006-17

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, *15*(2), R29. https://doi.org/10.1186/gb-2014-15-2-r29

Leake, S. L., Pagni, M., Falquet, L., Taroni, F., & Greub, G. (2016). The salivary microbiome for differentiating individuals: proof of principle. *Microbes and Infection*, *18*(6), 399–405. https://doi.org/https://doi.org/10.1016/j.micinf.2016.03.011

Lee, J.-H., Kweon, H. H.-I., Choi, J.-K., Kim, Y.-T., & Choi, S.-H. (2017). Association between periodontal disease and prostate cancer: results of a 12-year longitudinal cohort study in South Korea. *Journal of Cancer*, *8*(15), 2959.

Leek, J. T. (2014). Svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Research*, *42*(21), e161–e161.

Ley, R. E., Turnbaugh, P. J., Klein, S., & Gordon, J. I. (2006). Human gut microbes associated with obesity. *Nature*, *444*(7122), 1022–1023. https://doi.org/10.1038/4441022a

Li, H.-S., Ou-Yang, L., Zhu, Y., Yan, H., & Zhang, X.-F. (2022). scDEA: differential expression analysis in single-cell RNA-sequencing data via ensemble learning. *Briefings in Bioinformatics*, *23*(1), bbab402. https://doi.org/10.1093/bib/bbab402

Li, J., Quinque, D., Horz, H.-P., Li, M., Rzhetskaya, M., Raff, J. A., … Stoneking, M. (2014). Comparative analysis of the human saliva microbiome from different climate zones: Alaska, Germany, and Africa. *BMC Microbiology*, *14*(1), 316. https://doi.org/10.1186/s12866-014-0316-1

Lin, H., & Peddada, S. Das. (2020). Analysis of microbial compositions: a review of normalization and differential abundance analysis. *NPJ Biofilms and Microbiomes*, *6*(1), 60. https://doi.org/10.1038/s41522-020-00160-w

Ling, W., Lu, J., Zhao, N., Lulla, A., Plantinga, A. M., Fu, W., … Li, Z. (2022). Batch effects removal for microbiome data via conditional quantile regression. *Nature Communications*, *13*(1), 1–14.

Links, M. G., Dumonceaux, T. J., Hemmingsen, S. M., & Hill, J. E. (2012). The Chaperonin-60 Universal Target Is a Barcode for Bacteria That Enables De Novo Assembly of Metagenomic Sequence Data. *PLOS ONE*, *7*(11), e49755. Retrieved from https://doi.org/10.1371/journal.pone.0049755

Liss, M. A., White, J. R., Goros, M., Gelfond, J., Leach, R., Johnson-Pais, T., … Shah, D. P. (2018). Metabolic Biosynthesis Pathways Identified from Fecal Microbiome Associated with Prostate Cancer. *European Urology*, *74*(5), 575–582. https://doi.org/https://doi.org/10.1016/j.eururo.2018.06.033

Lloyd-Price, J., Abu-Ali, G., & Huttenhower, C. (2016). The healthy human microbiome. *Genome Medicine*, *8*(1), 51. https://doi.org/10.1186/s13073-016-0307-y

Lloyd-Price, J., Arze, C., Ananthakrishnan, A. N., Schirmer, M., Avila-Pacheco, J., Poon, T. W., … Investigators, I. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, *569*(7758), 655–662. https://doi.org/10.1038/s41586-019-1237-9

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

Lozupone, C. A., Li, M., Campbell, T. B., Flores, S. C., Linderman, D., Gebert, M. J., … Palmer, B. E. (2013). Alterations in the Gut Microbiota Associated with HIV-1 Infection. *Cell Host & Microbe*, *14*(3), 329–339. https://doi.org/10.1016/j.chom.2013.08.006

Lozupone, C, & Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol*, *71*. https://doi.org/10.1128/AEM.71.12.8228-8235.2005

Lozupone, Catherine, Lladser, M. E., Knights, D., Stombaugh, J., & Knight, R. (2011). UniFrac: an effective distance metric for microbial community comparison. *The ISME Journal*, *5*(2), 169–172.

Lu, H., Ren, Z., Li, A., Li, J., Xu, S., Zhang, H., … Zhou, K. (2019). Tongue coating microbiome data distinguish patients with pancreatic head cancer from healthy controls. *Journal of Oral Microbiology*, *11*(1), 1563409.

Lu, X., Junyan, Y., Ting, Y., Jiang, Z., Tingyu, L., Hong, W., & Jie, C. (2021). Fecal Microbiome Transplantation from Children with Autism Spectrum Disorder Modulates Tryptophan and Serotonergic Synapse Metabolism and Induces Altered Behaviors in Germ-Free Mice. *MSystems*, *6*(2), e01343-20. https://doi.org/10.1128/mSystems.01343-20

Lundmark, A., Hu, Y. O. O., Huss, M., Johannsen, G., Andersson, A. F., & Yucel-Lindberg, T. (2019). Identification of salivary microbiota and its association with host inflammatory mediators in periodontitis. *Frontiers in Cellular and Infection Microbiology*, *9*, 216.

Ma, S., Ren, B., Mallick, H., Moon, Y. S., Schwager, E., Maharjan, S., … Huttenhower, C. (2021). A statistical model for describing and simulating microbial community profiles. *PLOS Computational Biology*, *17*(9), e1008913. Retrieved from https://doi.org/10.1371/journal.pcbi.1008913

Madhusoodanan, J. (2019). Do hosts and their microbes evolve as a unit? *Proceedings of the National Academy of Sciences*, *116*(29), 14391–14394. https://doi.org/10.1073/pnas.1908139116

Magne, F., Gotteland, M., Gauthier, L., Zazueta, A., Pesoa, S., Navarrete, P., & Balamurugan, R. (2020). The firmicutes/bacteroidetes ratio: A relevant marker of gut dysbiosis in obese patients? *Nutrients*, *12*(5). https://doi.org/10.3390/nu12051474

Mallick, H., Rahnavard, A., McIver, L. J., Ma, S., Zhang, Y., Nguyen, L. H., … Huttenhower, C. (2021). Multivariable Association Discovery in Population-scale Meta-omics Studies. *BioRxiv*, 2021.01.20.427420. https://doi.org/10.1101/2021.01.20.427420

Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R., & Peddada, S. D. (2015). Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial Ecology in Health and Disease*, *26*(1), 27663. https://doi.org/10.3402/mehd.v26.27663

Mandel, I. D. (1987). The Functions of Saliva. *Journal of Dental Research*, *66*(1_suppl), 623–627. https://doi.org/10.1177/00220345870660S103

Marcos-Zambrano, L. J., Karaduzovic-Hadziabdic, K., Loncar Turukalo, T., Przymus, P., Trajkovik, V., Aasmets, O., … Truu, J. (2021). Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment . *Frontiers in Microbiology* , Vol. 12. Retrieved from https://www.frontiersin.org/articles/10.3389/fmicb.2021.634511

Mark Welch, J. L., Dewhirst, F. E., & Borisy, G. G. (2019). Biogeography of the Oral Microbiome: The Site-Specialist Hypothesis. *Annual Review of Microbiology*, *73*(1), 335–358. https://doi.org/10.1146/annurev-micro-090817-062503

Mark Welch, J. L., Rossetti, B. J., Rieken, C. W., Dewhirst, F. E., & Borisy, G. G. (2016). Biogeography of a human oral microbiome at the micron scale. *Proceedings of the National Academy of Sciences*, *113*(6), E791–E800. https://doi.org/10.1073/pnas.1522149113

Martin, B. D., Witten, D., & Willis, A. D. (2020). Modeling microbial abundances and dysbiosis with beta-binomial regression. *The Annals of Applied Statistics*, *14*(1), 94–115. https://doi.org/10.1214/19-AOAS1283

Martin, F., Peltonen, J., Laatikainen, T., Pulkkinen, M., & Adlercreutz, H. (1975). Excretion of progesteone metabolites and estriol in faeces from pregnant women during ampicillin administration. *Journal of Steroid Biochemistry*, *6*(9), 1339–1346. https://doi.org/https://doi.org/10.1016/0022-4731(75)90363-5

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal; Vol 17, No 1: Next Generation Sequencing Data Analysis*. Retrieved from http://journal.embnet.org/index.php/embnetjournal/article/view/200

Martino, C., Dilmore, A. H., Burcham, Z. M., Metcalf, J. L., Jeste, D., & Knight, R. (2022). Microbiota succession throughout life from the cradle to the grave. *Nature Reviews Microbiology*, 1–14. https://doi.org/10.1038/s41579-022-00768-z

Mason, M. R., Nagaraja, H. N., Camerlengo, T., Joshi, V., & Kumar, P. S. (2013). Deep Sequencing Identifies Ethnicity-Specific Bacterial Signatures in the Oral Microbiome. *PLOS ONE*, *8*(10), e77287. Retrieved from https://doi.org/10.1371/journal.pone.0077287

Matsen IV, F. A. (2015). Phylogenetics and the Human Microbiome. *Systematic Biology*, *64*(1), e26–e41. https://doi.org/10.1093/sysbio/syu053

Matsushita, M., Fujita, K., Motooka, D., Hatano, K., Fukae, S., Kawamura, N., … Takao, T. (2021). The gut microbiota associated with high-Gleason prostate cancer. *Cancer Science*, *112*(8), 3125–3135.

