

2024

Improving batch effect correction of metagenomic data: applications in the black women's health study

<https://hdl.handle.net/2144/47924>

"Downloaded from OpenBU. Boston University's institutional repository."

BOSTON UNIVERSITY
FACULTY OF COMPUTING & DATA SCIENCES

Dissertation

**IMPROVING BATCH EFFECT CORRECTION OF METAGENOMIC DATA:
APPLICATIONS IN THE BLACK WOMEN'S HEALTH STUDY**

by

HOWARD JAMES FAN

B.S., Johns Hopkins University, 2014
M.S., Boston University 2017

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2024

© 2024 by
HOWARD JAMES FAN
All rights reserved

Approved by

First Reader

W. Evan Johnson, Ph.D.
Adjunct Associate Professor of Medicine, Biostatistics, and
Bioinformatics

Second Reader

Trevor W. Siggers, Ph.D.
Associate Professor of Biology

Third Reader

Jessica L. Petrick, Ph.D., MPH
Assistant Professor of Medicine

Fourth Reader

Joshua D. Campbell, Ph.D.
Associate Professor of Medicine

Fifth Reader

Jennifer Bhatnagar, Ph.D.
Associate Professor of Biology

DEDICATION

I would like to dedicate this work to my lovely and patient wife Liyang and my amazing family and friends who encouraged and supporting me throughout this journey.

ACKNOWLEDGMENTS

These past five years have truly been a memorable journey for me, and I could not have made it to the finish line without the support of a whole community of people. Thank you to my primary advisor, W. Evan Johnson, for his incredible guidance and mentorship. With his constantly positive attitude, pure curiosity towards research, and kind attentiveness towards his students, he is an inspiring role model who has kept me afloat throughout the entirety of my PhD journey. Thank you to my collaborators, Julie and Jessica, for their patience and support in brainstorming innovative ideas and discussing the constantly changing analyses of the BWHS data. Thank you to my thesis committee, Trevor, Joshua, and Jennifer, for pushing me to become a better bioinformatics scientist. Thank you to all my friends at Boston University who made my time an enjoyable experience. Special thanks to Ahmed, Rebecca, Jackie, Ethel, and Lucas for being an awesome cohort. Thank you to Josh, my fellow Hopkins friend, to whom I am grateful for his infectious, positive energy. Last, but certainly not least, I must give a special thank you to my loving wife, without whom I could not have reached this point. For challenging my ideas, pushing me to meet my goals, and supporting me when I felt lost, you truly are my partner in life.

**IMPROVING BATCH EFFECT CORRECTION OF METAGENOMIC DATA:
APPLICATIONS IN THE BLACK WOMEN'S HEALTH STUDY**

HOWARD JAMES FAN

Boston University Faculty of Computing & Data Sciences, 2024

Major Professor: W. Evan Johnson, Ph.D., Adjunct Associate Professor of Medicine,
Biostatistics, and Bioinformatics

ABSTRACT

The microbiome has become a focus of research, particularly in the field of human health and precision medicine, due to its role in human development, immunity, and nutrition. Microbiome profiling studies have become more tractable and advanced in large part thanks to advancements in metagenomics. One such study is the Black Women's Health Study (BWHS), which aims to better understand health risks and disease development specific to Black women, who are more susceptible to certain health conditions. However, a major obstacle for reproducibility of microbiome research is the high sensitivity of microbial compositions to external factors and batch-to-batch technical variability, resulting in batch effects that often hinder analysis of factors of interest. While batch effect adjustment methods have been developed for other biomedical data, they do not appropriately account for two unique features of microbiome data: 1) its compositional nature, and 2) extreme overdispersion and zero-inflation.

My dissertation addresses these challenges by evaluating and improving batch effect correction methods for microbiome data and then applies these approaches to data from BWHS. First, I evaluated ComBat-Seq, along with existing microbiome-specific tools, in removing batch effects from both simulated 16S rRNA and real-world shotgun

metagenomic sequencing data while preserving effects belonging to biological factors of interest. Second, I applied ComBat-Seq in an epidemiological study in which I identified several oral health-related genera among adult Black women to be associated with the host's geographic location in the US. Finally, I introduced an extension to ComBat-Seq that improves its performance in batch effect correction on rare taxa with outliers via imputation. I demonstrated that, by replacing zeroes with predicted non-zero read counts that follow the observed compositional structure of the data, imputation effectively reduced the number of problematic cases in which outliers were intensified after batch effect correction.

Collectively, my thesis demonstrates that 1) when the specific features of microbiome data are accounted for, batch effect correction methods offer a promising solution to address batch effect in microbiome data and improve microbiome profiling studies and 2) it is important to consider social/environmental factors associated with the host's physical location when studying the oral microbiome.

TABLE OF CONTENTS

DEDICATION	iv
ACKNOWLEDGMENTS	v
ABSTRACT	vi
TABLE OF CONTENTS.....	viii
LIST OF TABLES.....	xii
LIST OF FIGURES	xiii
LIST OF ABBREVIATIONS.....	xv
Chapter 1: Background, Rationale, and Dissertation Aims	1
1.1 Introduction to the Human Microbiome	1
1.2 Introduction to Metagenomics	3
1.3 Introduction to Batch Effect.....	5
1.4 Dissertation Aims.....	7
1.4.1 Aim 1: Evaluate the performance and limitations of ComBat-Seq in removing batch effect from shotgun metagenomic sequencing data.....	7
1.4.2 Aim 2: Determine regional differences in oral microbial composition among adult Black women living in the US.....	8
1.4.3 Aim 3: Improve ComBat-Seq in removing batch effect from microbiome data using imputation	8

Chapter 2: Evaluation of ComBat-Seq in correcting batch effect in metagenomic data ..	10
2.1 Background	10
2.1.1 Batch Effect in Microbiome Data.....	10
2.1.2 Accounting for Batch Effect vs Correcting for Batch Effect	12
2.1.3 Current Batch Effect Correction Tools.....	12
2.2 Methods.....	15
2.2.1 Bioinformatic Tools for Batch Effect Correction.....	15
2.2.2 Simulated Dataset from MOMS-PI	16
2.2.3 Black Women’s Health Study Dataset	18
2.2.4 Evaluating Performance in Batch Effect Removal and Differential Abundance Analysis for Microbial Association with Smoking Status using ComBat-Seq- Corrected Data.....	20
2.3 Results	21
2.3.1 Comparing Batch Effect Tools using a Simulated 16S Sequencing Dataset ...	21
2.3.2 Evaluating ComBat-Seq using a Real Metagenomic Dataset from a Large-Scale Epidemiology Study	25
2.4 Discussion	27
Chapter 3: Regional variation analysis of the oral microbiota among adult Black women living in the US	30

3.1 Background	30
3.1.1 Oral Microbiome	30
3.1.2 Geographical Location and Microbial Composition	32
3.2 Methods	32
3.2.1 Study Population and Oral Wash Samples	32
3.2.2 DNA Extraction and Metagenomic Sequencing	33
3.2.3 Mapping Metagenomic Data	33
3.2.4 Filtering Data	34
3.2.5 Geographical Clusters	35
3.2.6 Correcting for Batch Effect	36
3.2.7 Differential Abundance Analysis	38
3.3 Results	39
3.4 Discussion	42
Chapter 4: Incorporation of Imputation to Improve ComBat-Seq in Removing Batch Effect from Rare Microbes	47
4.1 Background	47
4.1.1 Rare Microbes and their Roles in Microbial Ecosystems	47
4.1.2 Issue of Batch Effect Correction Methods in Managing Batch Effect among Rare Microbes	48

4.1.3 Imputation.....	49
4.2 Methods.....	51
4.2.1 Imputing Zeroes in Microbiome Data using zCompositions.....	51
4.2.2 Assessing Outliers and Degree of Adjustment after Batch Effect Correction .	52
4.2.3 16S rRNA Simulations	53
4.2.4 Real data application using shotgun metagenomic sequencing BWHS data ...	55
4.3 Results	56
4.3.1 iComBat-Seq retains more rare taxa compared to ComBat-Seq	56
4.3.2 Comparing iComBat-Seq with Existing Batch Effect Correction Methods using Simulated 16S Data	59
4.3.3 Applying iComBat-Seq in a Real-World Epidemiological BWHS Study	64
4.4 Discussion	66
Chapter 5: Conclusion and future directions	70
BIBLIOGRAPHY.....	73
VITA.....	92

LIST OF TABLES

Table 3.1	38
-----------------	----

LIST OF FIGURES

Figure 2.1 Evaluation on the simulated data (Condition FC > Batch FC).....	22
Figure 2.2 Evaluation on the simulated data (Condition FC < Batch FC).....	23
Figure 2.3 Comparison of batch effect correction tools (ConQuR without zero imputation).....	24
Figure 2.4 Removal of batch effect from BWHS dataset	26
Figure 3.1 Geographical clustering of samples across the US via K-Means Clustering ..	36
Figure 3.2 UMAP of metagenomic data before and after ComBat-Seq batch effect correction	37
Figure 3.3 Distribution of differentially abundant genera stratified by geographical cluster	40
Figure 3.4 Distribution of AHEI score stratified by geographical clusters	41
Figure 4.1 Comparison of outliers after batch effect correction using iComBat-Seq, ComBat-Seq, and MMUPHin.....	57
Figure 4.2 Read adjustments on logCPM scale using iComBat-Seq, ComBat-Seq, and MMUPHin	58
Figure 4.3 Proportion of variance explained by batch and condition effects	60
Figure 4.4 Performance of DA analysis using uncorrected and corrected data (Batch FC = 16, Condition FC = 64).....	62
Figure 4.5 Performance of DA analysis using uncorrected and corrected data (Batch FC = 64, Condition FC = 16).....	63

Figure 4.6 Read adjustments on BWHS dataset using iComBat-Seq, ComBat-Seq,
MMUPHin, and ConQuR 65

LIST OF ABBREVIATIONS

AHEI	American Healthy Eating Index
ALR	Additive Log Ratio
ANCOM-BC	Analysis of Compositions of Microbiomes with Bias Correction
BMI	Body Mass Index
BWHS	Black Women's Health Study
DA	Differential Abundance/Differentially Abundant
DNA	Deoxyribonucleic Acid
FC	Fold Change
FDR	False Discovery Rate
FN	False Negative
FP	False Positive
HMP	Human Microbiome Project
IBD	Inflammatory Bowel Disease
iHMP	integrative Human Microbiome Project
ILR	Isometric Log Ratio
logCPM	logarithm of Counts Per Million
MOMS-PI	Microbiome Study-Pregnancy Initiative
NB	Negative Binomial
NCBI	National Center of Biotechnology Information
NIH	National Institutes of Health
OTU	Operational Taxonomic Unit

PCR	Polymerase Chain Reaction
PERMANOVA	Permutational Multivariate Analysis of Variance
RNA	Ribonucleic Acid
rRNA	ribosomal Ribonucleic Acid
SES	Socioeconomic Score
TN	True Negative
TP	True Positive
UMAP	Uniform Manifold Approximation and Projection
US	United States
WGS	Whole Genome Sequencing

Chapter 1: Background, Rationale, and Dissertation Aims

1.1 Introduction to the Human Microbiome

The microbiome, a collection of microbial taxa or microbes and their genes in any given environment, has become a focus of research in recent years, particularly in the field of human health and precision medicine. Due to the human microbiome's role in human development (Mueller et al., 2015), immunity (Thaiss et al., 2016), and nutrition (Kau et al., 2011), researchers and doctors seek to utilize the microbial community to identify indicators of disease and improve personalized medicine. Starting from our birth, bacteria living in our bodies contribute significantly to our health by processing nutrients, driving away pathogens, and strengthening the immune system (Coelho et al., 2021; Yao et al., 2021). Hence, it is important to understand the connection between the microbiome and human health, starting with the association of changes in microbial structure to disease and various environmental factors. The Human Microbiome Project (HMP) was the first major research initiative from the National Institutes of Health (NIH) to deeply understand the human microbiome and its role in human health and disease.

On the human body, there are multiple sites where microbial colonization can occur: nose, mouth, lungs, stomach, colon, sexual organs, and skin. Depending on the location, the microbial community may contribute differently to its human host. The gut microbiome, for instance, is the home of many bacteria that mainly reside in the cecum of the large intestine (Zaborin et al., 2020). The microbes that grow in the gut are responsible for digesting breast milk and fiber (Bunyavanich et al., 2016; Clarke et al., 2014), helping control the host's immune system (Thaiss et al., 2016), and helping control brain health

(Martin et al., 2018). Indeed, many bacteria that reside in the gut serve to benefit the host, but other bacteria have been known to cause certain autoimmune diseases and gastrointestinal disorders, including irritable bowel syndrome (Pimentel & Lembo, 2020) and inflammatory bowel disease (Nishida et al., 2018). The ability to identify and differentiate between beneficial and harmful bacteria is paramount in potential therapeutic treatments that seek to influence the microbiota to treat disease and improve human health.

It is important to consider a range of factors that can influence the human microbial community. Due to the differences in environmental factors among these body sites, such as pH level (Miller et al., 2016), nutrition (Asnicar et al., 2021), and host-microbe interaction (Brown et al., 2013), microbial composition can vary significantly, which in turn influences its relationship with the human host. External factors can also greatly impact microbial composition. One of the more studied phenomena is the impact of antibiotics and its adverse effects on host health (Patangia et al., 2022). While their application in treating infectious diseases have been widely successful, they have come at the cost of disrupting the gut microbiome, which has led to a rise in cases of gastrointestinal disorders related to bacterial antibiotic resistance (Carlet, 2012; Huddleston, 2014). One of the most prominent examples is *C. difficile* infection. Commonly contracted among patients in hospital settings, it usually occurs when patients are prescribed antibiotics that wipe out their internal microbial communities and subsequently present an opportunity for *C. difficile* to flourish and dominate the gut microbiome (Chang et al., 2008). In addition to antibiotics, other factors can influence the microbial landscape, including diet, age, and environment (Yatsunencko et al., 2012). In our pursuits to better understand the relationship

between the microbiome and human health and disease and develop treatments through microbial interventions, it is important to identify and understand the numerous factors that can influence microbial composition.

1.2 Introduction to Metagenomics

In recent years, studies of the microbiome have become more tractable and advanced in large part thanks to advancements in metagenomics, the study of genetic material recovered directly from a collection of species within a sample (W.-L. Wang et al., 2015). There are several approaches to study the human microbiome: amplicon 16S sequencing and shotgun metagenomic sequencing.

16S sequencing is the gold standard for studying the microbiome and gained popularity due to its specific targeting of the unique ribosomal subunit within bacteria, making it particularly effective in targeting bacterial sequences (Clarridge, 2004; Fadrosch et al., 2014; Weisburg et al., 1991). The 16S rRNA gene consists of conserved and variable genomic regions. Because the conserved regions are present in most bacteria, it is possible to perform PCR amplification using universal primers that target the conserved regions (Baker et al., 2003; Lu et al., 2000; McCabe et al., 1999). After amplification, the hypervariable regions can then be used for species identification (Becker et al., 2004; Chakravorty et al., 2007). Sequence analysis of the 16S rRNA gene has been widely used to profile microbial communities and perform taxonomic studies, including identifying pathogens associated with human health and disease (Cox et al., 2013; Ley et al., 2005; Melito et al., 2001). However, 16S sequencing does have its disadvantages. First, because it does not capture any other genes besides the 16S subregion, it cannot detect activity and

hence cannot measure any activity both within the microbial community as well as between the community and its environment. Second, by targeting only the 16S variable regions, it is limited to genus and species in its taxonomy resolution. Nevertheless, thanks to its precise targeting of bacterial DNA and affordability, 16S rRNA sequencing remains a popular method for microbiome profiling studies.

Shotgun metagenomics, a second common method to study the microbiome, utilizes a “wide net” approach in which it captures any and all genetic material within a sample, both originating from bacteria as well as other sources such as host, fungi, and viruses (Quince et al., 2017). Like regular whole genomic sequencing, the library preparation workflow for shotgun sequencing involves random fragmentation and adapter ligation. Fragments of DNA are simultaneously and independently sequenced. Contigs of long DNA stretches are assembled from shorter, overlapping sequences. Once assembled, the DNA sequences are aligned to a reference database for taxonomic classification. Shotgun sequencing has several advantages when compared to 16S sequencing. First, because it captures all DNA within a sample, functional profiling becomes possible. Although it cannot measure actual functional activity like metatranscriptomics, we can determine functional potential from identifying and profiling all microbial genes present in a sample (Silva et al., 2016). Second, shotgun sequencing’s resolution can reach the strain-level by reading the entirety of a microbe’s genome, providing more accurate alignment to strain-level reference genomes (Buytaers et al., 2021). However, shotgun sequencing also has its disadvantages. For one, it has a higher risk of false positives. While 16S sequences are recovered with no error in the sequence, shotgun sequencing relies heavily on a

reference database when identifying microbes, which introduces several potential issues: 1) the database may not contain perfect genomes of all possible microbes, introducing considerable risk of false positives as some genes may be incorrectly mapped to other genomes (Campana et al., 2014); 2) unlike 16S sequencing, it is important to filter out unwanted reads and isolate only bacterial sequences if we want to study just the microbial community (Couto et al., 2018). Despite the drawbacks, however, the advantages of shotgun sequencing are apparent since, in return, we are rewarded with an insight into the activities that occur within the community.

1.3 Introduction to Batch Effect

While the microbial community's high sensitivity to its environment is a focus of interest in microbiome research, it also presents a major obstacle. A variety of external factors can profoundly affect the composition of the human microbiota, and technical and computational variations introduced during sample collection and processing are no exception (Y. Wang & Lêcao, 2020). Batch effects, defined as any unwanted, systematic source of variation that is unrelated to but obscures the biological factor of interest, can occur whenever samples are processed in different batches or at separate times. They can occur in various types of experiments, including genomics, proteomics, and transcriptomics, and can be caused by a variety of factors, such as variations in instrument settings, reagent lots, or technician experience (Leek et al., 2010). Batch effects can confound the results by introducing noise into the data and prevent the identification of true effects of biological factors of interest. Batch effect in microbiome data have led to erroneous claims in profiling studies. For instance, one study reported limited colonization

of *Micrococcus luteus* in the human intestine at mid-gestation, but that microbe was later found to be associated with batch effect instead (de Goffau et al., 2021; Rackaityte et al., 2020). It is therefore important to manage batch effects before downstream etiological analysis to ensure accurate and reliable results.

While batch effect has been known to affect many types of omics data, microbiome data presents a challenge due to a high abundance of zeros, over-dispersion of read counts, uneven library sizes, compositional nature, and inter-variable dependency (Y. Wang & Lêcao, 2020). An excess of zero counts may originate from either under-sampling or absence of the microbial community in the sample. In addition, data are often over-dispersed, with counts ranging from 0 to 10,000 per microbial variable. Uneven library sizes, which refer to differences in sequencing efficiency across samples, can further hinder sample comparisons. Although development of batch effect correction methods specifically for microbiome data is still in its early stages, existing methods made for other types of sequencing data have been used to remove batch effect from microbiome data. For instance, one study looking at the correlation of cigarette smoking and the oral microbiome used the `removeBatchEffect` function from the LIMMA package, originally designed for RNA-sequencing and microarray studies, to remove batch effects from oral microbiome 16S rRNA sequencing data (Wu et al., 2016). However, previously established methods do not account for the high sparsity, overdispersion, compositional nature, and inter-variable dependency of microbiome data and thus may be inappropriate to apply on microbiome data. Adjustments have been made to existing methods, such as adopting a negative binomial distribution model. For instance, ComBat-Seq, a batch effect adjustment tool for

RNA-seq count data, uses a negative binomial model to estimate the parameters of a systematic batch effect before adjusting the read counts to remove the batch effect (Y. Zhang et al., 2020). However, more work is needed to effectively remove batch effects from microbiome data while appropriately accounting for its specific characteristics.

1.4 Dissertation Aims

The aims of this dissertation seek to evaluate and optimize ComBat-Seq for removing batch effect from shotgun metagenomic sequencing data and to apply the method in an epidemiological study containing a metagenomic dataset of microbiota obtained from the human oral cavity. Together, these aims will illustrate the effectiveness of batch effect correction tools in increasing statistical power of microbiome studies.

1.4.1 Aim 1: Evaluate the performance and limitations of ComBat-Seq in removing batch effect from shotgun metagenomic sequencing data

In this aim, I evaluate the performance of ComBat-Seq, a batch effect adjustment tool for bulk RNA-seq count data, in removing batch effect from 16S rRNA and shotgun metagenomic sequencing data and compare it to other batch effect correction methods used in microbiome studies. Using both simulated 16S rRNA and real-world shotgun metagenomic sequencing datasets, I demonstrate that ComBat-Seq performs similarly to existing methods in removing batch effect from microbiome data while preserving the condition effect for downstream etiological analysis. Its application in microbiome profiling studies can improve the statistical power for biological discoveries connecting microbial communities and human health and disease.

1.4.2 Aim 2: Determine regional differences in oral microbial composition among adult

Black women living in the US

The human oral microbiome plays an essential role in oral health as resident bacteria have both pro- and anti-inflammatory activities crucial for supporting the host's immune system and maintaining homeostasis (Deo & Deshmukh, 2019/Jan-Apr; Kilian et al., 2016). Regions of the US differ significantly in environment, lifestyle, and diet, all of which affect oral microbiome composition. As dysbiosis has been previously associated with both oral disease and other health outcomes, it is important to consider how regional differences impact the oral microbiome composition. In this aim, I use whole metagenomic sequencing data extracted from oral wash samples to investigate the association between the oral microbiome and geographical region of residence among 640 women in the Black Women's Health Study. After correcting for batch effect using ComBat-Seq, I identify several oral health-related genera that are associated with the host's geographical location, highlighting the importance of considering social and environmental factors associated with the host's physical location when studying the oral microbiome.

1.4.3 Aim 3: Improve ComBat-Seq in removing batch effect from microbiome data using

imputation

While existing batch effect methods have shown fair success in removing batch effect from microbiome data, there are certain limitations due to the sparsity and high variability of microbiome data. ComBat-Seq, for instance, runs into issues when dealing with rare microbes with few non-zero read counts, of which one or two are outliers. In this aim, I address this issue by introducing iComBat-Seq, an extension of ComBat-Seq that

incorporates imputation to the batch effect correction technique, and evaluate the effectiveness of this change in removing batch effect while retaining rare microbes for downstream analysis. I demonstrate that iComBat-Seq improves upon the original method in removing batch effect from microbiome data, particularly among rare taxa, and outperforms existing methods in preserving relevant biological signals in both simulated 16S and real-world shotgun metagenomic sequencing data.

Chapter 2: Evaluation of ComBat-Seq in correcting batch effect in metagenomic data

2.1 Background

2.1.1 Batch Effect in Microbiome Data

Studies of the human microbiome have advanced in recent years thanks to the development of 16S rRNA and full metagenome sequencing technologies, which have enabled large-scale profiling studies involving hundreds to thousands of individuals (Graessler et al., 2013; J. S. Johnson et al., 2019). Large sample sizes facilitate more robust and powerful analyses of the role of microorganisms in health and disease. However, these large-scale studies are often susceptible to serious batch effect, caused by any unwanted biological, technical, or computational factors (Y. Wang & Lêcao, 2020).

One major obstacle in microbiome research is the high sensitivity of microbial compositions to their environment. While various biological factors, such as geography, age, sex, health status, stress, and diet (Foster et al., 2017; Kim et al., 2020; Singh et al., 2017; Yatsunenکو et al., 2012), are of interest when studying the microbiome, they may be obscured by technical elements introduced during the sample collection and processing stage. When samples are processed in multiple batches, any difference in laboratory conditions, reagent lots, and personnel between batches may affect measurements. Collectively, these unwanted sources of variation are known as batch effect and may produce spurious heterogeneity into the data, reducing replicability and muddling results (Leek et al., 2010). Therefore, it is crucial to manage batch effect during or before downstream etiological analysis to ensure accurate identification of effects of factors of

interest.

While procedures for managing batch effect in other types of omics data have been developed, microbiome data presents a particular challenge to batch effect correction due to its high abundance of zeros, over-dispersion of read counts, uneven library sizes, compositional structure, and inter-variable dependency (Gloor et al., 2017; Kaul et al., 2017; Y. Wang & Lêcao, 2020; Weiss et al., 2017). The excess of zero counts in microbiome data primarily stems from two sources: under-sampling and absence of certain microbes in the sampled microbial community. Under-sampling occurs when the limited size of the sample inhibits detection of taxa, resulting in technical or sampling zeros. This problem is particularly well-known in microbiome data because collected samples often represent only a small fraction of the true population microbial community. Rare taxa are especially susceptible to under-sampling due to their small presence in the true community. The other source of zeroes, the absence of taxa in the sample, produces an excess of biological or structural zeroes because certain taxa, particularly rare ones, are often only present in a few samples, resulting in zeroes among the remaining samples. Another distinct feature of microbiome data is its overdispersion: counts can range greatly for any given taxon. The publicly available 16S sequencing data for inflammatory bowel disease (IBD) from the integrative Human Microbiome Project (iHMP) contains taxa with non-zero read counts ranging from 1 to at least 10,000 reads (Lloyd-Price et al., 2019) and variances that go to the millions. Together, the high sparsity and overdispersion complicate the modeling step and assumptions performed in conventional batch effect adjustment methods.

2.1.2 Accounting for Batch Effect vs Correcting for Batch Effect

There are two main approaches when it comes to addressing batch effect. One approach is accounting for batch effect: during statistical analysis, batch effect can be included as a covariate in models (W. Chen et al., 2020). While this method avoids the issue of directly manipulating the data, its major disadvantage is decreased statistical power. The addition of a covariate in the model introduces additional degrees of freedom (and thus more variance), increasing the difficulty of rejecting the null hypothesis. As a result, it becomes harder to produce results with statistical certainty. The second approach is to correct for batch effect: before downstream analysis, batch effect can be removed entirely from the data (Tran et al., 2020). The main advantage here is that the methods are practical and enable broader application in a variety of analyses. With the removal of a covariate from the data, there will be more statistical power in the downstream analysis. However, batch effect correction has its own complications. For one, it assumes that all variance is either contained in the known variables or batch effect. In other words, any unknown variable is simply assumed to be due to batch effect. This can become a problem as it effectively removes our ability to contribute variance to any unknown source.

2.1.3 Current Batch Effect Correction Tools

Batch effects are not unique to microbiome data but rather widespread in all types of high-throughput experiments. While there is limited work on batch effect correction methods tailored specifically for microbiome data, existing standard tools developed for other genomic technologies have been applied to microbiome data with varying success. For instance, one study studying the correlation of cigarette smoking and the oral

microbiome used the `removeBatchEffect` function from the LIMMA package, originally designed for RNA-sequencing and microarray studies, to remove batch effects from oral microbiome 16S rRNA sequencing data (Wu et al., 2016). However, previously established tools do not account for the special characteristics of microbiome data and thus may be inappropriate to apply on microbiome data. For instance, ComBat, a popular batch effect adjustment method, assumes Gaussian distributions for the underlying distribution of the data (W. E. Johnson et al., 2007). However, microbiome data tend to more closely follow a zero-inflated negative binomial distribution, making it unsuitable for ComBat and related methods (Jiang et al., 2021). Thus, there is a need to adopt new methods that specifically account for zero inflation, over-dispersion, and complex distributions.

Fortunately, it is still possible to apply existing batch effect correction methods to a microbial abundance dataset, provided that the methods adhere to its characteristics. ComBat-Seq, for instance, while originally designed for RNA-sequencing count data, uses a negative binomial regression to model and remove batch effect, making it potentially applicable to microbiome data (Y. Zhang et al., 2020). Furthermore, by using quantile normalization to map the empirical distribution of count data to an expected “batch-free” distribution, it preserves the integer nature of count data after batch effect adjustment, allowing us to use common tools for differential abundance that require the input to be in counts, such as DESeq2, ANCOM-BC, and LEFse.

There has been recent work in batch effect correction methods tailored specifically to microbiome data. Developed in 2019, MMUPHin extends the method from ComBat to include a component accounting for the zero-inflated nature of microbial abundance data

(S. Ma et al., 2022). Specifically, it adds a binary zero-count indicator to allow for zero inflation in the model for sample read count with respect to batch variable and biologically relevant covariates. Just like in ComBat, MMUPHin then estimates hyperparameters with empirical Bayes estimators before subtracting the batch-associated variables to produce batch-corrected count data. MMUPHin has been used to enable microbial community meta-analysis for identifying novel taxa associated with IBD. Another microbiome-specific batch effect correction method is ConQuR, which uses a conditional quantile regression to model non-zero count data with respect to batch and additional covariates (Ling et al., 2022). By relying on a non-parametric approach, it can theoretically better handle irregular distributions in microbiome data compared to other methods that rely on parametric assumptions. Another key feature of ConQuR is the use of a logistic regression model to determine the likelihood of presence-absence status of microbes. During the matching step, it can then convert some zero counts to non-zero read counts and vice versa depending on the sparsity of the estimated batch-free distribution compared to the original distribution. However, there is debate on whether batch effect correction methods should change features' presence/absence across batches since zero counts can either be structural or sampling zeros (S. Ma et al., 2022). Nevertheless, ConQuR has shown promise in removing batch effect while preserving signals of interest, as demonstrated in their paper when applied to an epidemiology study on stool microbial associations with systolic blood pressure.

In this chapter, I evaluate the performance of ComBat-Seq compared to current microbiome-specific batch effect correction methods in removing batch effect from both

simulated 16S and real-world shotgun metagenomic sequencing data representing microbial communities while preserving biological effects of interest for downstream analysis.

2.2 Methods

2.2.1 Bioinformatic Tools for Batch Effect Correction

We evaluated three bioinformatic tools for batch effect correction: ComBat-Seq, MMUPHin, and ConQuR. All tools were run on R version 4.2.2.

For all corrections using ComBat-Seq, we used ComBat-Seq from the R package *sva* version 3.46.0, installed from Bioconductor. The taxon-by-sample counts table was read as a matrix, batch indicators as a factor, and covariates of interest as a data frame. By default, the condition of interest was set to be included in the model (`full_mod = TRUE`), and shrinkage on parameter estimation and dispersion was not performed (`shrink = FALSE`, `shrink.disp = FALSE`).

For all corrections using MMUPHin, we used MMUPHin version 1.12.0 from Bioconductor. The taxon-by-sample counts table was read as a matrix, and the metadata containing batch indicators and covariates was read as a data frame. Default control parameters were used to run MMUPHin: a zero-inflated model was used for correction (`zero_inflation = TRUE`), pseudo-count was set automatically to half of the minimal non-zero values observed in the counts table (`pseudo_count = NULL`), the convergence threshold for the method's iterative algorithm for shrinking batch effect parameters was set to $1e-4$ (`conv = 1e-4`), and the maximum number of iterations allowed was set to 1,000 (`maxit = 1000`).

For all corrections using ConQuR, we used ConQuR version 2.0 from GitHub. The taxa read count table was read as a sample-by-taxa data frame, batch indicator as a factor, and covariates as a data frame. The reference batch was set as “0” corresponding to the control batch without FC adjustment. By default, standard logistic and quantile regressions were used (`logistic_lasso = FALSE`, `quantile_type = “standard”`), and simple quantile-quantile matching was not used (`simple_match = FALSE`).

2.2.2 Simulated Dataset from MOMS-PI

To compare the effectiveness of batch effect correction among tools, we generated a simulated dataset with a set condition and batch effect. Starting with a publicly available 16S rRNA sequencing dataset from the Multi-Omic Microbiome Study-Pregnancy Initiative (MOMS-PI) (Fettweis et al., 2019), funded by the NIH Roadmap Human Microbiome Project (HMP) (Integrative HMP (iHMP) Research Network Consortium, 2019), we preprocessed the data to include only buccal mucosa samples that were obtained during the 4th visit and have a library size of at least 4,000 read counts. We then up sampled the counts table to the genus level and filtered out rare taxa that were present in less than 1% of samples. After preprocessing, the starting dataset contained 109 genera from 226 samples.

With the preprocessed data, we then proceeded to simulate batch effect and condition. We randomly assigned samples to a condition group (Condition 0 vs. 1) and batch (Batch 0 vs 1). Each assignment was performed using a Bernoulli distribution with an event probability of 0.5. To simulate association testing for individual taxa, we selected 20 taxa to be differentially abundant between Condition 1 and 0. Selected taxa ranged from

least to most abundant to evaluate the effectiveness of batch effect correction on association testing of taxa with varying abundances. Furthermore, the selected taxa were split evenly to be either more or less abundant in Condition 1 (relative to Condition 0). For samples belonging to Condition 1, we multiplied the initial counts of positively associated taxa by a pre-specified condition fold change (FC) and divided the initial counts of negatively associated taxa by the condition FC. To simulate batch effect, for samples belonging to Batch 1, we multiplied the counts of half of the taxa by a pre-specified batch FC to mimic increased abundance in Batch 1 (relative to Batch 0) and divided the counts of the remaining half by the batch FC for decreased abundance.

Due to the random selection of condition assignment and inclusion of rare taxa, there are cases in which one or both condition groups do not have any non-zero counts for a given taxon. In these cases, the taxon is considered to have structural zeroes associated with condition, and, if one of the two condition groups consists only of zeroes, it is automatically considered to be differentially abundant. Consequentially, none of the batch effect correction tools impacted downstream analysis in identifying such taxa since ComBat-Seq and MMUPHIn do not alter zero reads while ConQuR hits an error when encountering such taxa. As such, to compare the performance of ComBat-Seq, MMUPHIn, and ConQuR more accurately, taxa with condition-associated structural zeroes are excluded, keeping those that are more likely to differ across tools.

After incorporating a simulated condition and batch effect into the data, we performed batch effect correction using ComBat-Seq, MMUPHIn, and ConQuR. For each tool, we performed differential abundance analysis using ANCOM-BC2 (version 2.0.2) on

the resulting batch-corrected counts table to identify differentially abundant taxa associated with the simulated condition (Lin & Peddada, 2020). ANCOM-BC2 was set to identify structural zeroes based on group (`struct_zero = TRUE`), which removes taxa with condition-associated structural zeroes from further analysis. We calculated sensitivity and false discovery rate (FDR) using the predicted and true DA taxa to quantify the effectiveness of each tool in removing batch effect while preserving condition. We also evaluated the tools on overall removal of batch effect while preserving condition effects. Using PERMANOVA (Anderson, 2017) (`adonis2` from the *vegan* R package version 2.6-4), we calculated the R^2 metric to quantify the proportion of variability in the microbiome data explained by batch and condition. The robust Aitchison method was used to calculate pairwise distances for PERMANOVA (Martino et al., 2019).

To evaluate the robustness of the tools, we performed tests using 2 scenarios: 1) Condition effect > batch effect (condition FC = 64, Batch FC = 16); 2) Condition effect < batch effect (condition FC = 16, Batch FC = 64). For each scenario, we performed 100 simulations.

2.2.3 Black Women's Health Study Dataset

To assess ComBat-Seq in correcting batch effect in a real data, we used a shotgun metagenomic dataset obtained from the Black Women's Health Study (BWHS), a large follow-up study of the health of 59,000 Black American women who were ages 21 to 69 years at study baseline in 1995. Oral wash samples were obtained in 2004 through 2007 from approximately 50% of participants. For the current study, we included BWHS participants from two prior nested case-control studies of the oral microbiome in relation

to incidence of 1) pancreatic cancer and 2) lung cancer. After removing samples with a library size of less than 100,000 reads, a total of 640 participants were available. Samples were separated into two batches, and library preparation for each batch was performed in a different laboratory. Consequentially, batch effect associated with sequencing laboratory was observed in the metagenomic data, making it suitable for batch effect correction using ComBat-Seq. The extracted data were then sequenced using pair-ended whole metagenome shotgun sequencing using a minimum of 300 ng of DNA via the Illumina HiSeq2000 platform with a read length of 100 bp (insert size 350 bp). The metagenomic reads were then mapped to a comprehensive reference database containing 5,493 bacterial genomes, 9,364 viral genomes, 308 fungal genomes, and mammalian (human and mouse) libraries obtained from the NCBI database. The resulting dataset consisted of 30,686 OTUs representing 2,629 species and 1,741 genera. To reduce the sparsity of the data, during preprocessing, I removed individual OTUs with less than 1,000 total read counts and all associated OTUs belonging to a genus with less than an 0.01% average relative abundance. After filtering, the data had a median of 1,867,780 reads per sample assigned to bacterial, viral, and fungal reference genomes. From these, I detected 3,612 OTUs representing 362 unique species and 103 unique genera. Additional details on the BWHS dataset regarding sample collection, DNA extraction, sequencing, metagenomic mapping, and filtering can be found in Chapter 3.

2.2.4 Evaluating Performance in Batch Effect Removal and Differential Abundance Analysis for Microbial Association with Smoking Status using ComBat-Seq-Corrected Data

Even after removing rare microbes during the filtering step, we still observed batch effect. Thus, we corrected for batch effect using ComBat-Seq with sequencing lab as the batch variable. We performed batch effect correction on the metagenomic data at the genus level, which was the selected taxonomic level for downstream analysis. We also included age, gingivitis, alcohol consumption, smoking status, and BMI category as biological covariates to preserve in the corrected data. The covariates are included in the linear model used in ComBat-Seq and kept in the batch-free distribution model, resulting in the retention of signals from these variables in the data after correction.

To evaluate the effectiveness of ComBat-Seq in handling batch effect in a real dataset, we (1) measured the presence of batch effect in the metagenomic data before and after batch effect removal and (2) compared the results of differential abundance analysis for identifying taxa associated with smoking status. For (1), we used PERMANOVA R^2 using the robust Aitchison method to quantify the proportion of variance explained by batch effect in both the uncorrected and corrected data. For (2), we used ANCOM-BC2 to identify differentially abundant taxa associated with smoking status. We adjusted for age, gingivitis, alcohol consumption, and BMI, all of which are known to impact oral health, as categorical covariates. For the uncorrected data, we also included batch as a covariate to demonstrate the method of accounting for batch effect.

2.3 Results

2.3.1 Comparing Batch Effect Tools using a Simulated 16S Sequencing Dataset

When correcting for batch effect in a simulated 16S dataset in which variation due to condition is greater than that due to batch (condition FC = 64, batch FC = 4), ComBat-Seq performed adequately in removing batch effect correction while retaining enough variation associated with condition to identify the correct differentially abundant taxa with an average sensitivity of 0.66, compared to the baseline sensitivity of 0.70 (Figure 2.1). However, it did not perform well in reducing false positives with an average FDR of 0.055, which is slightly higher than the usual FDR cutoff of 0.05. Among the three methods tested, ConQuR performed the best both in identifying the correct differentially abundant taxa (sensitivity = 0.84) and reducing the number of false positives (FDR = 0.029).

When comparing overall effectiveness in removing batch effect while preserving condition effect, we found differing results. Using PERMANOVA with the robust Aitchison method, we found that ComBat-Seq was able to significantly reduce the batch effect (Batch $R^2 = 0.022$) while preserving most of the condition effect (Condition $R^2 = 0.056$). MMUPHin performed slightly better compared to ComBat-Seq (Batch $R^2 = 0.013$, Condition $R^2 = 0.061$). In contrast, ConQuR performed the worst in removing batch effect, retaining the highest percentage of variance that can be explained by batch (Batch $R^2 = 0.078$), but it also enhanced the condition effect (Condition $R^2 = 0.10$).

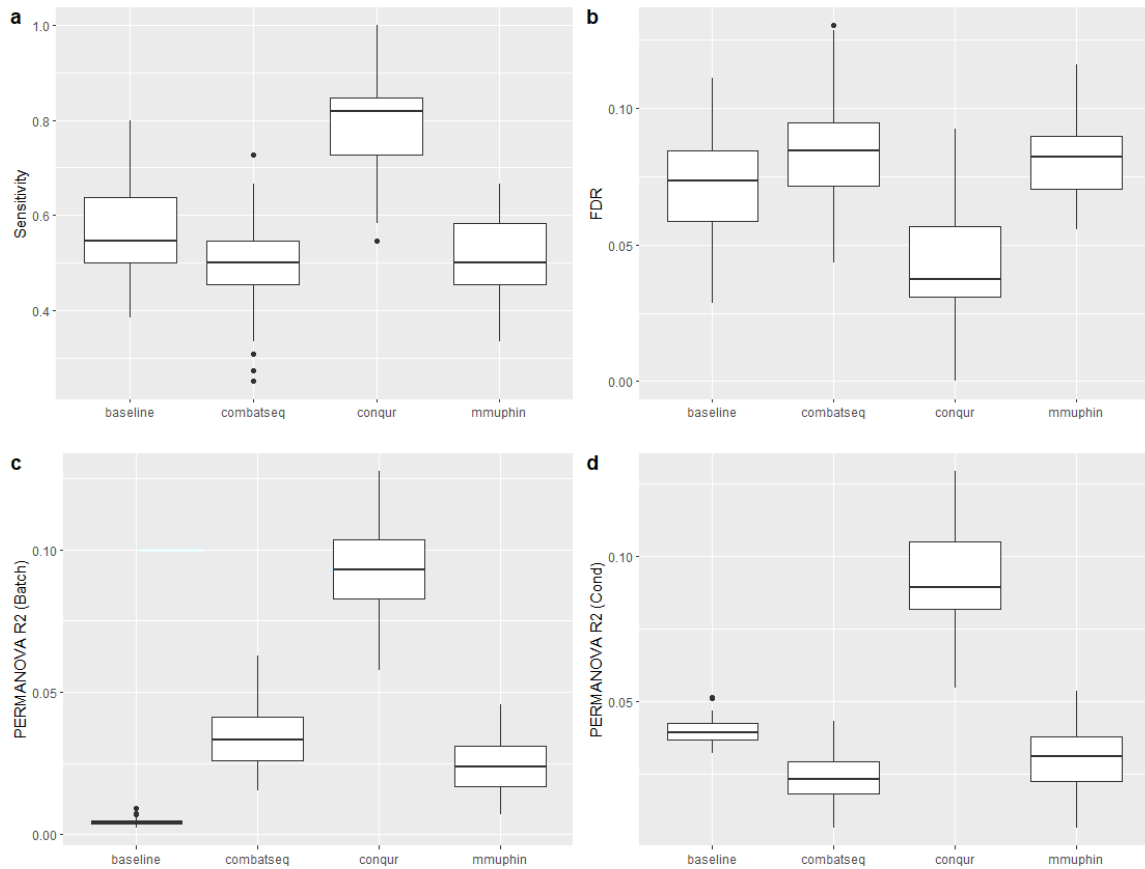


Figure 2.1 Evaluation on the simulated data (Condition FC > Batch FC) 100 simulations were run to compare the effectiveness of batch effect removal using ComBat-Seq, MMUPHin, and ConQuR. Sensitivity scores (a) and FDR (b) show that ComBat-Seq performed similarly to MMUPHin in identifying the correct differentially abundant taxa and avoiding false positives, while ConQuR performed the best. PERMANOVA R^2 statistics summarize the effect of batch (c) and condition (d) after correction.