McCormick, A. R., Hoellein, T. J., London, M. G., Hittie, J., Scott, J. W., & Kelly, J. J. (2016). Microplastic in surface waters of urban rivers: concentration, sources, and associated bacterial assemblages. *Ecosphere*, *7*(11), e01556. https://doi.org/https://doi.org/10.1002/ecs2.1556

McCoy, C. O., & Matsen IV, F. A. (2013). Abundance-weighted phylogenetic diversity measures distinguish microbial community states and are robust to sampling depth. *PeerJ*, *1*, e157. https://doi.org/10.7717/peerj.157

McInerney, P., Adams, P., & Hadi, M. Z. (2014). Error rate comparison during polymerase chain reaction by DNA polymerase. *Molecular Biology International*, *2014*.

McLaren, M. R., Nearing, J. T., Willis, A. D., Lloyd, K. G., & Callahan, B. J. (2022). Implications of taxonomic bias for microbial differential-abundance analysis. *BioRxiv*, 2022.08.19.504330. https://doi.org/10.1101/2022.08.19.504330

McLaren, M. R., Willis, A. D., & Callahan, B. J. (2019). Consistent and correctable bias in metagenomic sequencing experiments. *ELife*, *8*, e46923. https://doi.org/10.7554/eLife.46923

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

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

Mejía-León, M. E., Petrosino, J. F., Ajami, N. J., Domínguez-Bello, M. G., & de la Barca, A. M. C. (2014). Fecal microbiota imbalance in Mexican children with type 1 diabetes. *Scientific Reports*, *4*(1), 3814. https://doi.org/10.1038/srep03814

MIRARAB, S., NGUYEN, N., & WARNOW, T. (2011). SEPP: SATé-Enabled Phylogenetic Placement. In *Biocomputing 2012* (pp. 247–258). WORLD SCIENTIFIC. https://doi.org/doi:10.1142/9789814366496_0024

Moore, W. E. C., Stackebrandt, E., Kandler, O., Colwell, R. R., Krichevsky, M. I., Truper, H. G., … Moore, L. H. (1987). Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. *International Journal of Systematic and Evolutionary Microbiology*, *37*(4), 463–464. https://doi.org/10.1099/00207713-37-4-463

Morton, J. T., Donovan, S. M., & Taroncher-Oldenburg, G. (2022). Decoupling diet from microbiome dynamics results in model mis-specification that implicitly annuls potential associations between the microbiome and disease phenotypes—ruling out any role of the microbiome in autism (Yap &lt;em&gt;et al.&lt;/em&gt; 2021) . *BioRxiv*, 2022.02.25.482051. https://doi.org/10.1101/2022.02.25.482051

Morton, J. T., Marotz, C., Washburne, A., Silverman, J., Zaramela, L. S., Edlund, A., … Knight, R. (2019). Establishing microbial composition measurement standards with reference frames. *Nature Communications*, *10*(1), 2719. https://doi.org/10.1038/s41467-019-10656-5

Mougeot, J.-L. C., Beckman, M. F., Langdon, H. C., Lalla, R. V, Brennan, M. T., & Mougeot, F. K. B. (2021). Haemophilus pittmaniae and Leptotrichia spp. Constitute a Multi-Marker Signature in a Cohort of Human Papillomavirus-Positive Head and Neck Cancer Patients. *Frontiers in Microbiology*, *12*.

Murray, P. A., Prakobphol, A., Lee, T., Hoover, C. I., & Fisher, S. J. (1992). Adherence of oral streptococci to salivary glycoproteins. *Infection and Immunity*, *60*(1), 31 LP – 38. Retrieved from http://iai.asm.org/content/60/1/31.abstract

Nasidze, I., Li, J., Quinque, D., Tang, K., & Stoneking, M. (2009). Global diversity in the human salivary microbiome. *Genome Res*, *19*. https://doi.org/10.1101/gr.084616.108

Navas-Molina, J. A., Peralta-Sánchez, J. M., González, A., McMurdie, P. J., Vázquez-Baeza, Y., Xu, Z., … Song, S. J. (2013). Advancing our understanding of the human microbiome using QIIME. In *Methods in enzymology* (Vol. 531, pp. 371–444). Elsevier.

Nearing, J. T., Comeau, A. M., & Langille, M. G. I. (2021). Identifying biases and their potential solutions in human microbiome studies. *Microbiome*, *9*(1), 113. https://doi.org/10.1186/s40168-021-01059-0

Nearing, J. T., Connors, J., Whitehouse, S., Van Limbergen, J., Macdonald, T., Kulkarni, K., & Langille, M. G. I. (2019). Infectious Complications Are Associated With Alterations in the Gut Microbiome in Pediatric Patients With Acute Lymphoblastic Leukemia. *Frontiers in Cellular and Infection Microbiology*, *9*, 28. https://doi.org/10.3389/fcimb.2019.00028

Nearing, J. T., DeClercq, V., Van Limbergen, J., & Langille, M. G. I. (2020). Assessing the Variation within the Oral Microbiome of Healthy Adults. *MSphere*, *5*(5). https://doi.org/10.1128/mSphere.00451-20

Nearing, J. T., Douglas, G. M., Comeau, A. M., & Langille, M. G. I. (2018). Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. *PeerJ*, *6*, e5364. https://doi.org/10.7717/peerj.5364

Nearing, J. T., Douglas, G. M., Hayes, M. G., MacDonald, J., Desai, D. K., Allward, N., … Langille, M. G. I. (2022). Microbiome differential abundance methods produce different results across 38 datasets. *Nature Communications*, *13*(1), 342. https://doi.org/10.1038/s41467-022-28034-z

Noguera-Julian, M., Rocafort, M., Guillén, Y., Rivera, J., Casadellà, M., Nowak, P., … Paredes, R. (2016). Gut Microbiota Linked to Sexual Preference and HIV Infection. *EBioMedicine*, *5*, 135–146. https://doi.org/10.1016/j.ebiom.2016.01.032

Oberbeckmann, S., Osborn, A. M., & Duhaime, M. B. (2016). Microbes on a Bottle: Substrate, Season and Geography Influence Community Composition of Microbes Colonizing Marine Plastic Debris. *PLOS ONE*, *11*(8), e0159289. Retrieved from https://doi.org/10.1371/journal.pone.0159289

Ochi, A., Nguyen, A. H., Bedrosian, A. S., Mushlin, H. M., Zarbakhsh, S., Barilla, R., … Miller, G. (2012). MyD88 inhibition amplifies dendritic cell capacity to promote pancreatic carcinogenesis via Th2 cells. *Journal of Experimental Medicine*, *209*(9), 1671–1687. https://doi.org/10.1084/jem.20111706

Ogier, J.-C., Pagès, S., Galan, M., Barret, M., & Gaudriault, S. (2019). rpoB, a promising marker for analyzing the diversity of bacterial communities by amplicon sequencing. *BMC Microbiology*, *19*(1), 171. https://doi.org/10.1186/s12866-019-1546-z

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., … Wagner, H. (2018). *vegan: Community Ecology Package*. Retrieved from https://cran.r-project.org/package=vegan

Oliveira, F. S., Brestelli, J., Cade, S., Zheng, J., Iodice, J., Fischer, S., … Beiting, D. P. (2018). MicrobiomeDB: a systems biology platform for integrating, mining and analyzing microbiome experiments. *Nucleic Acids Research*, *46*(D1), D684–D691. https://doi.org/10.1093/nar/gkx1027

Omori, M., Kato-Kogoe, N., Sakaguchi, S., Kamiya, K., Fukui, N., Gu, Y.-H., … Ueno, T. (2022). Characterization of salivary microbiota in elderly patients with type 2 diabetes mellitus: a matched case–control study. *Clinical Oral Investigations*, *26*(1), 493–504. https://doi.org/10.1007/s00784-021-04027-y

Panek, M., Čipčić Paljetak, H., Barešić, A., Perić, M., Matijašić, M., Lojkić, I., … Verbanac, D. (2018). Methodology challenges in studying human gut microbiota – effects of collection, storage, DNA extraction and next generation sequencing technologies. *Scientific Reports*, *8*(1), 5143. https://doi.org/10.1038/s41598-018-23296-4