A similar result was also seen in the scenario in which variation attributed to batch is greater than variation attributed to condition (condition FC = 16, batch FC = 64). Unsurprisingly, performance dropped for all methods with overall decreased sensitivity and increased FDR, except for ConQuR, which retained a similar performance as in Scenario 1 (Figure 2.2).

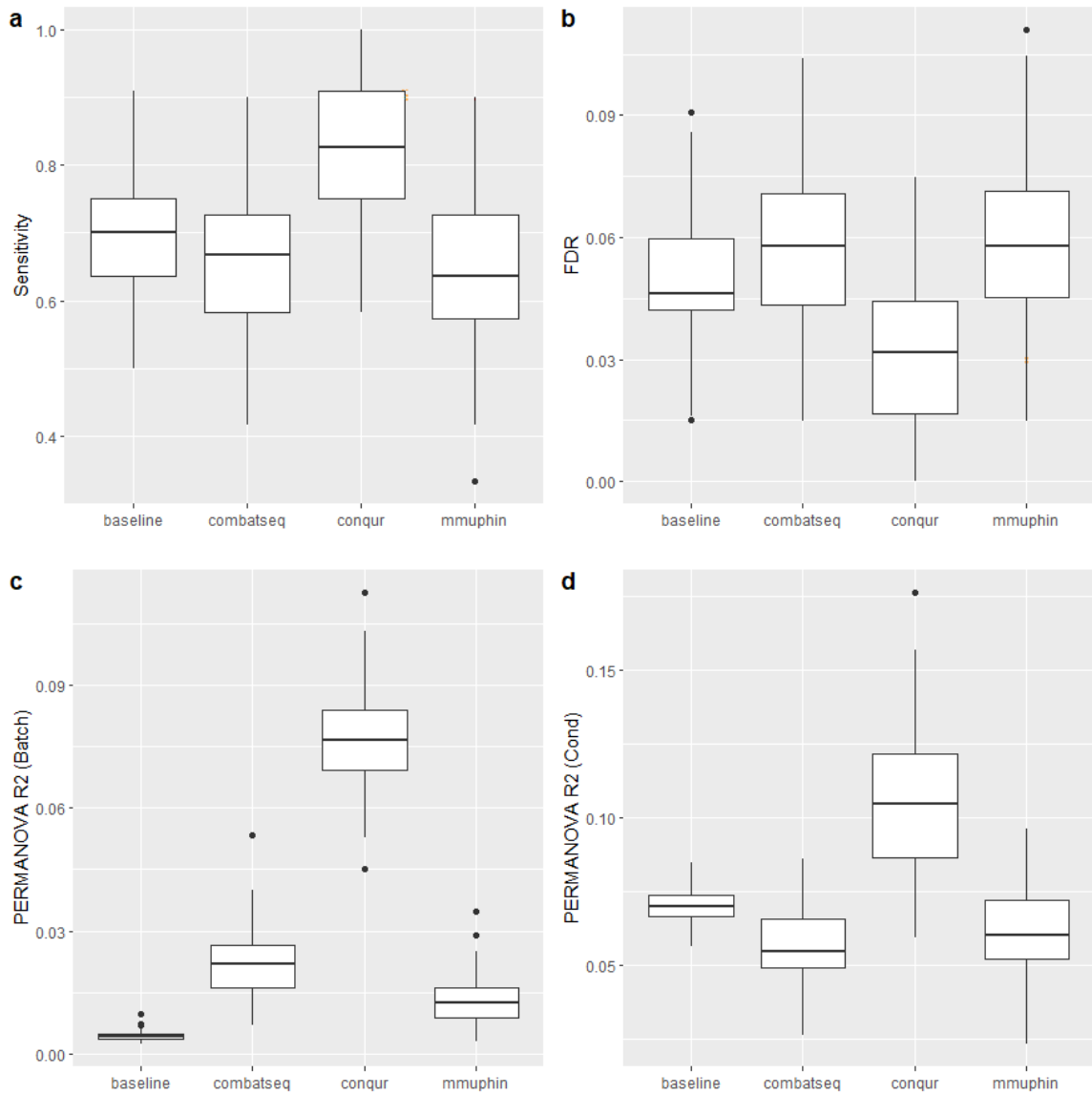


Figure 2.2 Evaluation on the simulated data (Condition FC < Batch FC) ComBat-Seq and MMUPHin performed worse overall in both sensitivity (a) and FDR (b) when the condition effect is less than the batch effect. ConQuR, on the other hand, retained a similar performance. In terms of overall effectiveness in removing batch effect (c) while preserving condition effect (d), ComBat-Seq and MMUPHin again performed worse. ConQuR, on the other hand, still performed the worst in removing batch effect, but was able to increase variance due to condition effect.

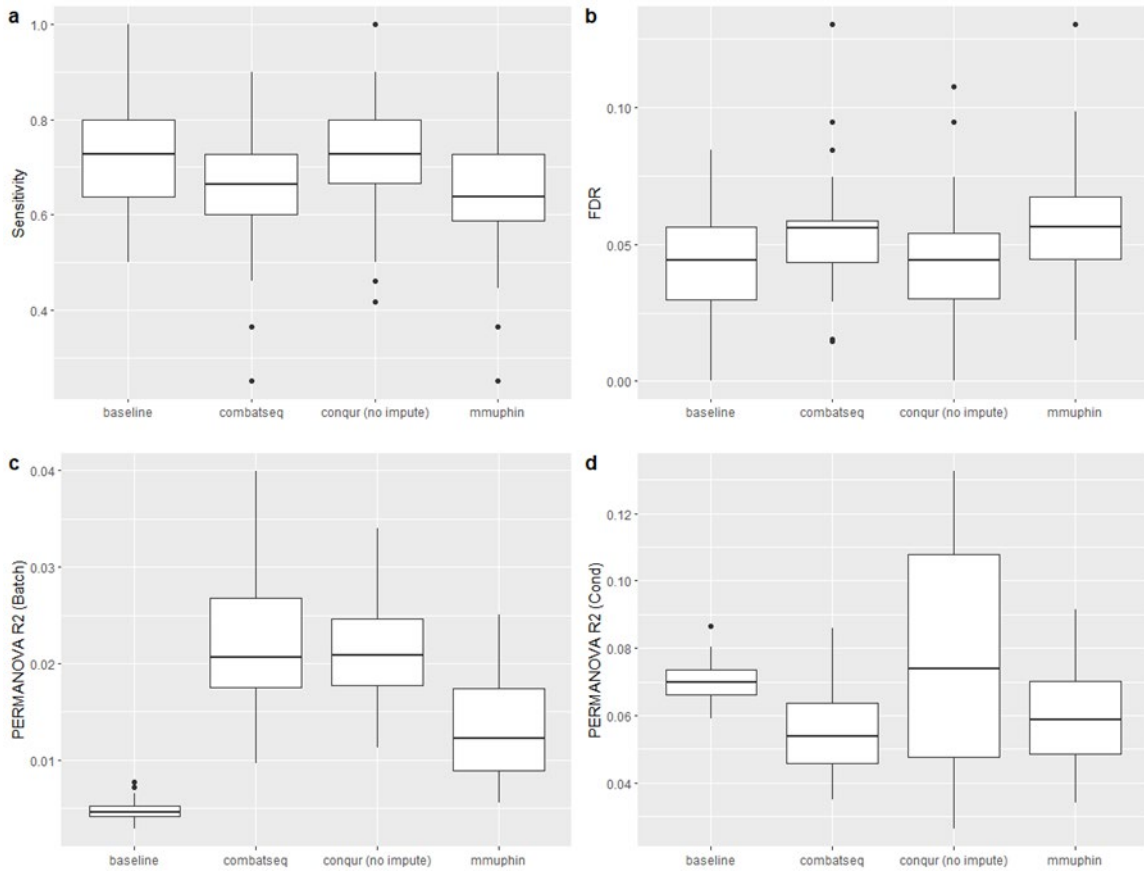


Figure 2.3 Comparison of batch effect correction tools (ConQuR without zero imputation) Scenario 1 (Condition FC > Batch FC) was rerun with original zeroes restored after correction using ConQuR. After restoring zeroes, ConQuR’s performance dropped both in detecting true differentially abundant taxa (a) and avoiding false positives (b). On the other hand, it was able to improve in removing overall batch effect (c), but also dropped in performance regarding preserving condition effect (d).

Interestingly, in both scenarios, ConQuR performed better than the baseline with no simulated batch effect, which suggests that ConQuR may not just be preserving the original variation associated with condition but in fact enhancing it. One distinctive feature of ConQuR is the inclusion of zero imputation, during which some zeroes are converted to non-zero read counts, which may improve downstream differential abundance analysis that relies on non-zero read counts. To test the impact of zero imputation on downstream

analysis, after batch effect correction using ConQuR, we restored the original zeroes in the count table before performing differential abundance analysis. The result was a drop in overall performance in both sensitivity (0.72) and FDR (0.044) for ConQuR (Figure 2.3). Nevertheless, ConQuR still performed the best in differential abundance analysis compared to ComBat-Seq and MMUPHin. For overall performance in batch effect removal, after restoring the original zero counts, ConQuR did improve in reducing batch effect (Batch $R^2 = 0.021$), but it also performed worse in preserving condition effect (Condition $R^2 = 0.078$).

2.3.2 Evaluating ComBat-Seq using a Real Metagenomic Dataset from a Large-Scale Epidemiology Study

We also tested ComBat-Seq on a real dataset from the BWHS study in which samples were separated into two batches during library preparation, causing batch effect. After correcting for batch effect using ComBat-Seq, we found that ComBat-Seq was able to significantly reduce batch effect (PERMANOVA R^2 for batch = 0.026 before correction, 0.0090 after correction) while preserving variance attributed to smoking status (PERMANOVA R^2 for smoking status = 0.0043 before correction, 0.0048 after correction). A dimension reduction using UMAP confirmed that batch effect was indeed reduced after correction (Figure 2.4). We also performed differential abundance analysis on the data to determine microbes that are significantly related to smoking status (current, past, and never) among our patients. To compare the differences in results between accounting for batch effect and correcting for batch effect, we did two independent analyses: an analysis on the unadjusted data where batch was included as a covariate in the modeling, and an analysis

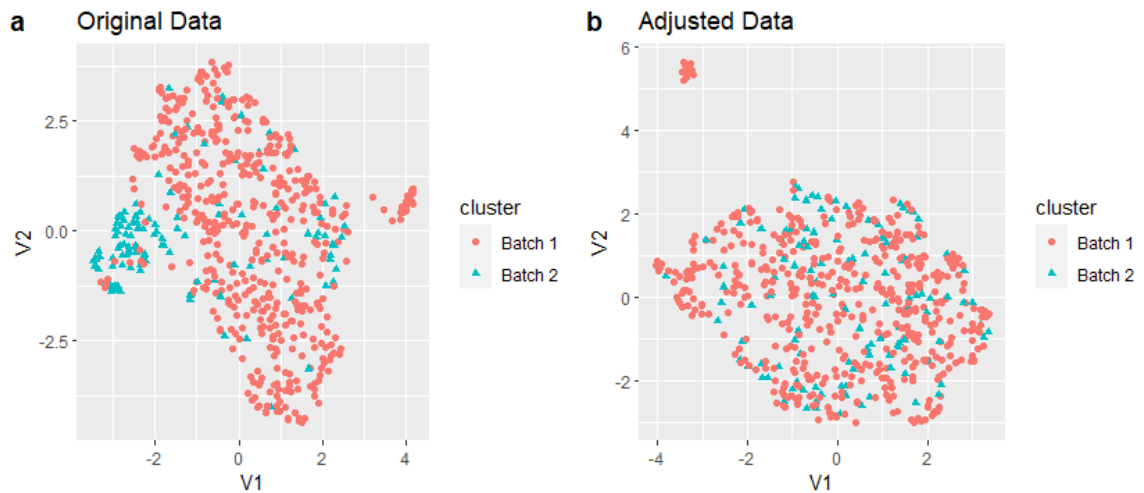


Figure 2.4 Removal of batch effect from BWHS dataset (a) Batch effect is observed in the original dataset, in which batch 1 clustered separately from batch 2. (b) after correction using ComBat-Seq, no clustering based on batch was observed.

on the adjusted data where batch was removed. Prior to correction, we identified 13 genera to be associated with current smokers compared to never smokers. After correction, we were able to detect 18 differentially abundant taxa. Taxa that were identified only after correction include *Peptoniphilus* ($P = 0.019$), *Nocardia* ($P = 0.0094$), *Verticillium* ($P = 0.023$), *Candida* ($P = 0.041$), *Debaryomyces* ($P = 0.018$), *Roseolovirus* ($P = 0.014$), *Xanthomonas* ($P = 0.024$), and *Mycolicibacterium* ($P = 0.035$). *Nocardia*, which is known to cause nocardiosis, was previously identified to be significantly associated with cigarette smoking among non-immunocompromised patients (Steinbrink et al., 2018). *Xanthomonas* belongs to the *Proteobacteria* phylum group, which is known to be depleted among current smokers (Wu et al., 2016). Overall, we found that the adjusted analysis identified more microbes compared to the unadjusted method, which suggests that batch effect correction is effective in increasing the statistical power to identify more microbes related to the biological factor of interest.

2.4 Discussion

Batch effect removal is a challenging task in microbiome data due to its zero-inflation, overdispersion, and compositional nature that distinguish it from other types of omics data. Therefore, it is important to select batch effect correction tools that account for the complex distribution of microbiome data. In this chapter, I evaluated the performance of ComBat-Seq in removing batch effect from both 16S rRNA and shotgun metagenomic sequencing data used in microbiome profiling studies.

Although ComBat-Seq was originally designed for bulk RNA-sequencing data, we have shown that its ability to remove batch effect from both 16S rRNA and shotgun metagenomic sequencing data is comparable to methods tailored specifically for microbiome data, such as MMUPHin and ConQuR. ComBat-Seq performed well in removing batch effect in the simulated data while preserving the condition effect, showing similar performance to MMUPHin and ConQuR (without imputation). It also demonstrated solid performance in correcting batch effect in a real shotgun metagenomic dataset, improving detection of individual taxa associated with relevant biological factors of interest. ComBat-Seq showed a higher sensitivity and lower FDR compared to MMUPHin. ConQuR did perform the best among the three methods, but its advantage was mainly due to its use of zero imputation to replace zeroes predicted to be due to under-sampling. However, zero imputation also resulted in unwanted noise associated with batch, resulting in increased variance due to batch after correction using ConQuR. Given the uncertainty of distinguishing sampling zeros from biological zeros, it may be more appropriate to focus on correcting non-zero abundance batch effects to avoid the complications of changing

features' presence/absence across batches.

ComBat-Seq's main advantage is its assumption that read counts follow a negative binomial distribution, which is commonly used to address over-dispersion and sparsity issues in count data (White & Bennetts, 1996). The use of a negative binomial distribution to model microbiome count data have been shown to be more powerful than conventional linear mixed models in detecting more significant taxa (X. Zhang et al., 2017). Zero-inflated Gaussian distributions (used in MMUPHin) can also be used, but it requires the count or proportion data be log-transformed first, which can entail the use of a pseudo count to avoid $\log(0)$. However, pseudo-counts can negatively impact downstream analysis by obscuring small signals associated with variables of interest.

There remain several limitations in using ComBat-Seq on microbiome data. First, like other batch removal methods, it struggles to correct for batch effect among rare taxa due to the lack of non-zero read counts needed to fit a stable model. The common solution is to remove rare taxa during the preprocessing step. However, while it improves the performance of ComBat-Seq in overall batch effect removal, it comes at the expense of lost data, particularly among rare microbes which may play significant roles in their microbial communities and host. Second, ComBat-Seq performs particularly badly when dealing with outliers, specifically within rare taxa. When outliers are present within a few non-zero read counts, ComBat-Seq tends to intensify the outliers, adjusting them to nonrealistic counts. Consequentially, this has a major impact on downstream analysis when counts are converted to relative abundances. For samples containing outliers in rare taxa, when the outlier is intensified, the relative abundances of such taxa increase significantly,

sometimes even becoming the most abundant within the sample, which the relative abundances of other taxa drop in response. Such significant changes in microbial composition can negatively affect downstream analysis, especially among studies with a small sample size.

Overall, ComBat-Seq is a viable method for removing batch effects from both 16S rRNA and shotgun metagenomic data, and its performance will improve with increased sample size and fewer outliers. Its application in microbiome profiling studies will improve the statistical power for biological discoveries connecting microbial communities and human health and disease.

Chapter 3: Regional variation analysis of the oral microbiota among adult Black women living in the US

3.1 Background

3.1.1 Oral Microbiome

First identified in the late 1600s by Antonie van Leeuwenhoek, known as the father of microbiology, and considered to be the second most diverse microbial community in the human body, the human oral microbiota consists of a wide range of symbiotic, commensal, and pathogenic microorganisms that colonize the oral cavity (Deo & Deshmukh, 2019/Jan-Apr). These microorganisms can be found on the hard surface of the teeth as well as the soft tissue of gums, tongue, and inside of the cheeks. With the highest alpha diversity among all the habitats on the human body, the oral microbiome is incredibly diverse, with hundreds of different species of bacteria alone. Using the Human Microbiome Project, one of the largest microbiome datasets to date, a study analyzed the microbial composition of nine intraoral sites (buccal mucosa, hard palate, keratinized gingiva, palatine tonsil, saliva, subgingival plaque, supragingival plaque, throat, and tongue dorsum) and found between 185 and 322 genera belonging to 13–19 phyla (Zhou et al., 2013). This plethora of bacteria can coat the surfaces of the oral cavity to form a structured, relatively stable biofilm (Marsh, 2006).

In contrast to early studies that presumed oral bacteria to be harmful, thanks to the advancements of next-generation gene-sequencing technologies, researchers have now determined the oral microbiota to play an essential role in oral health and other health outcomes. As it turns out, resident bacteria have both pro- and anti-inflammatory activities

crucial for supporting the host's immune system and maintaining homeostasis, both locally and throughout the body (Yu et al., 2019). For instance, *Streptococcus salivarius*, one of the first colonizers of the human oral cavity after birth, has been shown to significantly inhibit inflammation (Kaci et al., 2014). In addition, bacteria in the oral cavity help to regulate acidity, kill harmful pathogens, and even reduce blood pressure. On the other hand, dysbiosis of the oral microbiota have been linked to major oral diseases, including caries, periodontitis, and gingivitis (Davis et al., 2020; Yu et al., 2019), as well as several distal diseases, including obesity and pancreatic cancer (Gaiser et al., 2019; Yang et al., 2019).

As changes in oral microbial community compositions have been shown to be associated with health and disease, it is important to identify and examine factors that can influence the oral microbiome. The oral microbiota regularly interacts with the outside environment, and its composition is impacted by exposure to a wide variety of stimuli, including diet, oral hygiene practices, romantic partners, alcohol consumption, and tobacco use (Fan et al., 2018; Kato et al., 2017; Kort et al., 2014; Tribble et al., 2019; Wu et al., 2016). These external factors can significantly influence the overall function of the oral microbiome. Conventional oral hygiene, for example, was initially thought to be necessary for getting rid of as many oral bacteria as possible, including beneficial oral microbes. Many brands of mouthwash contain chlorhexidine that destroys much of the oral microbial community. However, a 2020 study found that the use of chlorhexidine mouthwash resulted in more acidity in the mouth and lower nitrite availability, leading to higher blood pressure (Bescos et al., 2020). For efforts in manipulating the oral microbiome to succeed, there is a need to better understand the various factors that contribute to it.

3.1.2 Geographical Location and Microbial Composition

One such factor that may influence the oral microbiota is the geographical location. A prior study examined the oral microbiome, using 16S rRNA gene sequencing, of native Alaskans, Germans, and Africans and reported that the alpha- and beta-diversity differed between the groups (Li et al., 2014). However, this study only considered geographical differences between countries and did not examine geographical heterogeneity within a country. Another study in China observed significant differences in the gut microbiome composition when comparing subjects from different provinces, even when they restricted to the Han ethnic group (Kwok et al., 2014). No such study has been conducted in the US, where there is heterogeneity in culture, diet, environment, and lifestyle across regions, all of which could potentially influence the healthy state of oral microbiome. In our study, we present a comparative analysis of oral microbiome diversity among participants of the Black Women's Health Study living in various regions in the US.

3.2 Methods

3.2.1 Study Population and Oral Wash Samples

The Black Women's Health Study (BWHS) is a large follow-up study of the health of 59,000 Black American women who were ages 21 to 69 years at study baseline in 1995. Oral wash samples were obtained in 2004 through 2007 from approximately 50% of participants (Adams-Campbell et al., 2016; Cozier et al., 2004). All samples were obtained via the mouthwash-swish method. Participants were asked to take a mouthful of Scope® mouthwash (at least ½ hour after eating or drinking), swish vigorously for 45 seconds, and spit the sample into a screw-top polypropylene jar. This method has been shown to produce

similar quality of DNA as the buccal swab method (García-Closas et al., 2001). The samples were then mailed via first class mail to the BWHS laboratory and processed on the day of receipt. Buccal cells from oral wash samples were centrifuged using TE buffer, aliquoted as a pellet into a 2ml vial, and stored at -80 degrees Celsius.

For the current study, we included BWHS participants from two prior nested case-control studies of the oral microbiome in relation to incidence of 1) pancreatic cancer and 2) lung cancer. All participants were cancer-free at the time of oral wash collection. A total of 640 participants were available.

3.2.2 DNA Extraction and Metagenomic Sequencing

Buccal cell samples were sent to research laboratories at either Harvard University or Vanderbilt University, where DNA was extracted using the PowerSoil Pro kit (MoBio Laboratories, Inc.), which has been shown to increase the ratio of bacterial and fungal to human DNA extracted (Goodrich et al., 2014). The extracted DNA were then sent to BGI Genomics, where pair-ended whole metagenome shotgun sequencing was performed using a minimum of 300 ng of DNA via the Illumina HiSeq2000 platform with a read length of 100 bp (insert size 350 bp).

3.2.3 Mapping Metagenomic Data

We used PathoScope 2.0 to map metagenomic reads to a comprehensive reference database of prokaryotic, eukaryotic, and viral genomes (Hong et al., 2014). The reference genome libraries (downloaded February 28, 2020) were generated from RefSeq's representative genomes and consisted of 5,493 bacterial genomes, 9,364 viral genomes, 308 fungal genomes, and mammalian (human and mouse) libraries. Reads that were

successfully mapped to bacterial, viral, and fungal genomes were included, while any reads that were mapped to human or mouse genomes were removed. From PathoScope's output of final best hit read numbers, each of which represents the number of reads that are mapped to a specific genome in the database, we created counts tables for each sample at the genus, species, and operational taxonomic unit (OTU) level (lowest taxonomic level possible).

3.2.4 Filtering Data

After alignment with PathoScope, the resulting data consisted of 30,686 OTUs representing 2,629 species and 1,741 genera. It was observed that 90.7% of observed read counts were zeroes, suggesting that many OTUs were observed in only a small subset of samples. Indeed, 18,553 OTUs (60.5%), 1,021 species (38.8%), and 614 genera (35.3%) were observed in less than 1% of samples (≤ 7 samples), of which 9,594 (31.3%), 384 (14.6%), and 210 (12.1%) were present in only 1 sample. The extremely small sample sizes for these taxa would significantly reduce the statistical power to detect significant differences across regions. Hence, to reduce the overabundance of zeroes present in the data and improve overall coverage of observed OTUs, two filtering criteria were applied: 1) remove individual OTUs with less than 1,000 total read counts and 2) remove all associated OTUs belonging to a genus with less than an 0.01% average relative abundance. The cut-off for average relative abundance was chosen based on a previous study that compared different filtering methods for removing contaminants (Cao et al., 2021). Following the above filtering criteria, we removed 26,536 OTUs that each had less than 1,000 total reads across all samples and an additional 538 OTUs that belonged to rare

genera with average relative abundances of less than 0.01%. After the filtering process, the samples had a median of 1,867,780 reads per sample assigned to bacterial, viral, and fungal reference genomes. From these, we detected 3,612 OTUs representing 362 unique species and 103 unique genera.

3.2.5 Geographical Clusters

K-means clustering was performed to categorize the participants into different geographical groups. We determined k-means clustering to be optimal for maximizing statistical power in downstream analyses due to the non-uniform distribution of participants across the US: many samples clustered around major cities on the East and West coasts and, more notably, we observed very few participants living in the Mountain region. We used longitude and latitude coordinates extrapolated from zip code data and the k-means function from the stats R package (maximum of 100,000 iterations and 10,000 random initial cluster centers). We determined 6 clusters to be optimal based on the average silhouette method (Rousseeuw, 1987). Using k-means clustering, we separated the participants into 6 geographical clusters: North-East coast (N = 231), South-East coast (N = 76), North Midwest (N = 174), South Midwest (N = 41), North-West coast (N = 50), and South-West coast (N = 68) (Figure 3.1). Most participants came from the Northeast coast and North Midwest regions.

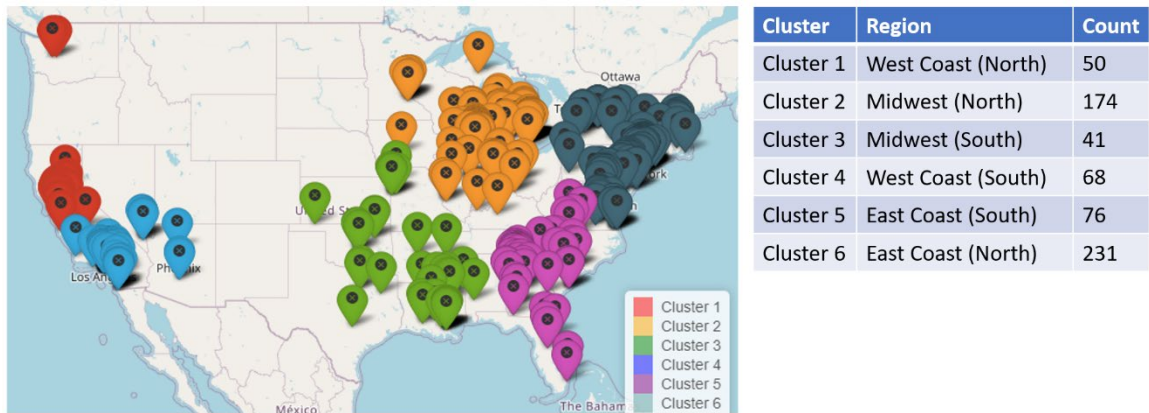


Figure 3.1 Geographical clustering of samples across the US via K-Means Clustering. Latitudinal and longitudinal data were extrapolated from zip code information, and K-means clustering was performing using latitude and longitude. 6 geographical clusters were identified: North-East coast (N = 231), South-East coast (N = 76), North Midwest (N = 174), South Midwest (N = 41), North-West coast (N = 50), and South-West coast (N = 68). No subjects living in the Mountain range (between the West coast and Midwest regions) were observed.

3.2.6 Correcting for Batch Effect

As the samples were processed at two different labs at three different timepoints, there are potential technical differences in the sample batches. Thus, we evaluated the metagenomic data for batch effects. In a UMAP plot of the data, we observed that samples whose library preparations were performed in one lab clustered separately from those with library preparations from the other lab. In addition, we observed that samples from one batch had significantly higher abundances of *Rothia* and *Actinomyces* compared to the other two batches. Even after removing rare microbes during the filtering step, we still noted batch effect, which suggests that batch effect affected even the more common microbes among our samples. Thus, we corrected the observed batch effect using ComBat-Seq with sequencing lab as the batch variable (Y. Zhang et al., 2020). We also included outcome, region, age, gingivitis, alcohol consumption, smoking status, BMI category, and

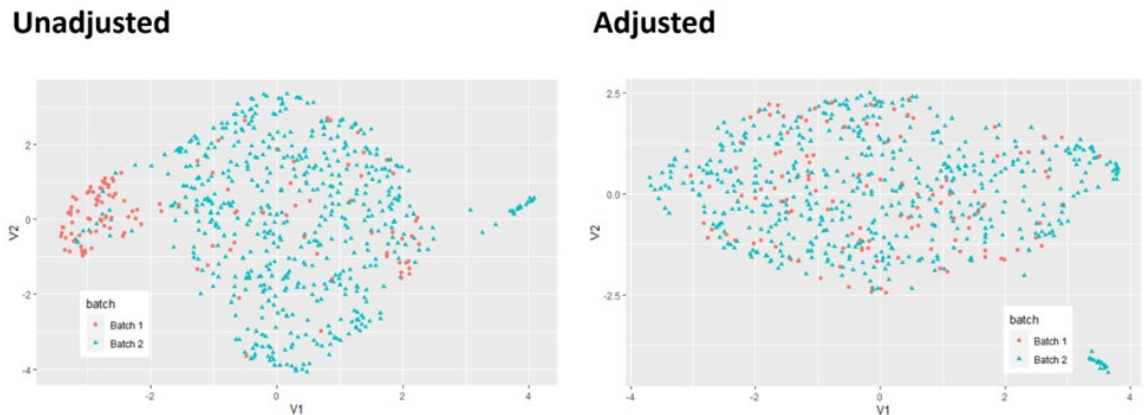


Figure 3.2 UMAP of metagenomic data before and after ComBat-Seq batch effect correction. Prior to adjustment, samples from Pancreatic Batch #1 (labeled Batch 1) clustered separately from samples from Pancreatic Batch #2 and Lung Batch (combined and labeled as Batch 2). After batch effect correction, we did not observe any separate clustering between Batch 1 and Batch 2.

neighborhood socioeconomic status (SES) quartile as biological variables to preserve in the corrected data. After correction, the adjusted data showed no clusters separating samples sequenced from Batch 1 and samples sequenced from Batch 2 (Figure 3.2). Differences in higher-abundance genera that were observed across batch groups in the unadjusted data were not observed in the adjusted data. A hierarchical clustering of the samples confirmed a significant difference in batch composition between two unsupervised clusters using a chi-square test ($P < 0.001$). After applying batch effect correction using ComBat-Seq and re-clustering the samples, we did not observe any significant separation for samples based on batch ($P = 0.90$), which suggests that we successfully removed the variation due to batch effect.

microbe	DESeq2		Limma		ANCOM-BC	
	P-Value	P-Value (Adjusted)	P-Value	P-Value (Adjusted)	P-Value	P-Value (Adjusted)
Bifidobacterium	< 0.0001	< 0.0001	0.0992	0.4592	0.0049	0.1681
Capnocytophaga	0.3750	0.4820	0.0039	0.0569	0.0222	0.3807
Corynebacterium	< 0.0001	< 0.0001	0.0032	0.0569	0.0000	0.0023
Eggerthia	0.0021	0.0091	0.3092	0.8167	0.0196	0.3807
Pauljensenia	< 0.0001	< 0.0001	0.0181	0.1730	0.7134	0.9350
Peptostreptococcus	0.0006	0.0036	0.2502	0.6995	0.0211	0.3807
Porphyromonas	0.0130	0.0371	0.0180	0.1730	0.6839	0.9350
Pseudopropionibacterium	0.7840	0.8540	0.0025	0.0569	0.0036	0.1681

Table 3.1 Differentially abundant genera identified using DESeq2, Limma, and ANCOM-BC. We adjusted for age, gingivitis, alcohol consumption, smoking status, BMI, and socioeconomic score (SES) as categorical covariates, all of which are known to impact oral health. Microbes were identified to be differentially abundant if it had a significant p-value for at least two of the three methods (adjusted p-value for DESeq2 and unadjusted p-value for Limma and ANCOM-BC).

3.2.7 Differential Abundance Analysis

To investigate differences between the geographic groups in oral microbiome composition, we performed differential abundance analysis using the R packages DESeq2, Limma, and ANCOM-BC (Lin & Peddada, 2020; Love et al., 2014; Ritchie et al., 2015). We adjusted for age, gingivitis, alcohol consumption, smoking status, BMI, and socioeconomic score (SES) as categorical covariates, all of which are known to impact oral health. Rather than relying on a single differential abundance analysis tool, which can produce substantially different results compared to other tools, we chose to use a consensus approach based on multiple methods to ensure robust results (Nearing et al., 2022). We observed that DESeq2 identified the most microbes to be differentially abundant even after multiple comparison correction while Limma and ANCOM-BC were particularly stringent in identifying microbes. Thus, to balance, we used an unadjusted p-value cutoff of 0.05 for Limma and ANCOM-BC and an adjusted p-value cutoff of 0.05 for DESeq2.

3.3 Results

Of the covariates included in our model, we found that age category and neighborhood SES quartiles were significantly different across geographical clusters. Notably, the South Midwest and South-West coast had higher proportions of subjects aged 70 and above, the South-East coast had higher proportion of subjects aged 60 to 64, and the North-West coast had higher proportions of subjects aged 65 to 69. For neighborhood SES, the Midwest clusters (both North and South) had more subjects who fell in the lowest quartile, the Southeast coast and Southwest coast had more people who fell in in the third quartile, and the Northwest Coast cluster had more subjects who fell in the highest quartile. We observed no differences between geographic location and BMI category, gingivitis, alcohol consumption, smoking status, number of packs smoked, or American Healthy Eating Index (AHEI) score.

We did not identify any significant difference in alpha diversity at the genus level among samples from different geographical clusters. However, when examining individual taxa, we identified 8 differentially abundant genera across the 6 geographical clusters: *Bifidobacterium*, *Capnocytophaga*, *Corynebacterium*, *Eggerthia*, *Paulgensenia*, *Peptostreptococcus*, *Porphyromanas*, and *Pseudopropionibacterium* (Table 3.1). Notably, we found *Bifidobacterium* to be more abundant among those living in the Northwest Coast compared to the other geographical clusters (Figure 3.3). *Corynebacterium* was also identified to be more abundant in the Northwest coast. In addition, several genera had consistent trends when comparing the North and the South parts of each region (East coast, Midwest, and West coast). *Capnocytophaga* was markedly more abundant in the North

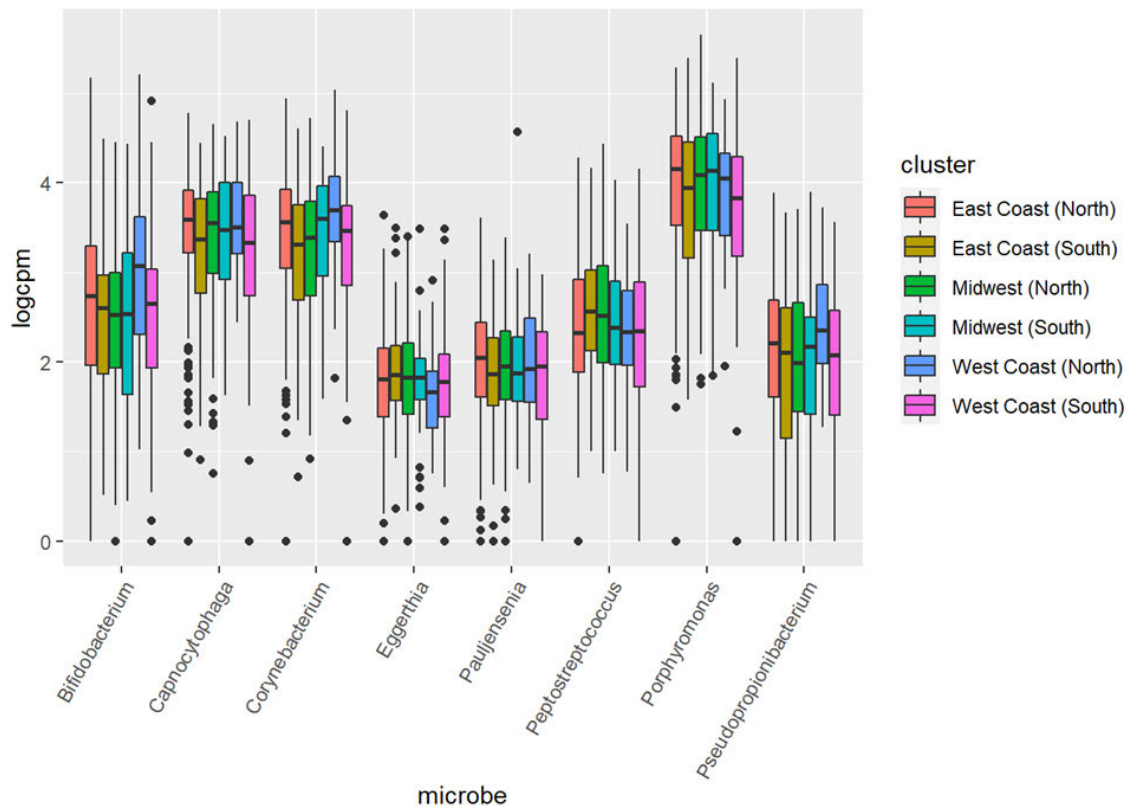


Figure 3.3 Distribution of differentially abundant genera stratified by geographical cluster: subjects living in the North-West coast showed overall higher abundances of *Bifidobacterium*, *Corynebacterium*, and *Pseudopropionibacterium* and lower abundances of *Eggerthia* compared to those living in other regions. Subjects living in the South-East and South-West coasts showed lower abundances of *Capnocytophaga* and *Porphyromonas* while those living in the North-East coast displayed higher abundances of *Pauljensenia*. Higher abundances of *Peptostreptococcus* were observed among those living in the South-East coast and North-Midwest. Age, gingivitis, alcohol consumption, smoking status, BMI, and socioeconomic score (SES) were controlled as categorical covariates.

regions compared to their South counterparts while *Eggerthia* was more abundant in the South regions. We did not observe a trend between *Capnocytophaga* and latitude when using a generalized linear model ($P = 0.37$). However, we did observe that among the subjects living in the West Coast, those living in higher latitudes were more likely to have higher abundances of *Capnocytophaga* compared to the those living in lower latitudes ($P = 0.016$).

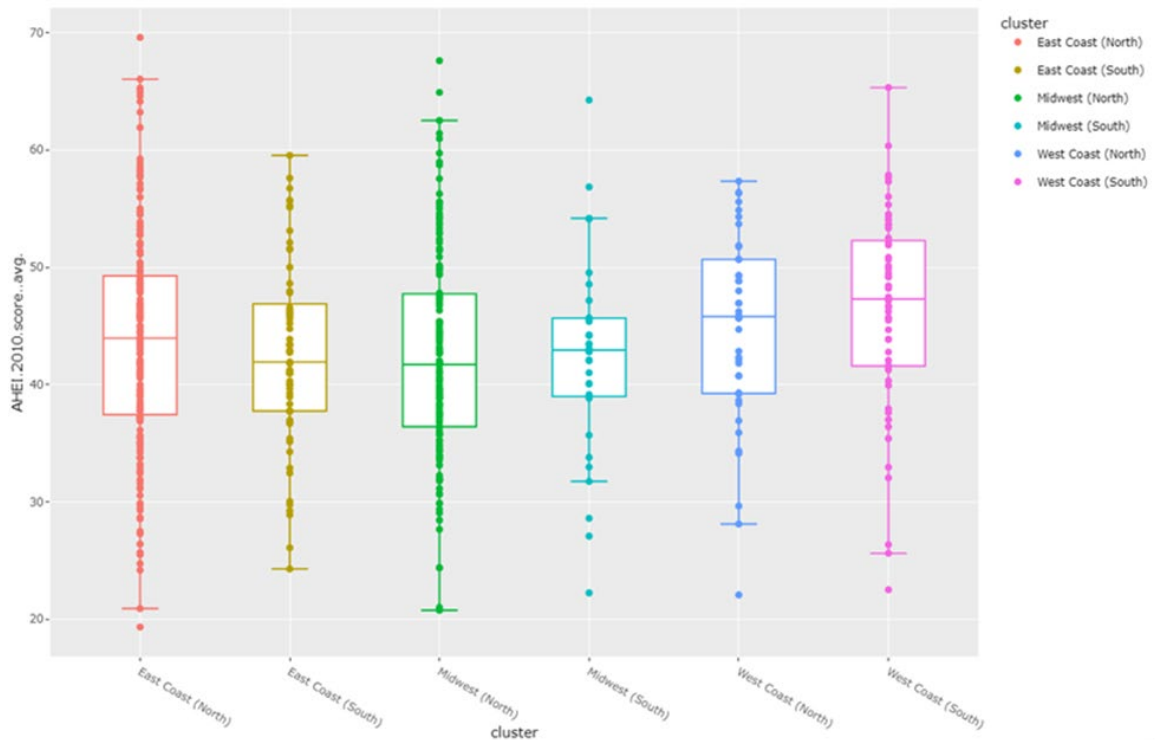


Figure 3.4 Distribution of AHEI score stratified by geographical clusters. there is a significant difference in distribution of AHEI score across the geographical clusters, with subjects living the North and South-West coast displaying overall higher AHEI scores compared to those living in the Midwest and East coast regions.

At the species level, we identified 29 differentially abundant species across geographical clusters. Notably, among them, we found 4 species from *Actinomyces* (*A. dentalis*, *A. israelii*, *A. massiliensis*, and *A. timonensis*), 3 species from *Bifidobacterium* (*B. breve*, *B. dentium*, and *B. mongoliense*), 2 species from *Capnocytophaga* (*C. canimorsus* and *C. granulosa*), 1 specie from *Corynebacterium* (*C. matruchotii*), 2 species from *Lactobacillus* (*L. aviarus* and *L. equicursoris*), and 2 species from *Streptococcus* (*S. gallolyticus*, *S. macacae*, and *S. urinalis*).