Papa, E., Docktor, M., Smillie, C., Weber, S., Preheim, S. P., Gevers, D., … Alm, E. J. (2012). Non-Invasive Mapping of the Gastrointestinal Microbiota Identifies Children with Inflammatory Bowel Disease. *PLOS ONE*, *7*(6), e39242. Retrieved from https://doi.org/10.1371/journal.pone.0039242

Parks, D. H., & Beiko, R. G. (2013). Measures of phylogenetic differentiation provide robust and complementary insights into microbial communities. *The ISME Journal*, *7*(1), 173–183. https://doi.org/10.1038/ismej.2012.88

Paster, B. J., Dewhirst, F. E., Olsen, I., & Fraser, G. J. (1994). Phylogeny of Bacteroides, Prevotella, and Porphyromonas spp. and related bacteria. *Journal of Bacteriology*, *176*(3), 725–732. https://doi.org/10.1128/jb.176.3.725-732.1994

Paulson, J. N., Stine, O. C., Bravo, H. C., & Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nature Methods*, *10*(12), 1200–1202. https://doi.org/10.1038/nmeth.2658

Peeters, K., & Willems, A. (2011). The gyrB gene is a useful phylogenetic marker for exploring the diversity of Flavobacterium strains isolated from terrestrial and aquatic habitats in Antarctica. *FEMS Microbiology Letters*, *321*(2), 130–140. https://doi.org/10.1111/j.1574-6968.2011.02326.x

Peres, M. A., Macpherson, L. M. D., Weyant, R. J., Daly, B., Venturelli, R., Mathur, M. R., … Watt, R. G. (2019). Oral diseases: a global public health challenge. *The Lancet*, *394*(10194), 249–260. https://doi.org/https://doi.org/10.1016/S0140-6736(19)31146-8

Peters, B. A., McCullough, M. L., Purdue, M. P., Freedman, N. D., Um, C. Y., Gapstur, S. M., … Ahn, J. (2018). Association of Coffee and Tea Intake with the Oral Microbiome: Results from a Large Cross-Sectional Study. *Cancer Epidemiology Biomarkers &amp;Amp; Prevention*, *27*(7), 814 LP – 821. https://doi.org/10.1158/1055-9965.EPI-18-0184

Petrosino, J. F. (2018). The microbiome in precision medicine: the way forward. *Genome Medicine*, *10*, 12. https://doi.org/10.1186/s13073-018-0525-6

Phipson, B., Lee, S., Majewski, I. J., Alexander, W. S., & Smyth, G. K. (2016). ROBUST HYPERPARAMETER ESTIMATION PROTECTS AGAINST HYPERVARIABLE GENES AND IMPROVES POWER TO DETECT DIFFERENTIAL EXPRESSION. *The Annals of Applied Statistics*, *10*(2), 946–963. https://doi.org/10.1214/16-AOAS920

Plottel, C. S., & Blaser, M. J. (2011). Microbiome and Malignancy. *Cell Host & Microbe*, *10*(4), 324–335. https://doi.org/https://doi.org/10.1016/j.chom.2011.10.003

Plummer, E., & Twin, J. (2015). A Comparison of Three Bioinformatics Pipelines for the Analysis of Preterm Gut Microbiota using 16S rRNA Gene Sequencing Data. *Journal of Proteomics & Bioinformatics*, *8*(12). https://doi.org/doi: 10.4172/jpb.1000381

Pollock, J., Glendinning, L., Wisedchanwet, T., & Watson, M. (2018). The Madness of Microbiome: Attempting To Find Consensus "Best Practice" for 16S Microbiome Studies. *Applied and Environmental Microbiology*, *84*(7), e02627-17. https://doi.org/10.1128/AEM.02627-17

Pop, M., Walker, A. W., Paulson, J., Lindsay, B., Antonio, M., Hossain, M. A., … Stine, O. C. (2014). Diarrhea in young children from low-income countries leads to large-scale alterations in intestinal microbiota composition. *Genome Biology*, *15*(6), R76. https://doi.org/10.1186/gb-2014-15-6-r76

Porter, C. M., Shrestha, E., Peiffer, L. B., & Sfanos, K. S. (2018). The microbiome in prostate inflammation and prostate cancer. *Prostate Cancer and Prostatic Diseases*, *21*(3), 345–354. https://doi.org/10.1038/s41391-018-0041-1

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., & Glöckner, F. O. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*, *35*(21), 7188–7196. https://doi.org/10.1093/nar/gkm864

Qiong, W., M., G. G., M., T. J., & R., C. J. (2007). Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Applied and Environmental Microbiology*, *73*(16), 5261–5267. https://doi.org/10.1128/AEM.00062-07

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., … Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, *41*(D1), 590–596. https://doi.org/10.1093/nar/gks1219

Quinn, T. P. (2021). Stool Studies Don't Pass the Sniff Test: A Systematic Review of Human Gut Microbiome Research Suggests Widespread Misuse of Machine Learning. *ArXiv Preprint ArXiv:2107.03611*.

R Development Core Team. (2008). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. Retrieved from http://www.r-project.org

Rajilić-Stojanović, M., & de Vos, W. M. (2014). The first 1000 cultured species of the human gastrointestinal microbiota. *FEMS Microbiology Reviews*, *38*(5), 996–1047. https://doi.org/10.1111/1574-6976.12075

Raju, S. C., Lagström, S., Ellonen, P., de Vos, W. M., Eriksson, J. G., Weiderpass, E., & Rounge, T. B. (2019). Gender-Specific Associations Between Saliva Microbiota and Body Size. *Frontiers in Microbiology*, *10*. https://doi.org/10.3389/fmicb.2019.00767

Renson, A., Jones, H. E., Beghini, F., Segata, N., Zolnik, C. P., Usyk, M., … Dowd, J. B. (2019). Sociodemographic variation in the oral microbiome. *Annals of Epidemiology*, *35*, 73-80.e2. https://doi.org/https://doi.org/10.1016/j.annepidem.2019.03.006

Rigottier-Gois, L. (2013). Dysbiosis in inflammatory bowel diseases: the oxygen hypothesis. *The ISME Journal*, *7*(7), 1256–1261. https://doi.org/10.1038/ismej.2013.80

Risso, D., Ngai, J., Speed, T. P., & Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology*, *32*(9), 896–902. https://doi.org/10.1038/nbt.2931

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47–e47. https://doi.org/10.1093/nar/gkv007

Rizzatti, G., Lopetuso, L. R., Gibiino, G., Binda, C., & Gasbarrini, A. (2017). Proteobacteria: A Common Factor in Human Diseases. *BioMed Research International*, *2017*, 9351507. https://doi.org/10.1155/2017/9351507

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 77. https://doi.org/10.1186/1471-2105-12-77

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3), R25. https://doi.org/10.1186/gb-2010-11-3-r25

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, *4*, e2584. https://doi.org/10.7717/peerj.2584

Rooks, M. G., Veiga, P., Wardwell-Scott, L. H., Tickle, T., Segata, N., Michaud, M., … Garrett, W. S. (2014). Gut microbiome composition and function in experimental colitis during active disease and treatment-induced remission. *The ISME Journal*, *8*(7), 1403–1417. https://doi.org/10.1038/ismej.2014.3

Rosato, A., Barone, M., Negroni, A., Brigidi, P., Fava, F., Xu, P., … Zanaroli, G. (2020). Microbial colonization of different microplastic types and biotransformation of sorbed PCBs by a marine anaerobic bacterial community. *The Science of the Total Environment*, *705*, 135790. https://doi.org/10.1016/j.scitotenv.2019.135790

Ross, M. C., Muzny, D. M., McCormick, J. B., Gibbs, R. A., Fisher-Hoch, S. P., & Petrosino, J. F. (2015). 16S gut community of the Cameron County Hispanic Cohort. *Microbiome*, *3*(1), 7. https://doi.org/10.1186/s40168-015-0072-y

Sabharwal, A., Ganley, K., Miecznikowski, J. C., Haase, E. M., Barnes, V., & Scannapieco, F. A. (2019). The salivary microbiome of diabetic and non-diabetic adults with periodontal disease. *Journal of Periodontology*, *90*(1), 26–34. https://doi.org/https://doi.org/10.1002/JPER.18-0167

Saeb, A. T. M., Al-Rubeaan, K. A., Aldosary, K., Udaya Raja, G. K., Mani, B., Abouelhoda, M., & Tayeb, H. T. (2019). Relative reduction of biological and phylogenetic diversity of the oral microbiota of diabetes and pre-diabetes patients. *Microbial Pathogenesis*, *128*, 215–229. https://doi.org/https://doi.org/10.1016/j.micpath.2019.01.009