To determine whether the identified differences in genera abundances were influenced by diet, we performed a differential abundance analysis with the American

Healthy Eating Index (AHEI), a measure of diet quality used to assess how well a set of foods align with key recommendations of the Dietary Guidelines for Americans. AHEI distribution differed across our clusters, which we confirmed using a Kruskal-Wallis rank sum test ($P = 0.015$) (Figure 3.4). Interestingly, when AHEI score was included in our model for geographic region, *Bifidobacterium* was no longer differentially abundant across regions.

3.4 Discussion

In this study of a sample of U.S. Black women, we identified several key microbes that were differentially abundant across specific regions of the US. Among the identified microbes were key microbes that play important roles in overall and oral-specific health, highlighting the importance in considering environmental factors stemming from the host's physical location when studying the oral microbiome.

The present study is the first major study of the oral microbiome in relation to geographic region of the US among adult Black women. Previous studies have examined global comparisons of microbiome composition among individuals. One study analyzed the saliva microbiome from a total of 152 healthy individuals living in different geographic and climatic environments: 76 native Alaskans, 10 Germans, and 66 Africans (Li et al., 2014). Using next-generation sequencing of partial 16S rRNA gene sequences, they found significant differences in both beta and alpha diversity at both the OTU and genus levels among the three groups, with Africans having the greatest beta diversity and lowest alpha diversity and Germans having the lowest beta diversity and greatest alpha diversity. They also found that native Alaskans and Germans shared more similarities in the saliva

microbiome than between either group and Africans, suggesting an association between saliva microbiome composition and latitude. However, while previous studies investigated more global comparisons of human microbiomes, they had conflicting results when it came to more granular levels (within a country). For example, the previously mentioned study did not find differences in salivary microbiota composition among the four Alaskan groups based on geographical location, whereas significant differences were observed among the three African groups (Li et al., 2014).

A study of the human gut microbiota in China considered both ethnicity and geographic region within China among 314 individuals belonging to 7 ethnic groups and living in 9 provinces (Kwok et al., 2014). They found that ethnic origin contributed to shaping the human gut microbiota and they also observed significant differences between Han Chinese living in the northernmost province (Heilongjiang) and Han Chinese living in the more central provinces (Henan, Jiangsu, and Sichuan). Considering the ongoing question of whether geographical location plays a crucial role in shaping oral microbiome composition, we sought to investigate the impact of geographical location on oral microbiome composition among adult Black women across the US.

A differential abundance analysis revealed several genera that were differentially abundant among our geographical cluster groups. Interestingly, Black women living in the Northwest Coast displayed higher abundances of *Bifidobacterium*, whose species are known to be probiotics that have been used to alleviate diarrhea, constipation, and other intestinal disorders (O'Mahony et al., 2005). Due to their health-promoting properties, *Bifidobacteria* have been used as active ingredients in many functional foods. One

explanation for our results is that people living on the Northwest Coast consume more food containing *Bifidobacteria* (e.g., yogurt, fermented foods) compared to those living in other parts of the US. Previous studies have found that plant-based diets that are generally naturally low in fat tend to be associated with a higher abundance of *Bifidobacteria* in the human gut microbiota (Tomova et al., 2019). As most BWHS participants in the Northwest Coast region live around the San Francisco Bay area or in the Seattle area, an explanation for the higher abundance of *Bifidobacterium* is the preference of plant-based/vegan diets in that region compared to other parts of the US. Indeed, when we included the AHEI variable as a covariate in the model, *Bifidobacterium* was no longer strongly identified as differentially abundant across clusters.

Another interesting observation was that *Corynebacterium* had differing abundances across our six geographical clusters. One detected species is *C. matruchotii*, a Gram-positive actinobacteria that has been previously identified to be among the most prevalent species in the adult human oral core microbiome (Eriksson et al., 2017). Most notably, studies have identified *C. matruchotii* to be the central filament to which other bacteria such as *Streptococcus* and *Actinomyces* bind to form the tooth biofilm, which, in an unhealthy state, can serve as the etiological agent for major dental diseases (Esberg et al., 2020; Marsh, 2006). The varying levels of abundances of *Corynebacterium* across geographical clusters could suggest differences in tooth biofilm formation and susceptibility to dental disease.

Finally, several studies have found a correlation between microbiome composition and distance from the equator. To that end, we identified several genera that had differing

levels of abundance between the North and South pairs of each of the three major regions. *Eggerthia* appeared to be more abundant in the South, while *Capnocytophaga* appeared to be more abundant in the North. Bacteria belonging to the *Capnocytophaga* genus are often dependent on increased levels of carbon dioxide (Kapke et al., 1980); therefore, they may be more likely to survive and grow in the Northern Regions that experience greater fluctuations in carbon dioxide levels compared to the Southern Regions. In addition, *Capnocytophaga* has been linked to periodontal infections including periodontitis, septicemia, and endocarditis (Idate et al., 2020; Sakai et al., 2019).

We acknowledge several limitations in the present study. First, samples were collected at one point in time via a self-collected oral wash method, which could lead to potential contamination or information bias in specimen quality and accuracy. Second, we were not able to control for oral hygiene practices, which play an important role in maintaining oral health via removal of plaque. Third, our findings are limited to the regions where most subjects resided. Fourth, DNA extraction was performed in different labs, which resulted in batch effects. However, any technical bias is likely to be minimal as batch effect was removed using ComBat-Seq prior to differential abundance analysis.

Our study benefitted from a large sample size limited specifically to adult Black women. This removed noise or confounding by sex/gender and by racial group (Dwiyanto et al., 2021; Z. S. Ma & Li, 2019). In addition, detailed data previously collected from this cohort allowed for the adjustment of various confounders including age, BMI, smoking status, alcohol consumption, gingivitis, neighborhood SES quartile, and AHEI.

Our results emphasize the importance of considering social/environmental factors

associated with the host's physical location when studying the oral microbiome. As more studies seek to identify key microbial biomarkers for oral health, it is important to control for geographical region as a confounder. Future studies inclusive of samples from differing geographic regions should examine how the baseline oral health differs by location and adjust if needed.

Chapter 4: Incorporation of Imputation to Improve ComBat-Seq in Removing Batch Effect from Rare Microbes

4.1 Background

4.1.1 Rare Microbes and their Roles in Microbial Ecosystems

Microbiome communities usually exhibit a skewed microbial abundance distribution, composing of a few dominant microbes and numerous low-abundance microbes that occupy a small proportion of the community (Dunbar et al., 2002; Lozupone et al., 2012). Traditionally, microbial studies remove rare microbial taxa from their data due to the statistical challenge of detecting effects within features with a small sample size of non-zero read counts. However, because of this filtering approach, these studies systematically overlook a substantial part of the microbial community.

Despite their low abundances, there has been increasing evidence that rare microbes play active roles in both environmental ecosystems and host-associated microbial communities. While they may not contribute significantly under normal conditions, rare microbes can become important under changing conditions by providing necessary traits or acting as partners in new interspecific interactions (Shade et al., 2014). These microbes have been found to play important roles in nutrient cycling, carbon storage, and other key processes that are critical to the health of ecosystems. In plant tissue, rare foliar endophytes have been found to affect host size and foliar nitrogen content when *Alternaria fulva*, a dominant fungal endophyte, is absent (Harrison et al., 2021). They have also been linked to the discovery of new antibiotics, enzymes, and other bioactive compounds that have the potential to benefit human health and biotechnology. Rare microbes also play significant

roles in human health and disease. Several studies have linked rare microbes to major diseases and human health. For instance, one study identified the expansion of *Actinobacteria* with a decrease in abundant taxa among patients with rheumatoid arthritis (Ma et al., 2022).

Overall, the study of rare microbes holds great promise for advancing our understanding of the natural world and improving our ability to address key environmental and health challenges. Hence, it is vital to develop improved methods or strategies that overcome challenges hindering the study of rare microbes.

4.1.2 Issue of Batch Effect Correction Methods in Managing Batch Effect among Rare Microbes

Rare microbes are incredibly valuable in terms of their potential significance; however, their sparsity poses significant challenges when it comes to accurately modeling and performing statistical analysis. One specific challenge is the impact of batch effects on rare microbes, and unfortunately, existing tools are unable to adequately address this issue. Current batch effect correction methods are unable to generate stable estimates with few non-zero read counts, which can lead to impractical and unreliable adjustments. When we attempted to use ComBat-Seq to correct for batch effect among very rare taxa, we found that, in a few cases, outliers can become intensified, resulting in unrealistic counts (Y. Zhang et al., 2020). Consequentially, this has a major impact on downstream analysis when counts are converted to relative abundances. For samples containing outliers in rare taxa, when the outlier is intensified, the relative abundances of such taxa may increase significantly, sometimes even becoming the most abundant within the sample, with the

relative abundances of other taxa dropping in response. Such significant changes in microbial composition can negatively affect downstream analysis, especially among studies with small sample sizes. Several ways to address this issue are to remove rare taxa during quality control or up-sample to higher taxonomic levels to reduce the proportion of zeros. MMUPHin, for instance, includes a filtering step to remove problematic taxa prior to batch effect correction (S. Ma et al., 2022). However, these practices result in the loss of valuable data, specifically that of rare microbes, which may play active roles in microbial communities. Given that we may want to keep rare taxa for downstream analysis, it is important to address this issue to retain as many rare taxa as possible.

4.1.3 Imputation

High sparsity in microbiome data is an example of the common occurrence of missing data often observed in real-world datasets. Missing data can introduce a significant degree of bias, make processing and analyzing data more difficult, and reduce efficiency. One way to deal with missing data is to remove observations with missing data, but as a result, valuable information is lost. A popular alternative is imputation, a widely used technique to replace missing data with estimated or predicted values and facilitate data analysis. First conceived in 1930, this technique has benefited various fields by providing a way to handle missing data and create more complete datasets for statistical analysis (Allan & Wishart, 1930). Imputation methods have success in recommendation systems, imaging, and finance (Bryzgalova et al., 2022; Dalca et al., 2018; Yuan et al., 2019). In medicine, imputation has been used to improve breast cancer prognosis accuracy, inform discharge assessment of patients with spontaneous supratentorial intracerebral hemorrhage,

and handle missing data in randomized clinical trials (Jakobsen et al., 2017; Jerez et al., 2010; H. Wang et al., 2022). As such, there is appeal in using imputation to solve the issue of sparsity in data and enhance the statistical power of identifying significant features or signals.

Several tools exist to impute missing values from compositional data, one of which is *zCompositions*, an R package used to impute missing values in multivariate data with left-censored values under a compositional approach (Palarea-Albaladejo & Martín-Fernández, 2015). Unlike standard approaches, it considers aspects such as scale invariance, sub-compositional coherence, and preservation of the relative variance structure. These properties are desirable under a compositional approach to data analysis. One of the challenges with imputing missing values in compositional data is that the sum of the parts must always equal 100%, which means that the values of the other parts must change if one part is imputed. *zCompositions* addresses this challenge by transforming the compositional data into an isometric log-ratio (ilr) space. By transforming each observation with respect to an orthonormal basis to get the ratios between components, the data is isometrically mapped into a space of real coordinates, which allows for the use of standard statistical methods and more straightforward imputation of missing values (Egozcue et al., 2003). By using the ilr transformation, *zCompositions* can impute missing values in compositional data while preserving the constraints of the compositional space. This can result in more accurate imputations and more reliable statistical analyses of compositional data. Given its ability to impute missing values in compositional data, *zCompositions* could potentially be used to tackle the challenge of ComBat-Seq by generating additional non-

zero read counts before implementing batch effect correction.

In this chapter, I introduce iComBat-Seq, which expands the ComBat-Seq adjustment framework to address the challenges in batch correction among features with high proportions of zeros. Specifically, I use imputation to replace zeroes in the data with estimated positive values prior to batch effect adjustment using a negative binomial regression model. I demonstrate that, by using imputation to enhance batch effect correction, iComBat-Seq can retain a greater number of rare taxa while preserving similar benefits in differential abundance analysis compared to other adjustment methods.

4.2 Methods

4.2.1 Imputing Zeroes in Microbiome Data using zCompositions

To enhance the performance of iComBat-Seq in correcting for batch effect, prior to batch effect correction, I performed imputation using the zCompositions R package (v 1.4.0-1) to replace zeroes in the dataset with estimated non-zero read counts. To avoid mixture of predicted non-zero read counts across batches, conditions, and covariates (if present), I split the count table on each variable. Since zCompositions requires non-zero values to estimate new values for zero replacement, any taxa with less than 2 non-zero read counts across all samples belonging to a given batch were excluded from the corresponding batch-specific count table. After filtering, zCompositions was used to replace all zeros in the batch-specific count table. After zero imputation, the removed taxa were added back to the count table. All the batch-specific count tables were then recombined to form a final zero-imputed count table, from which batch effect was removed using ComBat-Seq. Finally, to prevent possible bias caused by imputation in downstream analysis, the original

zeroes before imputation were restored in the corrected count table.

4.2.2 Assessing Outliers and Degree of Adjustment after Batch Effect Correction

As one of the main goals of improving ComBat-Seq is to reduce the prevalence and severity of outliers, I used several metrics to determine the overall performance of both the original ComBat-Seq and iComBat-Seq in avoiding outliers when performing batch effect correction. I also compared the two methods with MMUPHin, which was similarly based on ComBat. Because microbiome count data is often converted into compositional data for downstream analysis, I assessed outliers in both the count and logCPM tables corresponding to the absolute and relative measurements of the data, respectively.

To assess outliers in the absolute measurements, I calculated the relative difference between the adjusted reads and the original reads (excluding zeroes) with the original read as the reference. Using the calculated relative differences, I identified potential outliers using the interquartile range (IQR) criterion: any observation was identified as an outlier if it was above $q_{0.75} + (1.5 \times IQR)$ or below $q_{0.25} - (1.5 \times IQR)$, where $q_{0.25}$ and $q_{0.75}$ correspond to the first and third quartiles, respectively, and IQR is the difference between the third and first quartile.

For outlier detection in the logCPM table, I calculated the difference between the adjusted and unadjusted logCPM (excluding zeroes). Because the calculated differences followed a normal distribution (confirmed visually using a Q-Q plot), I converted the differences into z-scores, using the following equation:

$$z_i = \frac{x_i - \mu}{\sigma}$$

Where μ and σ are respectively, the mean and standard deviation of differences between adjusted and unadjusted logCPM across all taxa and z_i and x_i are the z-score and observed difference for taxon i , respectively. Observations with an absolute z-score greater than 3 were identified as outliers.

For overall degree of adjustment, using the original or unadjusted measurements as the baseline, I calculated the root-mean-square-error (RMSE) using the adjusted values as the “actual” observations and the original values as the “estimated” observations, given by the equation:

$$RMSE = \sqrt{\frac{1}{K} \sum_{i=0}^K (a_i - u_i)^2}$$

Where K is the number of taxa and a_i and u_i are the adjusted and unadjusted measurements of taxon i , respectively. This metric can be used to assess the degree to which the data is adjusted after batch effect correction. An RMSE of zero would indicate no difference between adjusted and original values while a greater RMSE would indicate a greater overall degree of adjustment.

4.2.3 16S rRNA Simulations

To evaluate the effectiveness of zero imputation in improving batch effect correction using ComBat-Seq, I used a simulated 16S rRNA dataset from the publicly available MOMS-PI dataset to measure the effectiveness of removing batch effect while preserving condition effects as well as correctly identifying differentially abundant taxa. Similar to the simulation-based analysis in Chapter 2, I preprocessed the dataset to include only buccal mucosa samples collected during the first visit and have a library size of at

least 4,000 reads and OTUs with less than 98% proportion of zeroes. The resulting dataset consisted of 361 samples and 853 OTUs. In addition to introducing a preset batch and condition effect to the data, I added three biological covariates, each with two levels: negative (0) and positive (1). For each covariate, I randomly divided the samples into two groups using a Bernoulli distribution with a probability of 0.5. For samples assigned to the positive group, I simulated 20 randomly selected taxa to be differentially abundant compared to the negative group; read counts of the 10 taxa selected to be more abundant were multiplied by a covariate FC of 2 and read counts of the 10 taxa selected to be less abundant were divided by the covariate FC of 2.

Performance in differential abundance analysis was measured by calculating sensitivity ($\frac{\text{true positive}}{\text{positive}}$), specificity ($\frac{\text{true negative}}{\text{negative}}$), precision ($\frac{\text{true positive}}{\text{true positive} + \text{false positive}}$), and F1 ($\frac{2(\text{true positive})}{2(\text{true positive}) + \text{false positive} + \text{false negative}}$). Performance in removing overall batch effect while preserving condition effect was measured using the R2 metric from PERMANOVA.

I repeated the simulation while varying the batch and condition FC. Specifically, for the evaluation on outliers, I changed the parameter for condition effect such that the condition FC is 4, 16, 32, and 64 while keeping the batch and covariate effects constant at 16 and 2, respectively. Results were averaged over 50 simulations for each condition FC. For the evaluation of differential abundance analysis, I used two different parameter settings: 1) condition FC of 64 and batch FC of 16, and 2) condition FC of 16 and batch FC of 64. In both settings, the covariate FC was set to 2. Results were averaged over 500 repeated simulations under each parameter setting.

For the evaluation of outliers, I compared the original ComBat-Seq, iComBat-Seq, and MMUPHin. For the evaluation of differential abundance analysis, I compared the original ComBat-Seq, iComBat-Seq, MMUPHin, and ConQuR with both a baseline simulated data (with condition and covariate effects and no batch effect) and an uncorrected simulated data (with condition, covariate, and batch effects). For the uncorrected simulated data, I included batch as a covariate to simulate accounting for batch effect. In addition, to test the impact of zero imputation on downstream analysis, I tested iComBat-Seq without restoring the original zeros and ConQuR with the restored original zeroes.

4.2.4 Real data application using shotgun metagenomic sequencing BWHS data

I applied the proposed iComBat-Seq approach on a real-world shotgun metagenomic dataset with batch effect from the Black Women's Health Study (BWHS). As described in Chapter 3, oral wash samples collected from 648 adult Black women were processed under one protocol, but library preparation was performed in three batches at different laboratories. After mapping metagenomic reads to a comprehensive reference database of prokaryotic, eukaryotic, and viral genomes from RefSeq to produce an OTU count table representing bacterial, viral, and fungal reads, I aggregated the data to the genus level and excluded any lineage with at least 98% proportion of zeroes across all samples. I also filtered out any samples with either a read depth of less than 4,000 reads or missing data for smoking status, age, BMI, gingivitis, and/or alcohol consumption. The resulting processed data included 962 genera and 635 samples.

Using the processed data, I removed batch effect using ComBat-Seq, iComBat-Seq, MMUPHin, and ConQuR. For all methods, I selected smoking status, age, BMI, gingivitis,

and alcohol consumptions as covariates to preserve. I evaluated each method's effectiveness for removing batch effect while preserving the desired condition effect (for smoking status) numerically using PERMANOVA R^2 (with the Robust Aitchison method). I performed differential abundance analysis using ANCOM-BC to identify differentially abundant taxa that are associated with smoking status (adjusting for age, BMI, gingivitis, and alcohol consumption as categorical variables).

4.3 Results

4.3.1 iComBat-Seq retains more rare taxa compared to ComBat-Seq

When correcting for batch effect in the simulated dataset after removing taxa with at least 95% proportion of zeroes across all samples, iComBat-Seq produced fewer outliers overall compared to the original ComBat-Seq and MMUPHin at both the absolute count and relative abundance scales. In all simulations at varying condition FC of 4, 16, 32, 48, and 64, iComBat-Seq produced the fewest number of outliers: on average, the proportions of identified outliers in the iComBat-Seq-corrected data were 2.0% on the absolute count scale and 0.05% on the logCPM scale (Figure 4.1). In contrast, the ComBat-Seq and MMUPHin-corrected data had 2.2% and 2.4% proportion of outliers, respectively, on the absolute count scale and 0.05% and 0.11%, respectively, on the logCPM scale. Most outliers from both methods derived from the same group of taxa with a high proportion of zero read counts, but iComBat-Seq reduced the number of outliers by an average of 13.5% ($P < 0.0001$) compared to the original ComBat-Seq and 19.3% ($P < 0.0001$) compared to MMUPHin. When the data was converted from raw counts to a compositional format such as logCPM, iComBat-Seq reduced the number of outliers by an average of 10.1% ($P <$

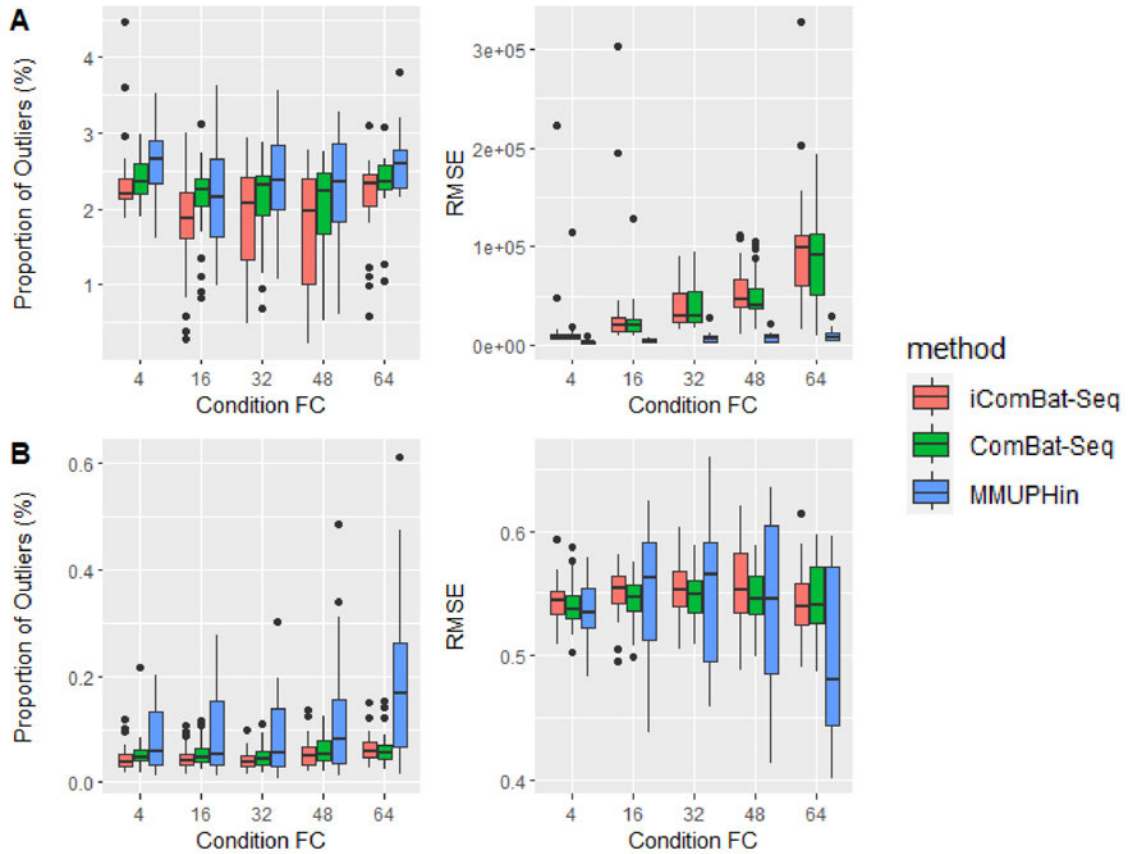


Figure 4.1 Comparison of outliers after batch effect correction using iComBat-Seq, ComBat-Seq, and MMUPHin. In both the A) absolute count and B) logCPM scales, the iComBat-Seq-corrected data produced the fewest outliers overall compared to the ComBat-Seq and Mmusi corrected data in all scenarios (condition FC = 4, 16, 32, 48, and 64) with 50 simulations each. With the lowest overall RMSE, MMUPHin was the most conservative in its read adjustments. In all scenarios, batch and covariate FCs were set to 16 and 2, respectively.

0.0001) compared to the original ComBat-Seq and 4.7% ($P < 0.0001$) compared to MMUPhin. Overall, iComBat-Seq produced the fewest number of outliers in the simulated 16S data compared to ComBat-Seq and MMUPHin, which suggests that it can retain a higher number of taxa for downstream analysis, though some pre-filtering is still required to further reduce the issue of outliers.

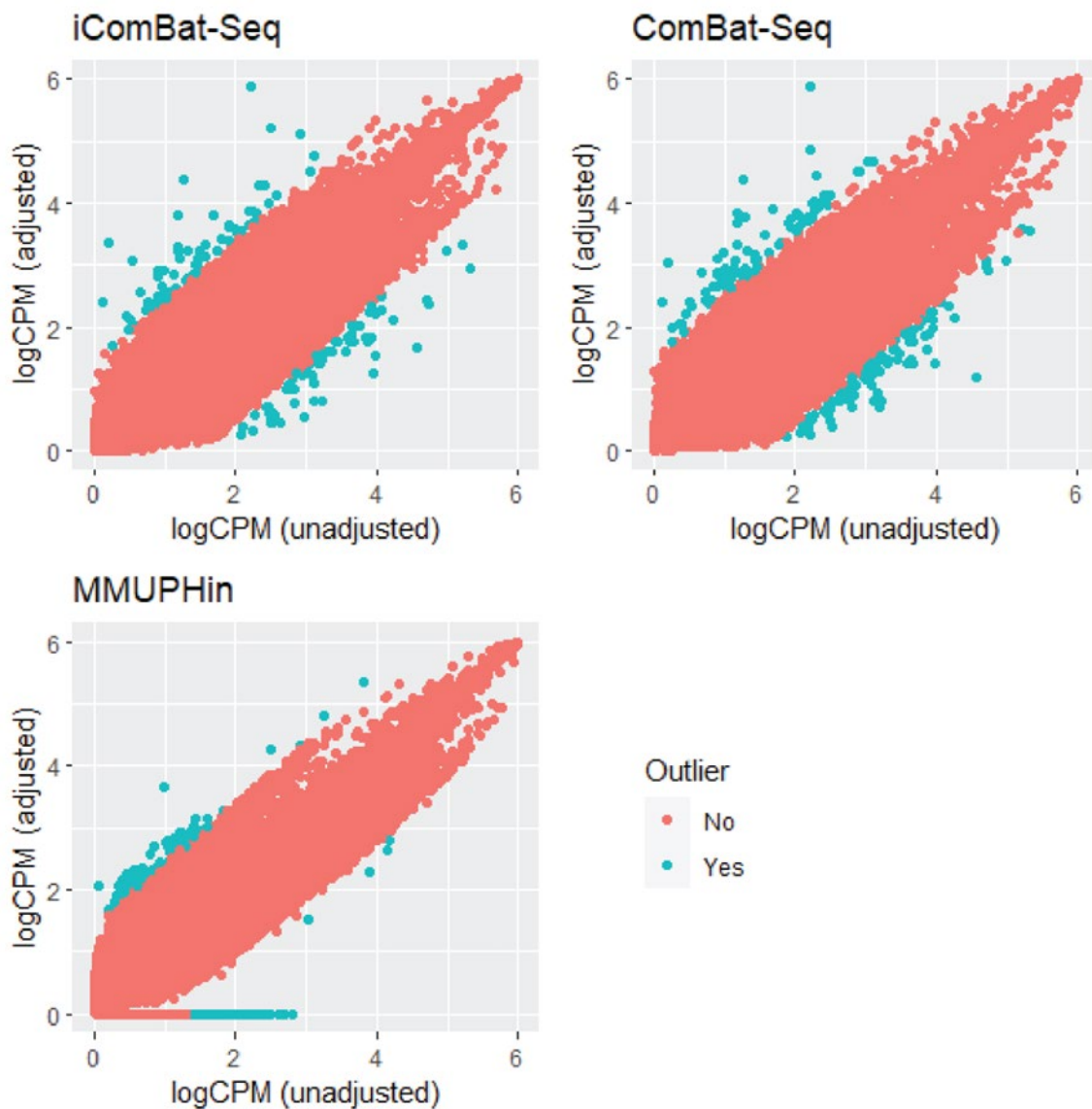


Figure 4.2 Read adjustments on logCPM scale using iComBat-Seq, ComBat-Seq, and MMUPHin. iComBat-Seq produced the fewest outliers in the logCPM scale compared to ComBat-Seq and MMUPHin. Although MMUPHin was more conservative in its adjustments, many of its outliers were reads being adjusted from moderate to extremely low abundances.

Interestingly, MMUPHin was the most conservative in its batch effect correction compared to iComBat-Seq and ComBat-Seq. On both the absolute count and relative abundance scales, MMUPHin had the lowest root mean square error (RMSE), suggesting

that most read counts had smaller adjustments after MMUPHin. However, despite having the lowest RMSE, MMUPHin produced the greatest number of outliers, most of which stem from reads being corrected from moderate to low abundances (Figure 4.2). These results suggest that MMUPHin is unsuitable for batch effect correction on microbiome data with outliers, particularly in studies with a small sample size.

4.3.2 Comparing iComBat-Seq with Existing Batch Effect Correction Methods using Simulated 16S Data

When performing batch effect correction on a simulated 16S dataset with a preset batch, condition, and covariate effect, iComBat-Seq performed similarly to ComBat-Seq and other existing methods when removing batch effect while preserving condition effect (Figure 4.2). MMUPHin performed the best in removing batch effect and better than ComBat-Seq and iComBat-Seq in preserving the condition effect. While ConQuR enhanced the condition effect the most, it also performed the worst in removing batch effect. A possible reason for ConQuR's poor performance is its use of zero imputation to replace sampling zeros with predicted non-zero reads. When zeros were restored in the ConQuR-corrected data, the proportions of variance explained by batch and condition effects dropped to similar levels as those observed in the other methods. When imputed values were not restored to zeros in the iComBat-Seq-corrected data, the proportions of variance for both batch and condition effects increased significantly. Together, these results confirm that, while imputation can improve detection of taxa associated with the condition effect, it also enhances rather than reduces the batch effect, which can result in an increase in false discoveries in downstream analysis.

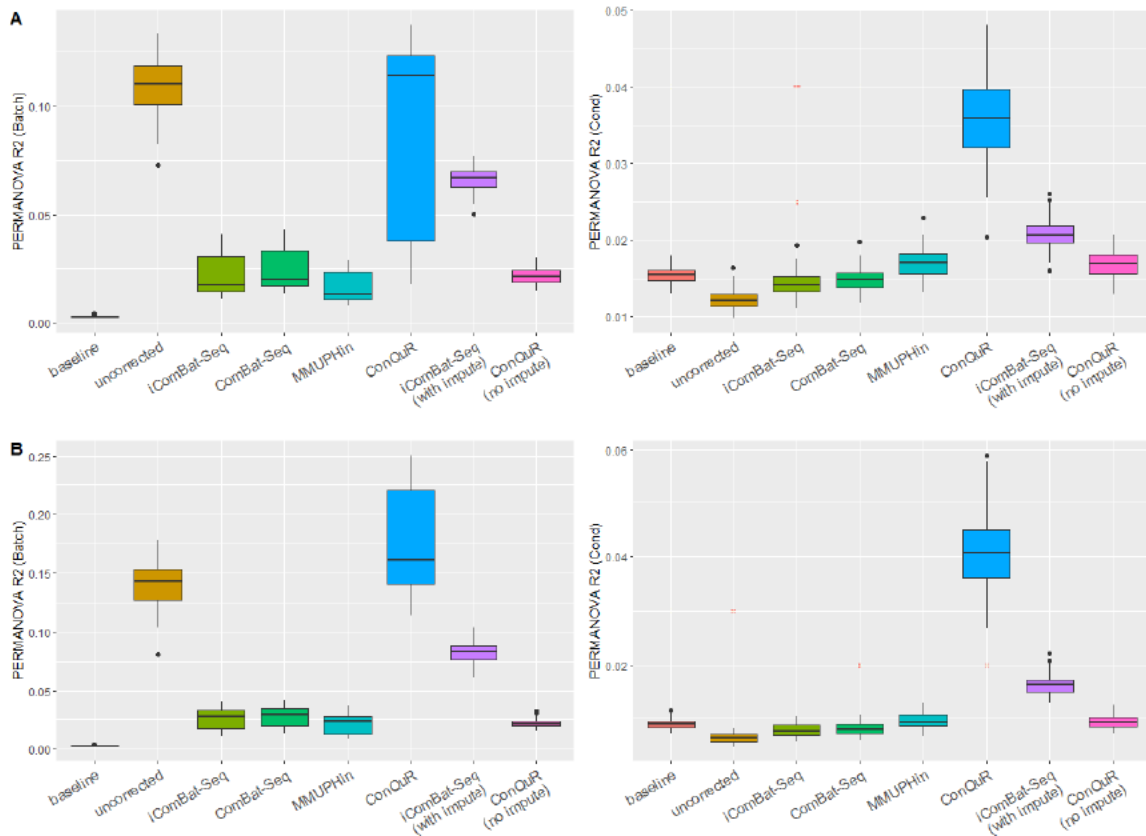


Figure 4.3 Proportion of variance explained by batch and condition effects. In simulations with A) batch FC = 16, condition FC = 64, and covariate FC = 2 and B) batch FC = 64, condition FC = 16, and covariate FC = 2, iComBat-Seq showed similar performances in removing batch effect and retaining batch effect compared to ComBat-Seq. MMUPHin performed the best in removing batch effect while preserving condition effect. Methods with imputed zeros (ConQuR and iComBat-Seq with impute) performed the worst in removing batch effect but enhanced the condition effect. Converting imputed zeros back to zeros in the ConQuR-corrected data resulted in a reduction of variance explained by both batch and condition effects.

In downstream analysis identifying differentially abundant taxa associated with the condition effect, ANCOM-BC performed the best in identifying the true differentially abundant taxa in the iComBat-Seq-corrected data compared to the other methods (Figure 4.4). Although iComBat-Seq had a slightly lower sensitivity compared to ComBat-Seq, it had a higher precision, resulting in an overall higher F1-score. In contrast, ConQuR performed the worst among the tested methods: although it had a highest sensitivity, it also

had a particularly low precision, meaning that it incorrectly identified many taxa as associated with the condition effect. iComBat-Seq also performed the best in identifying DA features among rare taxa (with average relative abundance $< 0.01\%$). While none of the methods (except for ConQuR) performed well in sensitivity, iComBat-Seq performed the best in precision, resulting in the best F1-score among the tested methods. The inability to detect rare DA taxa may be due to ANCOM-BC rather than the batch effect correction methods themselves since there are too few non-zero read counts among rare taxa to detect signals with adequate statistical power. Indeed, ConQuR and iComBat-Seq (with imputed values) showed higher sensitivity among rare taxa, confirming that detection of rare DA taxa can be achieved with more non-zero read counts.

Interestingly, using the iComBat-Seq-corrected data without restoring the original zeroes resulted in a higher false discovery rate (or lower precision). On the other hand, restoring the original zeros after imputation and batch effect correction resulted in slightly improved performances in compared to the original ComBat-Seq ($P = 0.11$) and MMUPHin ($P < 0.0001$). The same trend can be observed in the ConQuR-corrected data, which has imputed reads. ConQuR performed the worst in the F1-score due to having a particularly high false discovery rate (or low precision). However, after restoring the original zeros, the precision increased significantly, resulting in a higher F1-score for the ConQuR-corrected data. This pattern confirms that, although imputation can improve batch effect correction, if left in the corrected data, it can cause adverse effects on downstream analysis by introducing artificial signals, particularly among taxa with a higher initial proportion of zeroes.

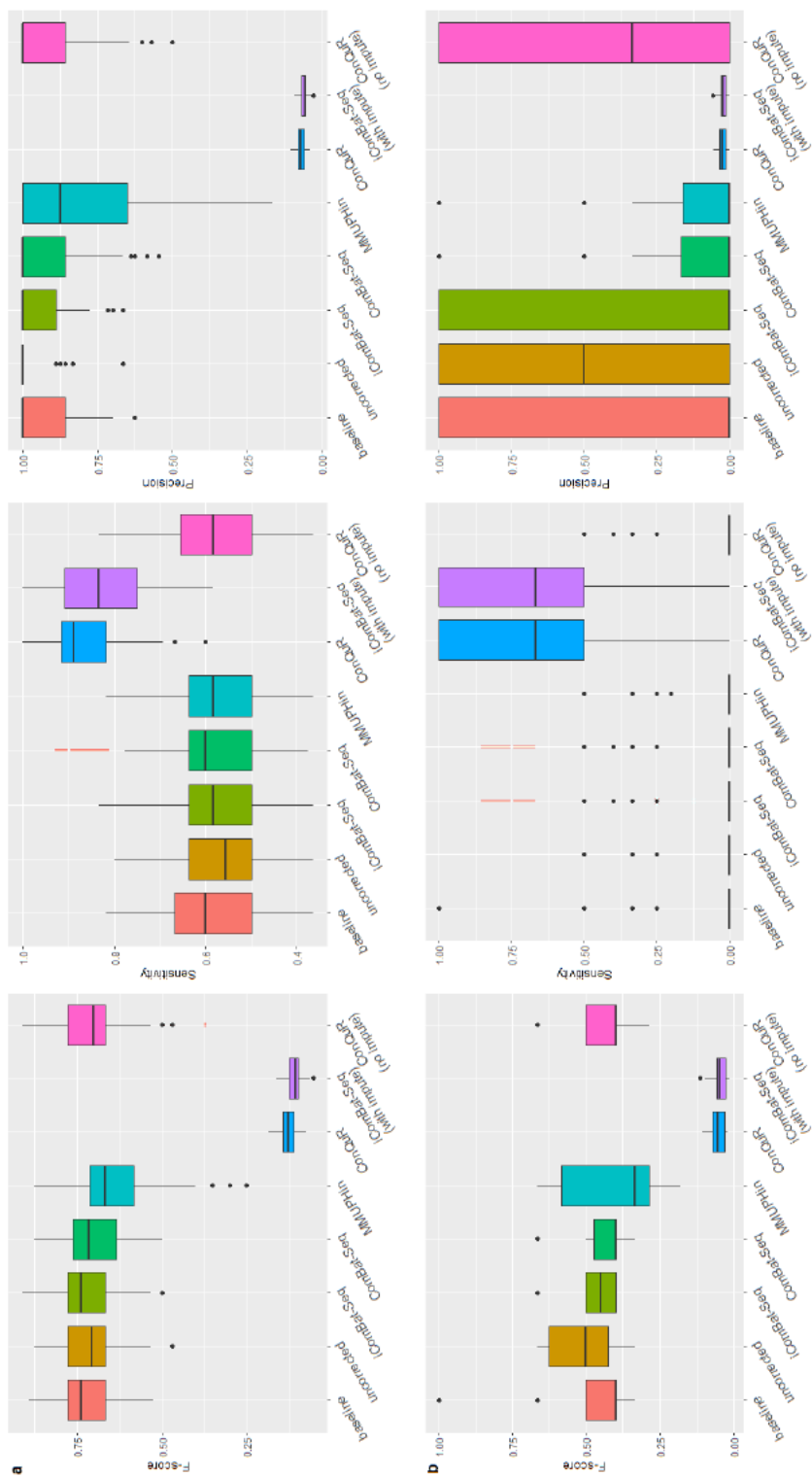


Figure 4.4 Performance of DA analysis using uncorrected and corrected data (Batch FC = 16, Condition FC = 64).
 A) iComBat-Seq performed the best overall in DA analysis on the simulated 16S dataset compared to other methods. Although its sensitivity was slightly lower than the original ComBat-Seq, iComBat-Seq performed better in precision, resulting in a higher F1 score. ConQuR performed the worst despite having the highest sensitivity due to having poor precision due to its use of imputed values. When imputed values were restored to zero, ConQuR's precision increased at the cost of lower sensitivity. iComBat-Seq showed similar poor performance as ConQuR when the imputed values in the iComBat-Seq-corrected data were not restored to zero. B) Among rare taxa (average relative abundances < 0.01%), iComBat-Seq performed the best compared to the other methods, most of which suffer from poor precision.

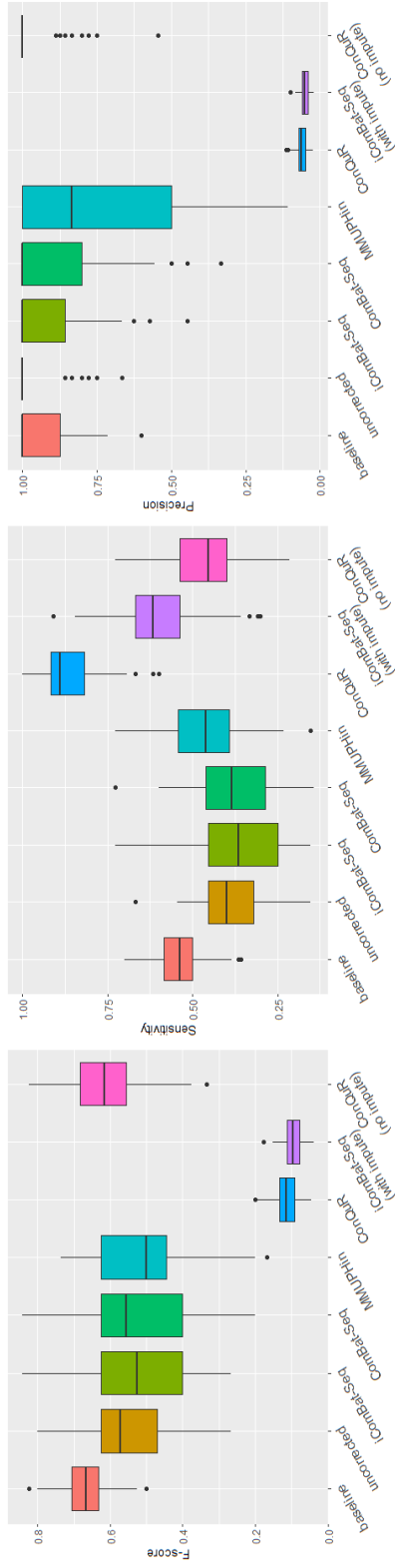


Figure 4.5 Performance of DA analysis using uncorrected and corrected data (Batch FC = 64, Condition FC = 16). In scenarios in which the batch FC is greater than the condition FC, iComBat-Seq performed the best in DA analysis compared to other methods. Although iComBat-Seq had a lowest sensitivity, it had the highest precision. While ConQuR performed the worst among the methods due to having the lowest precision, when imputed values were restored to zero in the ConQuR-corrected data, it showed the best overall performance.