Said, H. S., Suda, W., Nakagome, S., Chinen, H., Oshima, K., Kim, S., … Hattori, M. (2013). Dysbiosis of Salivary Microbiota in Inflammatory Bowel Disease and Its Association With Oral Immunological Biomarkers. *DNA Research*, *21*(1), 15–25. https://doi.org/10.1093/dnares/dst037

Samara, J., Moossavi, S., Alshaikh, B., Ortega, V. A., Pettersen, V. K., Ferdous, T., … Arrieta, M.-C. (2022). Supplementation with a probiotic mixture accelerates gut microbiome maturation and reduces intestinal inflammation in extremely preterm infants. *Cell Host & Microbe*, *30*(5), 696-711.e5. https://doi.org/10.1016/j.chom.2022.04.005

Sanapareddy, N., Legge, R. M., Jovov, B., McCoy, A., Burcal, L., Araujo-Perez, F., … Keku, T. O. (2012). Increased rectal microbial richness is associated with the presence of colorectal adenomas in humans. *The ISME Journal*, *6*(10), 1858–1868. https://doi.org/10.1038/ismej.2012.43

Sanders, H. L. (1968). Marine Benthic Diversity: A Comparative Study. *The American Naturalist*, *102*(925), 243–282. https://doi.org/10.1086/282541

Sanna, S., van Zuydam, N. R., Mahajan, A., Kurilshikov, A., Vich Vila, A., Võsa, U., … McCarthy, M. I. (2019). Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nature Genetics*, *51*(4), 600–605. https://doi.org/10.1038/s41588-019-0350-x

Scheperjans, F., Aho, V., Pereira, P. A. B., Koskinen, K., Paulin, L., Pekkonen, E., … Auvinen, P. (2015). Gut microbiota are related to Parkinson's disease and clinical phenotype. *Movement Disorders*, *30*(3), 350–358. https://doi.org/https://doi.org/10.1002/mds.26069

Scher, J. U., Sczesnak, A., Longman, R. S., Segata, N., Ubeda, C., Bielski, C., … Littman, D. R. (2013). Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. *ELife*, *2*, e01202. https://doi.org/10.7554/eLife.01202

Schloss, P. D. (2018). Identifying and Overcoming Threats to Reproducibility, Replicability, Robustness, and Generalizability in Microbiome Research. *MBio*, *9*(3), e00525-18. https://doi.org/10.1128/mBio.00525-18

Schloss, P. D. (2020). Removal of rare amplicon sequence variants from 16S rRNA gene sequence surveys biases the interpretation of community structure data. *BioRxiv*, 2020.12.11.422279. https://doi.org/10.1101/2020.12.11.422279

Schmidt, T. S. B., Hayward, M. R., Coelho, L. P., Li, S. S., Costea, P. I., Voigt, A. Y., … Bork, P. (2019). Extensive transmission of microbes along the gastrointestinal tract. *ELife*, *8*, e42693. https://doi.org/10.7554/eLife.42693

Schneider, D., Thürmer, A., Gollnow, K., Lugert, R., Gunka, K., Groß, U., & Daniel, R. (2017). Gut bacterial communities of diarrheic patients with indications of Clostridioides difficile infection. *Scientific Data*, *4*(1), 170152. https://doi.org/10.1038/sdata.2017.152

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., … Schindel, D. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences*, *109*(16), 6241–6246. https://doi.org/10.1073/pnas.1117018109

Schrøder, S. A., Bardow, A., Eickhardt-Dalbøge, S., Johansen, H. K., & Homøe, P. (2017). Is parotid saliva sterile on entry to the oral cavity? *Acta Oto-Laryngologica*, *137*(7), 762–764. https://doi.org/10.1080/00016489.2016.1272002

Schubert, A. M., Rogers, M. A. M., Ring, C., Mogle, J., Petrosino, J. P., Young, V. B., … Schloss, P. D. (2014). Microbiome Data Distinguish Patients with &lt;span class=&quot;named-content genus-species&quot; id=&quot;named-content-1&quot;&gt;Clostridium difficile&lt;/span&gt; Infection and Non-&lt;span class=&quot;named-content genus-species&quot; id=&quot;named-c. *MBio*, *5*(3), e01021-14. https://doi.org/10.1128/mBio.01021-14

Segata, N., Haake, S. K., Mannon, P., Lemon, K. P., Waldron, L., Gevers, D., … Izard, J. (2012). Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome Biol*, *13*. https://doi.org/10.1186/gb-2012-13-6-r42

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., & Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. *Genome Biology*, *12*(6), R60. https://doi.org/10.1186/gb-2011-12-6-r60

Sender, R., Fuchs, S., & Milo, R. (2016). Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLOS Biology*, *14*(8), e1002533. Retrieved from https://doi.org/10.1371/journal.pbio.1002533

Shaffer, J. P., Marotz, C., Belda-Ferre, P., Martino, C., Wandro, S., Estaki, M., … Knight, R. (2021). A comparison of DNA/RNA extraction protocols for high-throughput sequencing of microbial communities. *BioTechniques*, *70*(3), 149–159. https://doi.org/10.2144/btn-2020-0153

Sharon, G., Cruz, N. J., Kang, D.-W., Gandal, M. J., Wang, B., Kim, Y.-M., … Mazmanian, S. K. (2019). Human Gut Microbiota from Autism Spectrum Disorder Promote Behavioral Symptoms in Mice. *Cell*, *177*(6), 1600-1618.e17. https://doi.org/https://doi.org/10.1016/j.cell.2019.05.004

Shin, N.-R., Whon, T. W., & Bae, J.-W. (2015). Proteobacteria: microbial signature of dysbiosis in gut microbiota. *Trends in Biotechnology*, *33*(9), 496–503. https://doi.org/https://doi.org/10.1016/j.tibtech.2015.06.011

Shungin, D., Haworth, S., Divaris, K., Agler, C. S., Kamatani, Y., Keun Lee, M., … Johansson, I. (2019). Genome-wide analysis of dental caries and periodontitis combining clinical and self-reported data. *Nature Communications*, *10*(1), 2773. https://doi.org/10.1038/s41467-019-10630-1

Silverman, J. D., Washburne, A. D., Mukherjee, S., & David, L. A. (2017). A phylogenetic transform enhances analysis of compositional microbiota data. *ELife*, *6*, e21887. https://doi.org/10.7554/eLife.21887

Simón-Soro, A., & Mira, A. (2015). Solving the etiology of dental caries. *Trends in Microbiology*, *23*(2), 76–82. https://doi.org/https://doi.org/10.1016/j.tim.2014.10.010

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics*, *21*(20), 3940–3941. https://doi.org/10.1093/bioinformatics/bti623

Singh, P., Teal, T. K., Marsh, T. L., Tiedje, J. M., Mosci, R., Jernigan, K., … Manning, S. D. (2015). Intestinal microbial communities associated with acute enteric infections and disease recovery. *Microbiome*, *3*(1), 45. https://doi.org/10.1186/s40168-015-0109-2

Sinha, R., Abu-Ali, G., Vogtmann, E., Fodor, A. A., Ren, B., Amir, A., … Huttenhower, C. (2017). Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. *Nature Biotechnology*, *35*(11), 1077–1086. https://doi.org/10.1038/nbt.3981

Sneath, P. H. A., & Sokal, R. R. (1973). *Numerical taxonomy. The principles and practice of numerical classification.*

Socransky, S. S., Haffajee, A. D., Cugini, M. A., Smith, C., & Kent Jr., R. L. (1998). Microbial complexes in subgingival plaque. *Journal of Clinical Periodontology*, *25*(2), 134–144. https://doi.org/10.1111/j.1600-051X.1998.tb02419.x

Son, J. S., Zheng, L. J., Rowehl, L. M., Tian, X., Zhang, Y., Zhu, W., … Li, E. (2015). Comparison of Fecal Microbiota in Children with Autism Spectrum Disorders and Neurotypical Siblings in the Simons Simplex Collection. *PLOS ONE*, *10*(10), e0137725. Retrieved from https://doi.org/10.1371/journal.pone.0137725

Sonnenborn, U. (2016). Escherichia coli strain Nissle 1917—from bench to bedside and back: history of a special Escherichia coli strain with probiotic properties. *FEMS Microbiology Letters*, *363*(19), fnw212. https://doi.org/10.1093/femsle/fnw212

Stackebrandt, E., & Goebel, B. M. (1994). Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *International Journal of Systematic and Evolutionary Microbiology*, *44*(4), 846–849. https://doi.org/10.1099/00207713-44-4-846

Stahl, D. A., Lane, D. J., Olsen, G. J., & Pace, N. R. (1984). Analysis of Hydrothermal Vent-Associated Symbionts by Ribosomal RNA Sequences. *Science*, *224*(4647), 409–411. https://doi.org/10.1126/science.224.4647.409

Staley, J. T., & Konopka, A. (1985). Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annual Review of Microbiology*, *39*(1), 321–346.