Even in cases in which the batch effect was greater than the condition effect, iComBat-Seq performed the best in DA analysis compared to other methods (Figure 4.5). While iComBat-Seq did not have the highest sensitivity, it had a higher precision compared to ComBat-Seq, MMUPHin, and ConQuR, resulting in a higher F1 score.

Overall, these results confirm that iComBat-Seq perform better than existing batch effect correction methods in removing batch effect with varying degrees from 16S rRNA datasets and preserving biologically relevant signals to detect DA taxa in downstream analysis.

4.3.3 Applying iComBat-Seq in a Real-World Epidemiological BWHS Study

For the BWHS dataset, iComBat-Seq performed better than ComBat-Seq in reducing the variability explained by batch effect (0.87% vs 0.91% vs 1.15% in the iComBat-Seq-corrected, ComBat-Seq-corrected, and original data, respectively), while maintaining the variability that can be explained by smoking status (0.41% vs 0.42% vs 0.40%). ConQuR performed the best among the evaluated methods, reducing the batch effect to 0.52% and increasing the smoking status effect to 0.47%.

Of the four methods, ConQuR made the largest adjustments to the BWHS data (Figure 4.6). On the relative (logCPM) scale, 0.97% of the ConQuR-corrected data were outliers. Interestingly, a portion of the outliers were reduced from a positive to zero logCPM, suggesting significant reductions in relative proportions for genera that were originally more abundant in some samples. In contrast, 0.59% of the ComBat-Seq-corrected data and 0.55% of the iComBat-Seq-corrected data were outliers. MMUPHin was the most conservative in its adjustments, producing only 0.26% outliers.

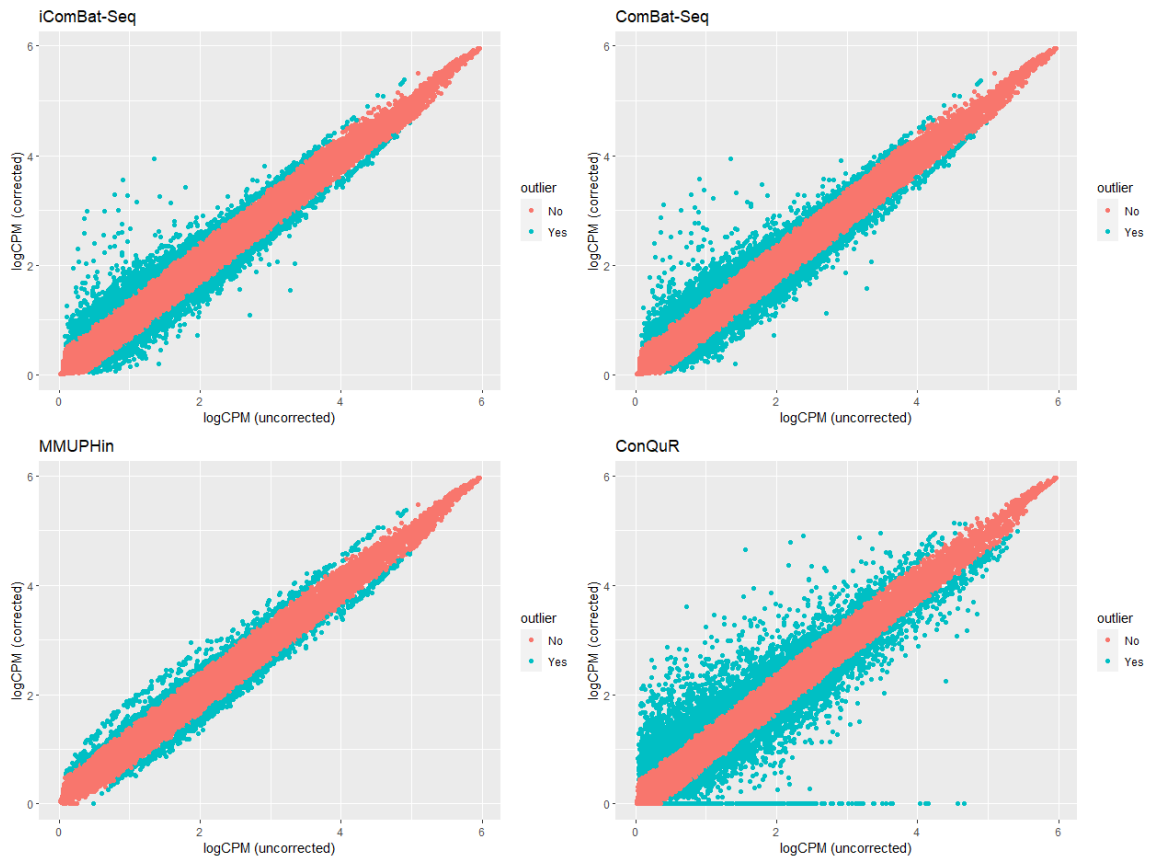


Figure 4.6 Read adjustments on BWHS dataset using iComBat-Seq, ComBat-Seq, MMUPHin, and ConQuR. iComBat-Seq reduced the number of outliers after correction on the BWHS dataset compared to the original ComBat-Seq (0.55% vs. 0.59% outliers, respectively). In contrast, ConQuR made the largest adjustments, resulting the highest proportion of outliers in the corrected data (0.97%), while MMUPHin was the most conservative, producing a corrected data with only 0.26% outliers.

For the differential abundance analysis, at FDR $\alpha = 0.05$, ANCOM-BC2 (adjusting for age, BMI, smoking status, gingivitis, and, for uncorrected data, batch) found 21 genera to be associated with smoking status in the iComBat-Seq-corrected data, 4 of which were not identified in the ComBat-Seq-corrected data. Most notably, *Aggregatibacter* (adjusted $P = 0.043$, also identified by MMUPHin and ConQuR) and *Slackia* (adjusted $P = 0.048$, also identified by MMUPHin and ConQuR) were detected to be DA in the iComBat-Seq-corrected data. *Slackia* is a known bacterium found to be more prevalent among smokers

compared to non-smokers (Karasneh et al., 2017). *Aggregatibacter* has been found to be decreased among cigarette smokers compared to non-smokers with moderate-high caries (Al-Marzooq et al., 2022). Overall, I confirm that iComBat-Seq performs better than the original ComBat-Seq in removing the confounding batch effect and revealing the true signals.

The MMUPHin and ConQuR-corrected data did reveal more genera to be associated with smoking status (30 and 32 genera, respectively). However, as evidenced in the DA analysis using the simulated 16S dataset, both methods suffer from lower precisions compared to iComBat-Seq, raising concerns that the additional genera identified from the MMUPHin and ConQuR-corrected data may be false positives. Thus, while iComBat-Seq detected fewer associated genera compared to MMUPHin and ConQuR, iComBat-Seq may be better in identifying the true associated genera.

4.4 Discussion

Batch effect correction methods have been effective in removing batch effect from all types of omics data, but challenges remain. Mainly, they are imperfect when dealing with features with high sparsity, which can cause outliers to be adjusted incorrectly. Improper adjustments can lead to erroneous conclusions in downstream etiological analysis, especially in studies with small sample sizes. On the other hand, the removal of such features during the preprocessing step can lead to overlooking potentially significant members of the microbial community. In this chapter, my proposed method, iComBat-Seq, demonstrates that imputation can reduce the issue of outliers in microbiome data and allow the retention of more rare microbes for downstream analysis.

When compared to the original ComBat-Seq and MMUPHin, iComBat-Seq produced fewer improper adjustments, measured by the number of outliers produced after correction. While ComBat-Seq intensifies certain outliers due to having too few non-zero read counts to fit a robust model, iComBat-Seq addresses this issue by replacing zeroes with predicted non-zero values while retaining the taxa's non-zero abundance distributions. The main issue is that, when modeling the read distribution for rare taxa, due to having only a few non-zero read counts, outliers have a significant influence on the model. While MMUPHin is intentionally conservative, only correcting batch differences that can be confidently inferred, it still produces outliers, mainly due to several taxa shrinking from moderate to low abundances. By generating more non-zero read counts, iComBat-Seq can simultaneously reduce the impact of outliers and provide more data for the model, resulting in more robust models to estimate and remove batch effect, improved shrinkage, and fewer cases in which outliers are inappropriately handled.

In addition to reducing the number of outliers after correction, iComBat-Seq improves the performance in selecting predictive taxa for conditions or variables of interest. In the DA analysis on the simulated 16S dataset, I show that, compared to the original ComBat-Seq, MMUPHin, and ConQuR, iComBat-Seq performs the best overall in identifying the true taxa associated with the condition while maintaining a high precision. In the DA analysis on a real-world shotgun metagenomic dataset, iComBat-Seq was able to identify more true genera associated with smoking status. The main advantage of iComBat-Seq is its capability to address zero inflation, which prevents stable modeling for rare taxa, a problem that affects many methods.

One downside I observed is that imputation can negatively impact downstream analysis. When the imputed values were kept in the iComBat-Seq-corrected data, although the DA analysis identified slightly more true taxa associated with the condition effect, it also identified many false positives, resulting in a lower precision and an overall lower F1-score compared to the iComBat-Seq-corrected data with the imputed values restored to zero. I saw a similar pattern with ConQuR, which performs its own version of imputation for zeroes predicted to be due to under-sampling. The ConQuR-corrected data also had a lower precision compared to other methods, but when I converted the imputed values back to zero, the precision increased significantly with the new ConQuR-corrected data. Together, these results confirm that imputation can introduce unwanted variance in the data, causing a higher rate of false positives in downstream analysis. However, when used solely for batch effect correction, it can strengthen the modeling performed to estimate and remove batch effect. An ideal use would be to apply imputation only for batch effect correction and restore the original zeroes before downstream analysis. This approach takes advantage of the benefits of imputation in batch effect correction while avoiding downstream issues of undesired artificial noise.

Despite the advantages of iComBat-Seq, it has several limitations. First, when additional covariates are considered, the overall impact of iComBat-Seq decreases. During the imputation step, it divides the data by batch and desired covariates to preserve. As more covariates are included, the data is split into smaller chunks, increasing the probability of having all zeroes for certain features in one or more of these chunks. In such cases, iComBat-Seq will not impute for those features, resulting in less zero replacement in the

imputed data. Second, while iComBat-Seq does improve the retention of more rare taxa compared to other methods, it is still imperfect for extremely low-frequency taxa. Thus, some filtering is still necessary.

The main purpose of incorporating imputation into ComBat-Seq is to retain more rare taxa for downstream analysis. While the statistical challenge of identifying association between rare taxa and condition remains, iComBat-Seq can preserve more rare taxa without suffering from erroneous adjustments, allowing for more useable data for downstream analysis.

Chapter 5: Conclusion

The advancement of microbiome research (and any study involving high-throughput data) is hampered by batch effect, often because of technical limitations in processing large amounts of data in a single batch, and a lack of ethnic diversity in past microbiome profiling studies. The driving motivation behind the research described in this dissertation is to 1) address the issue of batch effect in microbiome data through batch effect correction methods and 2) apply them in epidemiological studies to better understand the relationship between oral microbial composition and environmental exposures among adult Black women living in the US. While constraints in the lab preventing precise control of technical variables across batches persist, there is a need for computational solutions to address batch effects to prevent unwanted bias in downstream analysis, either by accounting for them in statistical models or outright removing them using batch effect correction methods. Furthermore, although there have been large-scale microbiome profiling studies such as the Human Microbiome Project (HMP), they are limited to a single ethnicity, with other ethnic groups having little to no representation. This disparity becomes an issue due to associations between ethnicity and certain diseases. To this end, I present the evaluation, application, and improvement of ComBat-Seq for the removal of batch effect in microbiome data to improve the identification of differentially abundant taxa that are relevant to biological variables of interest. By focusing on adult Black women in the US in an epidemiological study of the oral microbiome, my thesis provides new coverage of a previously under-represented ethnic group in studies aiming to understand the role of oral microbial composition and human health.

In Chapter 2, I performed a detailed evaluation of ComBat-Seq, comparing it with existing microbiome-specific methods, in removing batch effect from both simulated and real-world microbiome data while preserving effects belonging to biological factors of interest. I demonstrated that ComBat-Seq has comparable performance in removing batch effect from both 16S rRNA and shotgun metagenomic sequencing data while retaining enough of the condition effect to detect differentially abundant taxa associated with that condition in downstream analysis.

In Chapter 3, I applied ComBat-Seq in an epidemiological study in which I analyzed shotgun metagenomic sequencing data obtained from oral wash samples of adult Black women living in the US to explore the relationship between oral microbial composition and host's geographical location. After correcting for batch effect, a differential abundance analysis revealed several oral health-related genera to be associated with specific regions in the US, emphasizing the importance of considering social and environmental factors associated with the host's geographical location when studying the oral microbiome. The results derived from this study may guide future work studying associations between oral microbial composition and oral health from samples collected across differing geographic regions.

In Chapter 4, I introduced iComBat-Seq, an extension to ComBat-Seq that improves its performance in batch effect correction on rare taxa with outliers via imputation. By reducing the proportion of zeroes in the originally sparse microbial data and generating predicted non-zero read counts, which follow the observed compositional structure of the data, for more robust fitting of the negative binomial regression model, imputation

effectively reduced the number of problematic cases in which outliers were intensified after batch effect correction. This work demonstrates the potential of imputation in resolving a critical weakness among batch effect correction methods in handling low-frequency taxa (and highly sparse data in general).

Throughout my dissertation work, I present both current and new methods in batch effect correction specific for microbiome data and their applications in improving studies focused on identifying microbial compositions associated with biological variables. As microbiome profiling studies continue to grow in size, leading to increased chances of batch effect, computational and statistical approaches to address batch effect will become increasingly important. The work presented here represents significant contributions to the area of batch effect correction on 16S rRNA and shotgun metagenomics, which will empower future discoveries in the field of the microbiome.

Chapter 6: BIBLIOGRAPHY

- Adams-Campbell, L. L., Dash, C., Palmer, J. R., Wiedemeier, M. V., Russell, C. W., Rosenberg, L., & Cozier, Y. C. (2016). Predictors of biospecimen donation in the Black Women's Health Study. *Cancer Causes & Control: CCC*, 27(6), 797–803.
- Allan, F. E., & Wishart, J. (1930). A method of estimating the yield of a missing plot in field experimental work. *The Journal of Agricultural Science*, 20(3), 399–406.
- Al-Marzooq, F., Al Kawas, S., Rahman, B., Shearston, J. A., Saad, H., Benzina, D., & Weitzman, M. (2022). Supragingival microbiome alternations as a consequence of smoking different tobacco types and its relation to dental caries. *Scientific Reports*, 12(1), 2861.
- Anderson, M. J. (2017). Permutational multivariate analysis of variance (PERMANOVA). In *Wiley StatsRef: Statistics Reference Online* (pp. 1–15). Wiley. <https://doi.org/10.1002/9781118445112.stat07841>
- Asnicar, F., Berry, S. E., Valdes, A. M., Nguyen, L. H., Piccinno, G., Drew, D. A., Leeming, E., Gibson, R., Le Roy, C., Khatib, H. A., Francis, L., Mazidi, M., Mompeo, O., Valles-Colomer, M., Tett, A., Beghini, F., Dubois, L., Bazzani, D., Thomas, A. M., ... Segata, N. (2021). Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. *Nature Medicine*, 27(2), 321–332.
- Baker, G. C., Smith, J. J., & Cowan, D. A. (2003). Review and re-analysis of domain-specific 16S primers. *Journal of Microbiological Methods*, 55(3), 541–555.

- Becker, K., Harmsen, D., Mellmann, A., Meier, C., Schumann, P., Peters, G., & von Eiff, C. (2004). Development and evaluation of a quality-controlled ribosomal sequence database for 16S ribosomal DNA-based identification of *Staphylococcus* species. *Journal of Clinical Microbiology*, *42*(11), 4988–4995.
- Bescos, R., Ashworth, A., Cutler, C., Brookes, Z. L., Belfield, L., Rodiles, A., Casas-Agustench, P., Farnham, G., Liddle, L., Burleigh, M., White, D., Easton, C., & Hickson, M. (2020). Effects of Chlorhexidine mouthwash on the oral microbiome. *Scientific Reports*, *10*(1), 5254.
- Brown, E. M., Sadarangani, M., & Finlay, B. B. (2013). The role of the immune system in governing host-microbe interactions in the intestine. *Nature Immunology*, *14*(7), 660–667.
- Bryzgalova, S., Lerner, S., Lettau, M., & Pelger, M. (2022). Missing financial data. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4106794>
- Bunyavanich, S., Shen, N., Grishin, A., Wood, R., Burks, W., Dawson, P., Jones, S. M., Leung, D. Y. M., Sampson, H., Sicherer, S., & Clemente, J. C. (2016). Early-life gut microbiome composition and milk allergy resolution. *The Journal of Allergy and Clinical Immunology*, *138*(4), 1122–1130.
- Buytaers, F. E., Saltykova, A., Mattheus, W., Verhaegen, B., Roosens, N. H. C., Vanneste, K., Laisnez, V., Hammami, N., Pochet, B., Cantaert, V., Marchal, K., Denayer, S., & De Keersmaecker, S. C. J. (2021). Application of a strain-level shotgun metagenomics approach on food samples: resolution of the source of a *Salmonella*

food-borne outbreak. *Microbial Genomics*, 7(4).

<https://doi.org/10.1099/mgen.0.000547>

Campana, M. G., Robles García, N., Rühli, F. J., & Tuross, N. (2014). False positives complicate ancient pathogen identifications using high-throughput shotgun sequencing. *BMC Research Notes*, 7(1), 111.

Cao, Q., Sun, X., Rajesh, K., Chalasani, N., Gelow, K., Katz, B., Shah, V. H., Sanyal, A. J., & Smirnova, E. (2021). Effects of Rare Microbiome Taxa Filtering on Statistical Analysis. *Frontiers in Microbiology*, 11.

<https://doi.org/10.3389/fmicb.2020.607325>

Carlet, J. (2012). The gut is the epicentre of antibiotic resistance. *Antimicrobial Resistance and Infection Control*, 1(1), 39.

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.

Chang, J. Y., Antonopoulos, D. A., Kalra, A., Tonelli, A., Khalife, W. T., Schmidt, T. M., & Young, V. B. (2008). Decreased diversity of the fecal Microbiome in recurrent *Clostridium difficile*-associated diarrhea. *The Journal of Infectious Diseases*, 197(3), 435–438.

Chen, J., Wright, K., Davis, J. M., Jeraldo, P., Marietta, E. V., Murray, J., Nelson, H., Matteson, E. L., & Taneja, V. (2016). An expansion of rare lineage intestinal microbes characterizes rheumatoid arthritis. *Genome Medicine*, 8(1), 43.

- Chen, W., Zhang, S., Williams, J., Ju, B., Shaner, B., Easton, J., Wu, G., & Chen, X. (2020). A comparison of methods accounting for batch effects in differential expression analysis of UMI count based single cell RNA sequencing. *Computational and Structural Biotechnology Journal*, *18*, 861–873.
- Clarke, G., Stilling, R. M., Kennedy, P. J., Stanton, C., Cryan, J. F., & Dinan, T. G. (2014). Minireview: Gut microbiota: the neglected endocrine organ. *Molecular Endocrinology*, *28*(8), 1221–1238.
- Clarridge, J. E., 3rd. (2004). Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical Microbiology Reviews*, *17*(4), 840–862, table of contents.
- Coelho, G. D. P., Ayres, L. F. A., Barreto, D. S., Henriques, B. D., Prado, M. R. M. C., & Passos, C. M. D. (2021). Acquisition of microbiota according to the type of birth: an integrative review. *Revista Latino-Americana de Enfermagem*, *29*, e3446.
- Couto, N., Schuele, L., Raangs, E. C., Machado, M. P., Mendes, C. I., Jesus, T. F., Chlebowicz, M., Rosema, S., Ramirez, M., Carriço, J. A., Autenrieth, I. B., Friedrich, A. W., Peter, S., & Rossen, J. W. (2018). Critical steps in clinical shotgun metagenomics for the concomitant detection and typing of microbial pathogens. *Scientific Reports*, *8*(1), 13767.
- Cox, M. J., Cookson, W. O. C. M., & Moffatt, M. F. (2013). Sequencing the human microbiome in health and disease. *Human Molecular Genetics*, *22*(R1), R88-94.