Suau, A., Bonnet, R., Sutren, M., Godon, J.-J., Gibson, G. R., Collins, M. D., & Doré, J. (1999). Direct Analysis of Genes Encoding 16S rRNA from Complex Communities Reveals Many Novel Molecular Species within the Human Gut. *Applied and Environmental Microbiology*, *65*(11), 4799–4807. https://doi.org/10.1128/AEM.65.11.4799-4807.1999

Sul, W. J., Cole, J. R., Jesus, E. da C., Wang, Q., Farris, R. J., Fish, J. A., & Tiedje, J. M. (2011). Bacterial community comparisons by taxonomy-supervised analysis independent of sequence alignment and clustering. *Proceedings of the National Academy of Sciences*, *108*(35), 14637–14642.

Sun, H., Zhao, X., Zhou, Y., Wang, J., Ma, R., Ren, X., … Zou, L. (2020). Characterization of oral microbiome and exploration of potential biomarkers in patients with pancreatic cancer. *BioMed Research International*, *2020*.

Sun, S., Luo, L., Liang, W., Yin, Q., Guo, J., Rush, A. M., … Wang, F. (2020). Bifidobacterium alters the gut microbiota and modulates the functional metabolism of T regulatory cells in the context of immune checkpoint blockade. *Proceedings of the National Academy of Sciences*, *117*(44), 27509–27515. https://doi.org/10.1073/pnas.1921223117

Sweeney, E., Cui, Y., DeClercq, V., Devichand, P., Forbes, C., Grandy, S., … Dummer, T. J. B. (2017). Cohort Profile: The Atlantic Partnership for Tomorrow's Health (Atlantic PATH) Study. *International Journal of Epidemiology*, *46*(6), 1762-1763i. Retrieved from http://dx.doi.org/10.1093/ije/dyx124

Szymanski, M., Barciszewska, M. Z., Erdmann, V. A., & Barciszewski, J. (2002). 5S Ribosomal RNA Database. *Nucleic Acids Research*, *30*(1), 176–178. https://doi.org/10.1093/nar/30.1.176

Takahashi, N., & Nyvad, B. (2010). The Role of Bacteria in the Caries Process: Ecological Perspectives. *Journal of Dental Research*, *90*(3), 294–303. https://doi.org/10.1177/0022034510379602

Takeshita, T., Kageyama, S., Furuta, M., Tsuboi, H., Takeuchi, K., Shibata, Y., … Yamashita, Y. (2016). Bacterial diversity in saliva and oral health-related conditions: the Hisayama Study. *Scientific Reports*, *6*(1), 22164. https://doi.org/10.1038/srep22164

Tamaki, H., Nakase, H., Inoue, S., Kawanami, C., Itani, T., Ohana, M., … Shibatouge, M. (2016). Efficacy of probiotic treatment with Bifidobacterium longum 536 for induction of remission in active ulcerative colitis: A randomized, double-blinded, placebo-controlled multicenter trial. *Digestive Endoscopy*, *28*(1), 67–74. https://doi.org/https://doi.org/10.1111/den.12553

Tange, O. (2011). GNU Parallel: the command-line power tool. *;;Login: The USENIX Magazine*, *36*(1), 42–47. https://doi.org/10.5281/zenodo.16303

Thomas, A. M., Manghi, P., Asnicar, F., Pasolli, E., Armanini, F., Zolfo, M., … Segata, N. (2019). Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nature Medicine*, *25*(4), 667–678. https://doi.org/10.1038/s41591-019-0405-7

Thorsen, J., Brejnrod, A., Mortensen, M., Rasmussen, M. A., Stokholm, J., Al-Soud, W. A., … Waage, J. (2016). Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. *Microbiome*, *4*(1), 62. https://doi.org/10.1186/s40168-016-0208-8

Tikkanen, M. J., Adlercreutz, H., & Pulkkinen, M. O. (1973). Effects of antibiotics on oestrogen metabolism. *British Medical Journal*, *2*(5862), 369 LP – 369. https://doi.org/10.1136/bmj.2.5862.369

Tikkanen, M. J., Pulkkinen, M. O., & Adlercreutz, H. (1973). Effect of ampicillin treatment on the urinary excretion of estriol conjugates in pregnancy. *Journal of Steroid Biochemistry*, *4*(4), 439–440. https://doi.org/https://doi.org/10.1016/0022-4731(73)90015-0

Tipton, L., Darcy, J. L., & Hynson, N. A. (2019). A developing symbiosis: enabling cross-talk between ecologists and microbiome scientists. *Frontiers in Microbiology*, *10*, 292.

Tomás, I., Diz, P., Tobías, A., Scully, C., & Donos, N. (2012). Periodontal health status and bacteraemia from daily oral activities: systematic review/meta-analysis. *Journal of Clinical Periodontology*, *39*(3), 213–228. https://doi.org/10.1111/j.1600-051X.2011.01784.x

Torres, P. J., Fletcher, E. M., Gibbons, S. M., Bouvet, M., Doran, K. S., & Kelley, S. T. (2015). Characterization of the salivary microbiome in patients with pancreatic cancer. *PeerJ*, *3*, e1373.

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., … Gordon, J. I. (2009). A core gut microbiome in obese and lean twins. *Nature*, *457*(7228), 480–484. https://doi.org/10.1038/nature07540

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The human microbiome project. *Nature*, *449*. https://doi.org/10.1038/nature06244

Uemura, N., Okamoto, S., Yamamoto, S., Matsumura, N., Yamaguchi, S., Yamakido, M., … Schlemper, R. J. (2001). Helicobacter pylori infection and the development of gastric cancer. *New England Journal of Medicine*, *345*(11), 784–789.

Vebø, H. C., Karlsson, M. K., Avershina, E., Finnby, L., & Rudi, K. (2016). Bead-beating artefacts in the Bacteroidetes to Firmicutes ratio of the human stool metagenome. *Journal of Microbiological Methods*, *129*, 78–80. https://doi.org/https://doi.org/10.1016/j.mimet.2016.08.005

Vincent, C., Stephens, D. A., Loo, V. G., Edens, T. J., Behr, M. A., Dewar, K., & Manges, A. R. (2013). Reductions in intestinal Clostridiales precede the development of nosocomial *Clostridium difficile* infection. *Microbiome*, *1*(1), 18. https://doi.org/10.1186/2049-2618-1-18

Vogtmann, E., Han, Y., Caporaso, J. G., Bokulich, N., Mohamadkhani, A., Moayyedkazemi, A., … Suman, S. (2020). Oral microbial community composition is associated with pancreatic cancer: A case-control study in Iran. *Cancer Medicine*, *9*(2), 797–806.

Vogtmann, E., Hua, X., Yu, G., Purandare, V., Hullings, A. G., Shao, D., … Abnet, C. C. (2022a). The oral microbiome and lung cancer risk: An analysis of 3 prospective cohort studies. *JNCI: Journal of the National Cancer Institute*, djac149. https://doi.org/10.1093/jnci/djac149

Vogtmann, E., Hua, X., Yu, G., Purandare, V., Hullings, A. G., Shao, D., … Abnet, C. C. (2022b). The oral microbiome and lung cancer risk: An analysis of 3 prospective cohort studies. *JNCI: Journal of the National Cancer Institute*, djac149. https://doi.org/10.1093/jnci/djac149

Vos, P., Garrity, G., Jones, D., Krieg, N. R., Ludwig, W., Rainey, F. A., … Whitman, W. B. (2011). *Bergey's manual of systematic bacteriology: Volume 3: The Firmicutes* (Vol. 3). Springer Science & Business Media.

Waardenberg, A. J., & Field, M. A. (2019). consensusDE: an R package for assessing consensus of multiple RNA-seq algorithms with RUV correction. *PeerJ*, *7*, e8206. https://doi.org/10.7717/peerj.8206

Wade, W. G. (2013). The oral microbiome in health and disease. *Pharmacological Research*, *69*(1), 137–143. https://doi.org/https://doi.org/10.1016/j.phrs.2012.11.006

Wan, Y., Zuo, T., Xu, Z., Zhang, F., Zhan, H., CHAN, D., … Ng, S. C. (2022). Underdevelopment of the gut microbiota and bacteria species as non-invasive markers of prediction in children with autism spectrum disorder. *Gut*, *71*(5), 910 LP – 918. https://doi.org/10.1136/gutjnl-2020-324015

Wang, H., Altemus, J., Niazi, F., Green, H., Calhoun, B. C., Sturgis, C., … Eng, C. (2017). Breast tissue, oral and urinary microbiomes in breast cancer. *Oncotarget; Vol 8, No 50*. Retrieved from https://www.oncotarget.com/article/21490/text/

Wang, Yao, Zhang, Y., Qian, Y., Xie, Y.-H., Jiang, S.-S., Kang, Z.-R., … Fang, J.-Y. (2021). Alterations in the oral and gut microbiome of colorectal cancer patients and association with host clinical factors. *International Journal of Cancer*, *149*(4), 925–935. https://doi.org/https://doi.org/10.1002/ijc.33596

Wang, Yiwen, & LêCao, K.-A. (2020). Managing batch effects in microbiome data. *Briefings in Bioinformatics*, *21*(6), 1954–1970. https://doi.org/10.1093/bib/bbz105

Washburne, A. D., Silverman, J. D., Leff, J. W., Bennett, D. J., Darcy, J. L., Mukherjee, S., … David, L. A. (2017). Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. *PeerJ*, *5*, e2969. https://doi.org/10.7717/peerj.2969

Wei, Y., Zhong, Y., Wang, Y., & Huang, R. (2021). Association between periodontal disease and prostate cancer: a systematic review and meta-analysis. *Medicina Oral, Patología Oral y Cirugía Bucal*, *26*(4), e459.

Weiss, S. J., Xu, Z., Amir, A., Peddada, S., Bittinger, K., Gonzalez, A., … Birmingham, A. (2015). *Effects of library size variance, sparsity, and compositionality on the analysis of microbiome data*. PeerJ PrePrints.

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., … Knight, R. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, *5*(1), 27. https://doi.org/10.1186/s40168-017-0237-y

Westcott, S. L., & Schloss, P. D. (2015). De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ*, *3*, e1487. https://doi.org/10.7717/peerj.1487

Whittaker, R. H. (1972). EVOLUTION AND MEASUREMENT OF SPECIES DIVERSITY. *TAXON*, *21*(2–3), 213–251. https://doi.org/https://doi.org/10.2307/1218190

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. Retrieved from http://ggplot2.org

Willis, A., Bunge, J., & Whitman, T. (2017). Improved detection of changes in species richness in high diversity microbial communities. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *66*(5), 963–977. https://doi.org/https://doi.org/10.1111/rssc.12206

Willis, A. D. (2019). Rarefaction, alpha diversity, and statistics. *Frontiers in Microbiology*, *10*, 2407.

Willis, C., Desai, D., & LaRoche, J. (2019). Influence of 16S rRNA variable region on perceived diversity of marine microbial communities of the Northern North Atlantic. *FEMS Microbiology Letters*, *366*(13), fnz152. https://doi.org/10.1093/femsle/fnz152

Willman, K., & Pulkkinen, M. O. (1971). Reduced maternal plasma and urinary estriol during ampicillin treatment. *American Journal of Obstetrics and Gynecology*, *109*(6), 893–896. https://doi.org/https://doi.org/10.1016/0002-9378(71)90803-9

Wilson, K. H., & Blitchington, R. B. (1996). Human colonic biota studied by ribosomal DNA sequence analysis. *Applied and Environmental Microbiology*, *62*(7), 2273–2278. https://doi.org/10.1128/aem.62.7.2273-2278.1996

Woese, Carl R, & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences*, *74*(11), 5088–5090. https://doi.org/10.1073/pnas.74.11.5088

Woese, Carl Richard. (1987). Bacterial evolution. *Microbiological Reviews*, *51*(2), 221–271. https://doi.org/10.1128/mr.51.2.221-271.1987

Woese, Carl Richard, Fox, G. E., Zablen, L., Uchida, T., Bonen, L., Pechman, K., … Stahl, D. (1975). Conservation of primary structure in 16S ribosomal RNA. *Nature*, *254*(5495), 83–86. https://doi.org/10.1038/254083a0

Wolf, S. A., Epping, L., Andreotti, S., Reinert, K., & Semmler, T. (2021). SCORE: Smart Consensus Of RNA Expression—a consensus tool for detecting differentially expressed genes in bacteria. *Bioinformatics*, *37*(3), 426–428.

Wong, R. G., Wu, J. R., & Gloor, G. B. (2016). Expanding the UniFrac Toolbox. *PLOS ONE*, *11*(9), e0161196. Retrieved from https://doi.org/10.1371/journal.pone.0161196

Wu, J., Peters, B. A., Dominianni, C., Zhang, Y., Pei, Z., Yang, L., … Ahn, J. (2016). Cigarette smoking and the oral microbiome in a large study of American adults. *The Isme Journal*, *10*, 2435. Retrieved from http://dx.doi.org/10.1038/ismej.2016.37

Wu, L., Ning, D., Zhang, B., Li, Y., Zhang, P., Shan, X., … Consortium, G. W. M. (2019). Global diversity and biogeography of bacterial communities in wastewater treatment plants. *Nature Microbiology*, *4*(7), 1183–1195. https://doi.org/10.1038/s41564-019-0426-5

Wu, W.-K., Chen, C.-C., Panyod, S., Chen, R.-A., Wu, M.-S., Sheen, L.-Y., & Chang, S.-C. (2019). Optimization of fecal sample processing for microbiome study — The journey from bathroom to bench. *Journal of the Formosan Medical Association*, *118*(2), 545–555. https://doi.org/https://doi.org/10.1016/j.jfma.2018.02.005

Wu, Yongjian, Cheng, X., Jiang, G., Tang, H., Ming, S., Tang, L., … Huang, X. (2021). Altered oral and gut microbiota and its association with SARS-CoV-2 viral load in COVID-19 patients during hospitalization. *Npj Biofilms and Microbiomes*, *7*(1), 61. https://doi.org/10.1038/s41522-021-00232-5

Wu, Yujia, Chi, X., Zhang, Q., Chen, F., & Deng, X. (2018). Characterization of the salivary microbiome in people with obesity. *PeerJ*, *6*, e4458. https://doi.org/10.7717/peerj.4458

Wu, Z., Byrd, D. A., Wan, Y., Ansong, D., Clegg-Lamptey, J.-N., Wiafe-Addai, B., … Vogtmann, E. (2022). The oral microbiome and breast cancer and nonmalignant breast disease, and its relationship with the fecal microbiome in the Ghana Breast Health Study. *International Journal of Cancer*, *151*(8), 1248–1260. https://doi.org/https://doi.org/10.1002/ijc.34145

Xu, X., He, J., Xue, J., Wang, Y., Li, K., Zhang, K., … Zhou, X. (2015). Oral cavity contains distinct niches with dynamic microbial communities. *Environmental Microbiology*, *17*(3), 699–710. https://doi.org/https://doi.org/10.1111/1462-2920.12502

Xun, Z., Zhang, Q., Xu, T., Chen, N., & Chen, F. (2018). Dysbiosis and ecotypes of the salivary microbiome associated with inflammatory bowel diseases and the assistance in diagnosis of diseases using oral bacterial profiles. *Frontiers in Microbiology*, *9*(MAY). https://doi.org/10.3389/fmicb.2018.01136

Yang, C.-Y., Yeh, Y.-M., Yu, H.-Y., Chin, C.-Y., Hsu, C.-W., Liu, H., … Chang, K.-P. (2018). Oral microbiota community dynamics associated with oral squamous cell carcinoma staging. *Frontiers in Microbiology*, *9*, 862.

Yang, Y., Cai, Q., Shu, X.-O., Steinwandel, M. D., Blot, W. J., Zheng, W., & Long, J. (2019). Prospective study of oral microbiome and colorectal cancer risk in low-income and African American populations. *International Journal of Cancer*, *144*(10), 2381–2389. https://doi.org/https://doi.org/10.1002/ijc.31941

Yang, Y., Cai, Q., Zheng, W., Steinwandel, M., Blot, W. J., Shu, X.-O., & Long, J. (2019). Oral microbiome and obesity in a large study of low-income and African-American populations. *Journal of Oral Microbiology*, *11*(1), 1650597. https://doi.org/10.1080/20002297.2019.1650597

Yaohua, Y., Wei, Z., Qiuyin, C., J., S. M., Zhiheng, P., Robert, B., … Jirong, L. (2019). Racial Differences in the Oral Microbiome: Data from Low-Income Populations of African Ancestry and European Ancestry. *MSystems*, *4*(6), e00639-19. https://doi.org/10.1128/mSystems.00639-19

Yap, C. X., Henders, A. K., Alvares, G. A., Wood, D. L. A., Krause, L., Tyson, G. W., … Gratten, J. (2021). Autism-related dietary preferences mediate autism-gut microbiome associations. *Cell*, *184*(24), 5916-5931.e17. https://doi.org/10.1016/j.cell.2021.10.015

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., … Gordon, J. I. (2012). Human gut microbiome viewed across age and geography. *Nature*, *486*, 222. Retrieved from http://dx.doi.org/10.1038/nature11053

Ye, M., Robson, P. J., Eurich, D. T., Vena, J. E., Xu, J.-Y., & Johnson, J. A. (2017). Cohort Profile: Alberta’s Tomorrow Project. *International Journal of Epidemiology*, *46*(4), 1097-1098l. https://doi.org/10.1093/ije/dyw256

Yow, M. A., Tabrizi, S. N., Severi, G., Bolton, D. M., Pedersen, J., BioResource, A. P. C., … Southey, M. C. (2017). Characterisation of microbial communities within aggressive prostate cancer tissues. *Infectious Agents and Cancer*, *12*, 4. https://doi.org/10.1186/s13027-016-0112-7

Yurgel, S. N., Douglas, G. M., Comeau, A. M., Mammoliti, M., Dusault, A., Percival, D., & Langille, M. G. I. (2017). Variation in Bacterial and Eukaryotic Communities Associated with Natural and Managed Wild Blueberry Habitats. *Phytobiomes*, *1*(2), 102–113. https://doi.org/10.1094/pbiomes-03-17-0012-r

Zaura, E., Keijser, B. J., Huse, S. M., & Crielaard, W. (2009). Defining the healthy “core microbiome” of oral microbial communities. *BMC Microbiol*, *9*. https://doi.org/10.1186/1471-2180-9-259

Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kultima, J. R., Costea, P. I., … Bork, P. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular Systems Biology*, *10*(11), 766–766. https://doi.org/10.15252/msb.20145645

Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, *30*(5), 614–620. https://doi.org/10.1093/bioinformatics/btt593

Zhang, S., Kong, C., Yang, Y., Cai, S., Li, X., Cai, G., & Ma, Y. (2020). Human oral microbiome dysbiosis as a novel non-invasive biomarker in detection of colorectal cancer. *Theranostics*, *10*(25), 11595–11606. https://doi.org/10.7150/thno.49515

Zhang, Yong, Kang, N., Xue, F., Qiao, J., Duan, J., Chen, F., & Cai, Y. (2021). Evaluation of salivary biomarkers for the diagnosis of periodontitis. *BMC Oral Health*, *21*(1), 266. https://doi.org/10.1186/s12903-021-01600-5

Zhang, Yuqing, Parmigiani, G., & Johnson, W. E. (2020). ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genomics and Bioinformatics*, *2*(3), lqaa078.

Zhao, L., Cho, W. C., & Nicolls, M. R. (2021). Colorectal cancer-associated microbiome patterns and signatures. *Frontiers in Genetics*, *12*.

Zhu, J., Liao, M., Yao, Z., Liang, W., Li, Q., Liu, J., … Mo, Z. (2018). Breast cancer in postmenopausal women is associated with an altered gut metagenome. *Microbiome*, *6*(1), 136. https://doi.org/10.1186/s40168-018-0515-3

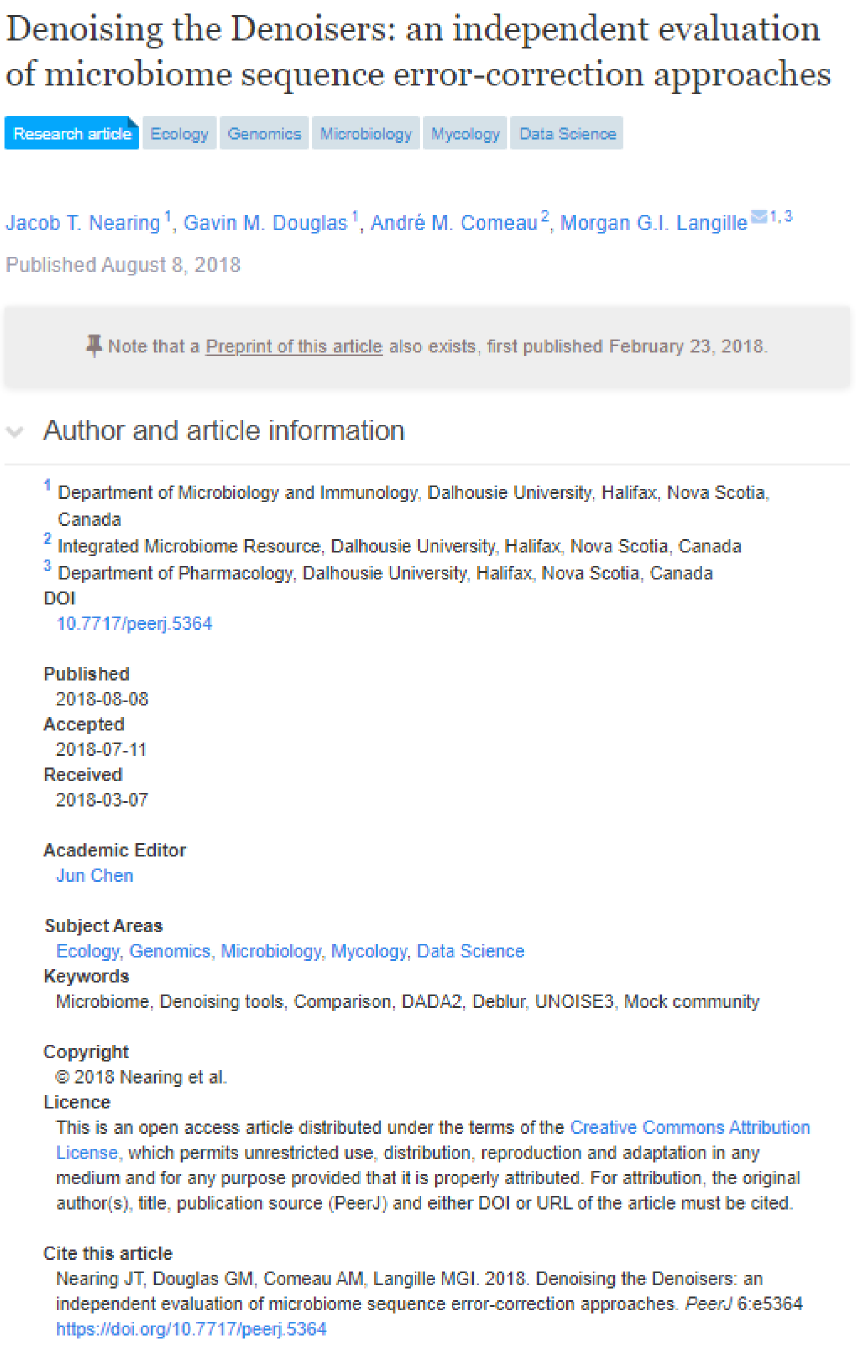
Zhu, L., Baker, S. S., Gill, C., Liu, W., Alkhouri, R., Baker, R. D., & Gill, S. R. (2013). Characterization of gut microbiomes in nonalcoholic steatohepatitis (NASH) patients: A connection between endogenous alcohol and NASH. *Hepatology*, *57*(2), 601–609. https://doi.org/https://doi.org/10.1002/hep.26093

Zhu, W., Shen, W., Wang, J., Xu, Y., Zhai, R., Zhang, J., … Liu, L. (2022). Capnocytophaga gingivalis is a Potential Tumor Promotor in Oral Cancer. *Oral Diseases*, *n/a*(n/a). https://doi.org/https://doi.org/10.1111/odi.14376

Zupancic, M. L., Cantarel, B. L., Liu, Z., Drabek, E. F., Ryan, K. A., Cirimotich, S., … Fraser, C. M. (2012). Analysis of the gut microbiota in the old order amish and its relation to the metabolic syndrome. *PLoS ONE*, *7*(8), e43052. https://doi.org/10.1371/journal.pone.0043052

Appendices

Copy Right Permissions



Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches

Jacob T. Nearing[1], Gavin M. Douglas[1], André M. Comeau[2], Morgan G.I. Langille[1,3]

Published August 8, 2018

Note that a Preprint of this article also exists, first published February 23, 2018.

Author and article information

[1] Department of Microbiology and Immunology, Dalhousie University, Halifax, Nova Scotia, Canada
[2] Integrated Microbiome Resource, Dalhousie University, Halifax, Nova Scotia, Canada
[3] Department of Pharmacology, Dalhousie University, Halifax, Nova Scotia, Canada

DOI
10.7717/peerj.5364

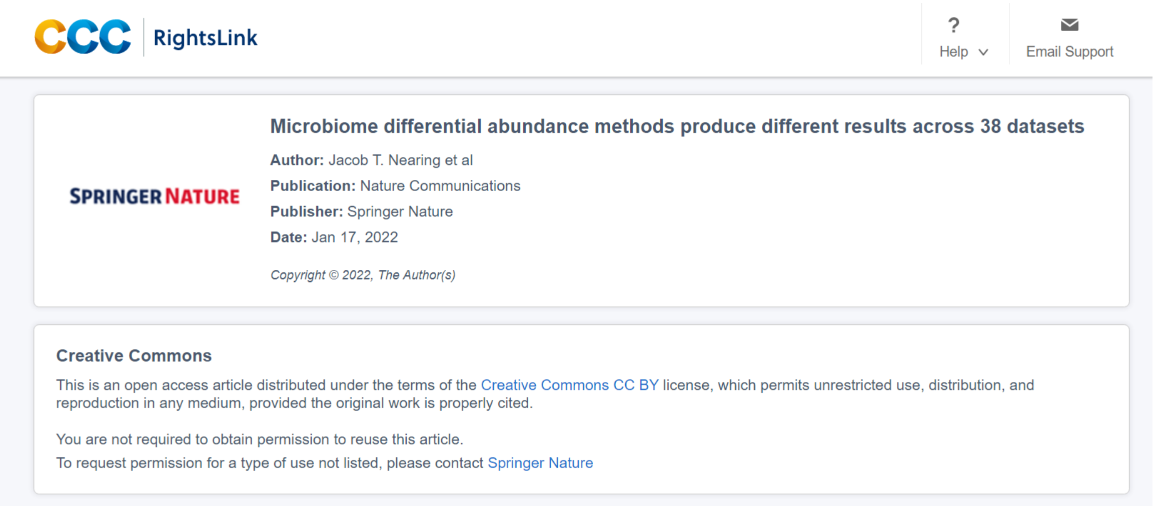
Published
2018-08-08
Accepted
2018-07-11
Received
2018-03-07

Academic Editor
Jun Chen

Subject Areas
Ecology, Genomics, Microbiology, Mycology, Data Science
Keywords
Microbiome, Denoising tools, Comparison, DADA2, Deblur, UNOISE3, Mock community

Copyright
© 2018 Nearing et al.
Licence
This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, reproduction and adaptation in any medium and for any purpose provided that it is properly attributed. For attribution, the original author(s), title, publication source (PeerJ) and either DOI or URL of the article must be cited.

Cite this article
Nearing JT, Douglas GM, Comeau AM, Langille MGI. 2018. Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. *PeerJ* 6:e5364 https://doi.org/10.7717/peerj.5364



Microbiome differential abundance methods produce different results across 38 datasets

Author: Jacob T. Nearing et al

Publication: Nature Communications

Publisher: Springer Nature

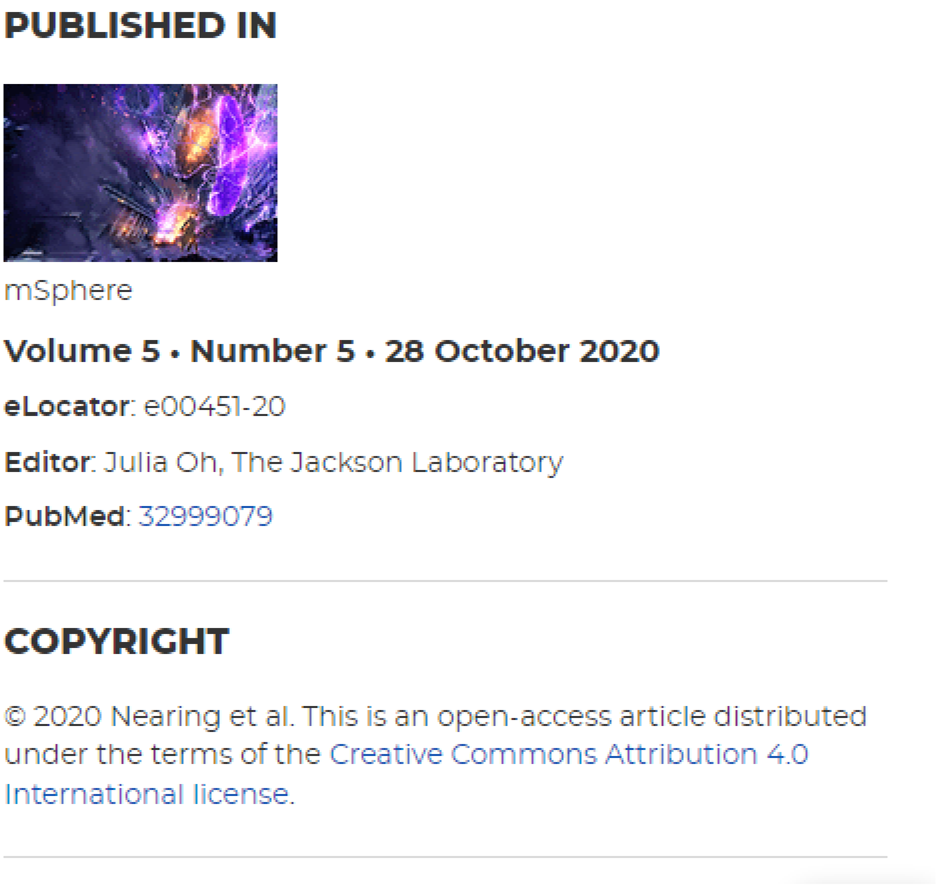
Date: Jan 17, 2022

Copyright © 2022, The Author(s)

Creative Commons

This is an open access article distributed under the terms of the Creative Commons CC BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.
To request permission for a type of use not listed, please contact Springer Nature



PUBLISHED IN

mSphere

Volume 5 • Number 5 • 28 October 2020

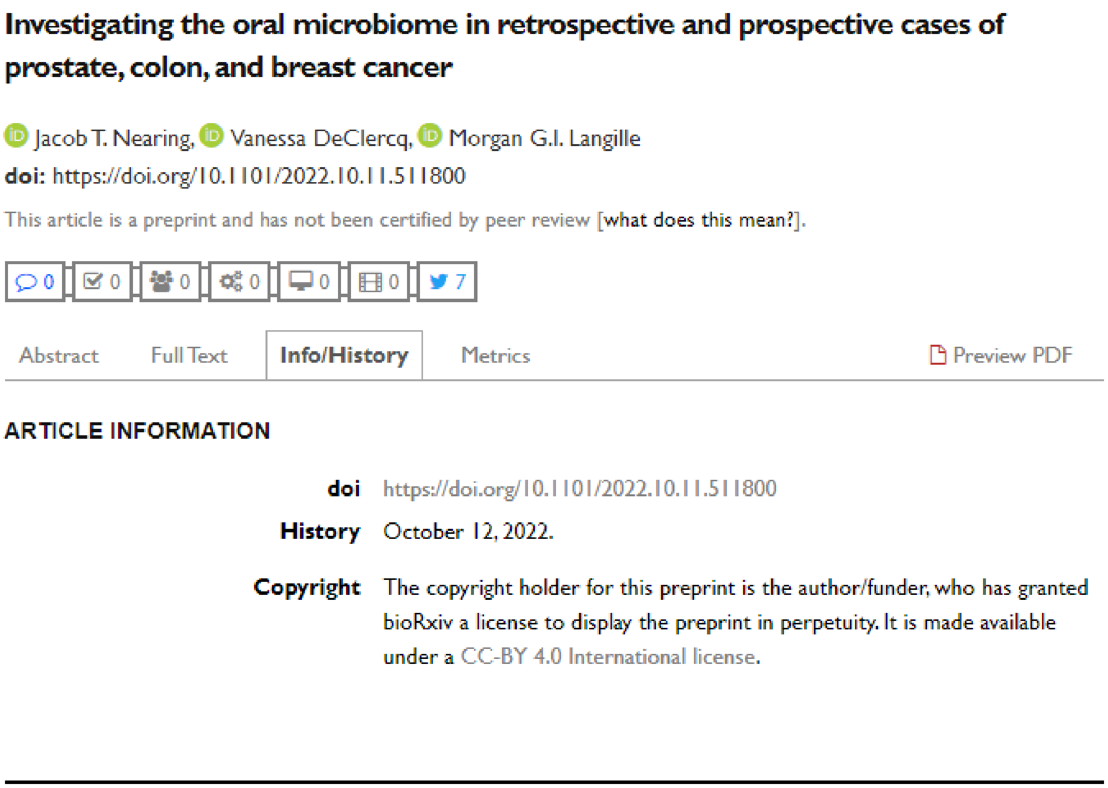
eLocator: e00451-20

Editor: Julia Oh, The Jackson Laboratory

PubMed: 32999079

COPYRIGHT

© 2020 Nearing et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.



Investigating the oral microbiome in retrospective and prospective cases of prostate, colon, and breast cancer

Jacob T. Nearing, Vanessa DeClercq, Morgan G.I. Langille

doi: https://doi.org/10.1101/2022.10.11.511800

This article is a preprint and has not been certified by peer review [what does this mean?].

ARTICLE INFORMATION

doi	https://doi.org/10.1101/2022.10.11.511800
History	October 12, 2022.
Copyright	The copyright holder for this preprint is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.