- Cozier, Y. C., Palmer, J. R., & Rosenberg, L. (2004). Comparison of methods for collection of DNA samples by mail in the Black Women's Health Study. *Annals of Epidemiology, 14*(2), 117–122.
- Dalca, A. V., Bouman, K. L., Freeman, W. T., Rost, N. S., Sabuncu, M. R., & Golland, P. (2018). Medical image imputation from image collections. *IEEE Transactions on Medical Imaging, 38*(2), 504–514.
- Davis, E., Bakulski, K. M., Goodrich, J. M., Peterson, K. E., Marazita, M. L., & Foxman, B. (2020). Low levels of salivary metals, oral microbiome composition and dental decay. *Scientific Reports, 10*(1), 14640.
- de Goffau, M. C., Charnock-Jones, D. S., Smith, G. C. S., & Parkhill, J. (2021). Batch effects account for the main findings of an in utero human intestinal bacterial colonization study [Review of *Batch effects account for the main findings of an in utero human intestinal bacterial colonization study*]. *Microbiome, 9*(1), 6. Springer Science and Business Media LLC.
- Deo, P. N., & Deshmukh, R. (2019/Jan-Apr). Oral microbiome: Unveiling the fundamentals. *Journal of Oral and Maxillofacial Pathology, 23*(1), 122–128.
- Dunbar, J., Barns, S. M., Ticknor, L. O., & Kuske, C. R. (2002). Empirical and theoretical bacterial diversity in four Arizona soils. *Applied and Environmental Microbiology, 68*(6), 3035–3045.
- Dwiyanto, J., Hussain, M. H., Reidpath, D., Ong, K. S., Qasim, A., Lee, S. W. H., Lee, S. M., Foo, S. C., Chong, C. W., & Rahman, S. (2021). Ethnicity influences the gut

- microbiota of individuals sharing a geographical location: a cross-sectional study from a middle-income country. *Scientific Reports*, *11*(1), 2618.
- Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., & Barceló-Vidal, C. (2003). Isometric Logratio Transformations for Compositional Data Analysis. *Mathematical Geology*, *35*(3), 279–300.
- Eriksson, L., Lif Holgerson, P., & Johansson, I. (2017). Saliva and tooth biofilm bacterial microbiota in adolescents in a low caries community. *Scientific Reports*, *7*(1), 5861.
- Esberg, A., Barone, A., Eriksson, L., Lif Holgerson, P., Teneberg, S., & Johansson, I. (2020). *Corynebacterium matruchotii* Demography and Adhesion Determinants in the Oral Cavity of Healthy Individuals. *Microorganisms*, *8*(11), 1780.
- Fadrosh, D. W., Ma, B., Gajer, P., Sengamalay, N., Ott, S., Brotman, R. M., & Ravel, J. (2014). An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. *Microbiome*, *2*(1), 6.
- Fan, X., Peters, B. A., Jacobs, E. J., Gapstur, S. M., Purdue, M. P., Freedman, N. D., Alekseyenko, A. V., Wu, J., Yang, L., Pei, Z., Hayes, R. B., & Ahn, J. (2018). Drinking alcohol is associated with variation in the human oral microbiome in a large study of American adults. *Microbiome*, *6*(1). <https://doi.org/10.1186/s40168-018-0448-x>
- Fettweis, J. M., Serrano, M. G., Brooks, J. P., Edwards, D. J., Girerd, P. H., Parikh, H. I., Huang, B., Arodz, T. J., Edupuganti, L., Glascock, A. L., Xu, J., Jimenez, N. R., Vivadelli, S. C., Fong, S. S., Sheth, N. U., Jean, S., Lee, V., Bokhari, Y. A., Lara,

- A. M., ... Buck, G. A. (2019). The vaginal microbiome and preterm birth. *Nature Medicine*, 25(6), 1012–1021.
- Foster, J. A., Rinaman, L., & Cryan, J. F. (2017). Stress & the gut-brain axis: Regulation by the microbiome. *Neurobiology of Stress*, 7, 124–136.
- Gaiser, R. A., Halimi, A., Alkharaan, H., Lu, L., Davanian, H., Healy, K., Hugerth, L. W., Ateeb, Z., Valente, R., Fernández Moro, C., Del Chiaro, M., & Sällberg Chen, M. (2019). Enrichment of oral microbiota in early cystic precursors to invasive pancreatic cancer. *Gut*, 68(12), 2186–2194.
- García-Closas, M., Egan, K. M., Abruzzo, J., Newcomb, P. A., Titus-Ernstoff, L., Franklin, T., Bender, P. K., Beck, J. C., Le Marchand, L., Lum, A., Alavanja, M., Hayes, R. B., Rutter, J., Buetow, K., Brinton, L. A., & Rothman, N. (2001). Collection of genomic DNA from adults in epidemiological studies by buccal cytobrush and mouthwash. *Cancer Epidemiology, Biomarkers & Prevention*, 10(6), 687–696.
- 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. <https://doi.org/10.3389/fmicb.2017.02224>
- Goodrich, J. K., Waters, J. L., Poole, A. C., Sutter, J. L., Koren, O., Blekhman, R., Beaumont, M., Van Treuren, W., Knight, R., Bell, J. T., Spector, T. D., Clark, A. G., & Ley, R. E. (2014). Human genetics shape the gut microbiome. *Cell*, 159(4), 789–799.
- Graessler, J., Qin, Y., Zhong, H., Zhang, J., Licinio, J., Wong, M.-L., Xu, A., Chavakis, T., Bornstein, A. B., Ehrhart-Bornstein, M., Lamounier-Zepter, V., Lohmann, T., Wolf,

- T., & Bornstein, S. R. (2013). Metagenomic sequencing of the human gut microbiome before and after bariatric surgery in obese patients with type 2 diabetes: correlation with inflammatory and metabolic parameters. *The Pharmacogenomics Journal*, *13*(6), 514–522.
- Harrison, J. G., Beltran, L. P., Buerkle, C. A., Cook, D., Gardner, D. R., Parchman, T. L., Poulson, S. R., & Forister, M. L. (2021). A suite of rare microbes interacts with a dominant, heritable, fungal endophyte to influence plant trait expression. *The ISME Journal*, *15*(9), 2763–2778.
- Hong, C., Manimaran, S., Shen, Y., Perez-Rogers, J. F., Byrd, A. L., Castro-Nallar, E., Crandall, K. A., & Johnson, W. E. (2014). PathoScope 2.0: A complete computational framework for strain identification in environmental or clinical sequencing samples. *Microbiome*, *2*(1). <https://doi.org/10.1186/2049-2618-2-33>
- Huddleston, J. R. (2014). Horizontal gene transfer in the human gastrointestinal tract: potential spread of antibiotic resistance genes. *Infection and Drug Resistance*, *7*, 167–176.
- Idate, U., Bhat, K., Kotrashetti, V., Kugaji, M., & Kumbar, V. (2020). Molecular identification of Capnocytophaga species from the oral cavity of patients with chronic periodontitis and healthy individuals. *Journal of Oral and Maxillofacial Pathology*, *24*(2), 397.
- Integrative HMP (iHMP) Research Network Consortium. (2019). The Integrative Human Microbiome Project. *Nature*, *569*(7758), 641–648.

- Jakobsen, J. C., Gluud, C., Wetterslev, J., & Winkel, P. (2017). When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. *BMC Medical Research Methodology*, *17*(1). <https://doi.org/10.1186/s12874-017-0442-1>
- Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine*, *50*(2), 105–115.
- Jiang, S., Xiao, G., Koh, A. Y., Kim, J., Li, Q., & Zhan, X. (2021). A Bayesian zero-inflated negative binomial regression model for the integrative analysis of microbiome data. *Biostatistics*, *22*(3), 522–540.
- Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., Leopold, S. R., Hanson, B. M., Agresta, H. O., Gerstein, M., Sodergren, E., & Weinstock, G. M. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications*, *10*(1), 5029.
- Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, *8*(1), 118–127.
- Kaci, G., Goudercourt, D., Dennin, V., Pot, B., Doré, J., Ehrlich, S. D., Renault, P., Blottière, H. M., Daniel, C., & Delorme, C. (2014). Anti-inflammatory properties of *Streptococcus salivarius*, a commensal bacterium of the oral cavity and digestive tract. *Applied and Environmental Microbiology*, *80*(3), 928–934.

- Kapke, P. A., Brown, A. T., & Lillich, T. T. (1980). Carbon dioxide metabolism by *Campylobacter jejuni*: identification, characterization, and regulation of a phosphoenolpyruvate carboxykinase. *Infection and Immunity*, 27(3), 756–766.
- Karasneh, J. A., Al Habashneh, R. A., Marzouka, N. A. S., & Thornhill, M. H. (2017). Effect of cigarette smoking on subgingival bacteria in healthy subjects and patients with chronic periodontitis. *BMC Oral Health*, 17(1), 64.
- Kato, I., Vasquez, A., Moyerbrailean, G., Land, S., Djuric, Z., Sun, J., Lin, H.-S., & Ram, J. L. (2017). Nutritional Correlates of Human Oral Microbiome. *Journal of the American College of Nutrition*, 36(2), 88–98.
- Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L., & Gordon, J. I. (2011). Human nutrition, the gut microbiome and the immune system. *Nature*, 474(7351), 327–336.
- 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.
- Kilian, M., Chapple, I. L. C., Hannig, M., Marsh, P. D., Meuric, V., Pedersen, A. M. L., Tonetti, M. S., Wade, W. G., & Zaura, E. (2016). The oral microbiome – an update for oral healthcare professionals. *British Dental Journal*, 221(10), 657–666.
- Kim, Y. S., Unno, T., Kim, B. Y., & Park, M. S. (2020). Sex differences in gut Microbiota. *The World Journal of Men's Health*, 38(1), 48–60.
- Kort, R., Caspers, M., van de Graaf, A., van Egmond, W., Keijser, B., & Roeselers, G. (2014). Shaping the oral microbiota through intimate kissing. *Microbiome*, 2, 41.

- Kwok, L.-Y., Zhang, J., Guo, Z., Gesudu, Q., Zheng, Y., Qiao, J., Huo, D., & Zhang, H. (2014). Characterization of fecal microbiota across seven Chinese ethnic groups by quantitative polymerase chain reaction. *PLoS One*, *9*(4), e93631.
- Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., & Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews. Genetics* *11*(10), 733–739. <https://doi.org/10.1038/nrg2825>
- Ley, R. E., Bäckhed, F., Turnbaugh, P., Lozupone, C. A., Knight, R. D., & Gordon, J. I. (2005). Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(31), 11070–11075.
- Li, J., Quinque, D., Horz, H.-P., Li, M., Rzhetskaya, M., Raff, J. A., Hayes, M. G., & Stoneking, M. (2014). Comparative analysis of the human saliva microbiome from different climate zones: Alaska, Germany, and Africa. *BMC Microbiology*, *14*(1), 316.
- Lin, H., & Peddada, S. D. (2020). Analysis of compositions of microbiomes with bias correction. *Nature Communications*, *11*(1), 3514.
- Ling, W., Lu, J., Zhao, N., Lulla, A., Plantinga, A. M., Fu, W., Zhang, A., Liu, H., Song, H., Li, Z., Chen, J., Randolph, T. W., Koay, W. L. A., White, J. R., Launer, L. J., Fodor, A. A., Meyer, K. A., & Wu, M. C. (2022). Batch effects removal for microbiome data via conditional quantile regression. *Nature Communications*, *13*(1), 5418.

- Lloyd-Price, J., Arze, C., Ananthakrishnan, A. N., Schirmer, M., Avila-Pacheco, J., Poon, T. W., Andrews, E., Ajami, N. J., Bonham, K. S., Brislawn, C. J., Casero, D., Courtney, H., Gonzalez, A., Graeber, T. G., Hall, A. B., Lake, K., Landers, C. J., Mallick, H., Plichta, D. R., ... Huttenhower, C. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, *569*(7758), 655–662.
- 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.
- Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K., & Knight, R. (2012). Diversity, stability and resilience of the human gut microbiota. *Nature*, *489*(7415), 220–230.
- Lu, J. J., Perng, C. L., Lee, S. Y., & Wan, C. C. (2000). Use of PCR with universal primers and restriction endonuclease digestions for detection and identification of common bacterial pathogens in cerebrospinal fluid. *Journal of Clinical Microbiology*, *38*(6), 2076–2080.
- Ma, S., Shungin, D., Mallick, H., Schirmer, M., Nguyen, L. H., Kolde, R., Franzosa, E., Vlamakis, H., Xavier, R., & Huttenhower, C. (2022). Population structure discovery in meta-analyzed microbial communities and inflammatory bowel disease using MMUPHin. *Genome Biology*, *23*(1), 208.
- Ma, Z. S., & Li, W. (2019). How and why men and women differ in their microbiomes: Medical ecology and network analyses of the microgenderome. *Advanced Science*, *6*(23), 1902054.

- Marsh, P. D. (2006). Dental plaque as a biofilm and a microbial community – implications for health and disease. *BMC Oral Health*, 6(1), S14.
- Martin, C. R., Osadchiy, V., Kalani, A., & Mayer, E. A. (2018). The brain-gut-microbiome axis. *Cellular and Molecular Gastroenterology and Hepatology*, 6(2), 133–148.
- Martino, C., Morton, J. T., Marotz, C. A., Thompson, L. R., Tripathi, A., Knight, R., & Zengler, K. (2019). A novel sparse compositional technique reveals microbial perturbations. *MSystems*, 4(1). <https://doi.org/10.1128/mSystems.00016-19>
- McCabe, K. M., Zhang, Y. H., Huang, B. L., Wagar, E. A., & McCabe, E. R. (1999). Bacterial species identification after DNA amplification with a universal primer pair. *Molecular Genetics and Metabolism*, 66(3), 205–211.
- Melito, P. L., Munro, C., Chipman, P. R., Woodward, D. L., Booth, T. F., & Rodgers, F. G. (2001). *Helicobacter winghamensis* sp. nov., a novel *Helicobacter* sp. isolated from patients with gastroenteritis. *Journal of Clinical Microbiology*, 39(7), 2412–2417.
- Miller, E. A., Beasley, D. E., Dunn, R. R., & Archie, E. A. (2016). Lactobacilli dominance and vaginal pH: Why is the human vaginal microbiome unique? *Frontiers in Microbiology*, 7, 1936.
- Mueller, N. T., Bakacs, E., Combellick, J., Grigoryan, Z., & Dominguez-Bello, M. G. (2015). The infant microbiome development: mom matters. *Trends in Molecular Medicine*, 21(2), 109–117.
- Nearing, J. T., Douglas, G. M., Hayes, M. G., MacDonald, J., Desai, D. K., Allward, N., Jones, C. M. A., Wright, R. J., Dhanani, A. S., Comeau, A. M., & Langille, M. G.

- I. (2022). Microbiome differential abundance methods produce different results across 38 datasets. *Nature Communications*, *13*(1), 342.
- Nishida, A., Inoue, R., Inatomi, O., Bamba, S., Naito, Y., & Andoh, A. (2018). Gut microbiota in the pathogenesis of inflammatory bowel disease. *Clinical Journal of Gastroenterology*, *11*(1), 1–10.
- O’Mahony, L., McCarthy, J., Kelly, P., Hurley, G., Luo, F., Chen, K., O’Sullivan, G. C., Kiely, B., Collins, J. K., Shanahan, F., & Quigley, E. M. M. (2005). Lactobacillus and bifidobacterium in irritable bowel syndrome: Symptom responses and relationship to cytokine profiles. *Gastroenterology*, *128*(3), 541–551.
- Palarea-Albaladejo, J., & Martín-Fernández, J. A. (2015). zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems*, *143*, 85–96.
- Patangia, D. V., Anthony Ryan, C., Dempsey, E., Paul Ross, R., & Stanton, C. (2022). Impact of antibiotics on the human microbiome and consequences for host health. *MicrobiologyOpen*, *11*(1), e1260.
- Pimentel, M., & Lembo, A. (2020). Microbiome and its role in irritable bowel syndrome. *Digestive Diseases and Sciences*, *65*(3), 829–839.
- Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, *35*(9), 833–844.
- Rackaityte, E., Halkias, J., Fukui, E. M., Mendoza, V. F., Hayzelden, C., Crawford, E. D., Fujimura, K. E., Burt, T. D., & Lynch, S. V. (2020). Viable bacterial colonization is highly limited in the human intestine in utero. *Nature Medicine*, *26*(4), 599–607.

- 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.
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.
- Sakai, J., Imanaka, K., Kodana, M., Ohgane, K., Sekine, S., Yamamoto, K., Nishida, Y., Kawamura, T., Matsuoka, T., Maesaki, S., Oka, H., & Ohno, H. (2019). Infective endocarditis caused by *Capnocytophaga canimorsus*; a case report. *BMC Infectious Diseases*, *19*(1), 927.
- Shade, A., Jones, S. E., Caporaso, J. G., Handelsman, J., Knight, R., Fierer, N., & Gilbert, J. A. (2014). Conditionally rare taxa disproportionately contribute to temporal changes in microbial diversity. *MBio*, *5*(4), e01371-14.
- Silva, G. G. Z., Green, K. T., Dutilh, B. E., & Edwards, R. A. (2016). SUPER-FOCUS: a tool for agile functional analysis of shotgun metagenomic data. *Bioinformatics*, *32*(3), 354–361.
- Singh, R. K., Chang, H.-W., Yan, D., Lee, K. M., Ucmak, D., Wong, K., Abrouk, M., Farahnik, B., Nakamura, M., Zhu, T. H., Bhutani, T., & Liao, W. (2017). Influence of diet on the gut microbiome and implications for human health. *Journal of Translational Medicine*, *15*(1), 73.
- Steinbrink, J., Leavens, J., Kauffman, C. A., & Miceli, M. H. (2018). Manifestations and outcomes of nocardia infections: Comparison of immunocompromised and nonimmunocompromised adult patients. *Medicine*, *97*(40), e12436.

- Thaiss, C. A., Zmora, N., Levy, M., & Elinav, E. (2016). The microbiome and innate immunity. *Nature*, *535*(7610), 65–74.
- Tomova, A., Bukovsky, I., Rembert, E., Yonas, W., Alwarith, J., Barnard, N. D., & Kahleova, H. (2019). The Effects of Vegetarian and Vegan Diets on Gut Microbiota. *Frontiers in Nutrition*, *6*. <https://doi.org/10.3389/fnut.2019.00047>
- Tran, H. T. N., Ang, K. S., Chevrier, M., Zhang, X., Lee, N. Y. S., Goh, M., & Chen, J. (2020). A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biology*, *21*(1). <https://doi.org/10.1186/s13059-019-1850-9>
- Tribble, G. D., Angelov, N., Weltman, R., Wang, B.-Y., Eswaran, S. V., Gay, I. C., Parthasarathy, K., Dao, D.-H. V., Richardson, K. N., Ismail, N. M., Sharina, I. G., Hyde, E. R., Ajami, N. J., Petrosino, J. F., & Bryan, N. S. (2019). Frequency of Tongue Cleaning Impacts the Human Tongue Microbiome Composition and Enterosalivary Circulation of Nitrate. *Frontiers in Cellular and Infection Microbiology*, *9*, 39.
- Wang, H., Tang, J., Wu, M., Wang, X., & Zhang, T. (2022). Application of machine learning missing data imputation techniques in clinical decision making: taking the discharge assessment of patients with spontaneous supratentorial intracerebral hemorrhage as an example. *BMC Medical Informatics and Decision Making*, *22*(1), 13.

- Wang, W.-L., Xu, S.-Y., Ren, Z.-G., Tao, L., Jiang, J.-W., & Zheng, S.-S. (2015). Application of metagenomics in the human gut microbiome. *World Journal of Gastroenterology: WJG*, *21*(3), 803–814.
- Wang, Y., & Lêcao, K. A. (2020). Managing batch effects in microbiome data. *Briefings in Bioinformatics*, *21*(6), 1954–1970.
- Weisburg, W. G., Barns, S. M., Pelletier, D. A., & Lane, D. J. (1991). 16S ribosomal DNA amplification for phylogenetic study. *Journal of Bacteriology*, *173*(2), 697–703.
- Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E. R., & Knight, R. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, *5*(1). <https://doi.org/10.1186/s40168-017-0237-y>
- White, G. C., & Bennetts, R. E. (1996). Analysis of frequency count data using the negative binomial distribution. *Ecology*, *77*(8), 2549–2557.
- Wu, J., Peters, B. A., Dominianni, C., Zhang, Y., Pei, Z., Yang, L., Ma, Y., Purdue, M. P., Jacobs, E. J., Gapstur, S. M., Li, H., Alekseyenko, A. V., Hayes, R. B., & Ahn, J. (2016). Cigarette smoking and the oral microbiome in a large study of American adults. *The ISME Journal*, *10*(10), 2435–2446.
- Yang, Y., Cai, Q., Zheng, W., Steinwandell, 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). <https://doi.org/10.1080/20002297.2019.1650597>

- Yao, Y., Cai, X., Ye, Y., Wang, F., Chen, F., & Zheng, C. (2021). The role of Microbiota in infant health: From early life to adulthood. *Frontiers in Immunology*, *12*, 708472.
- Yatsunenکو, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R. N., Anokhin, A. P., Heath, A. C., Warner, B., Reeder, J., Kuczynski, J., Caporaso, J. G., Lozupone, C. A., Lauber, C., Clemente, J. C., Knights, D., ... Gordon, J. I. (2012). Human gut microbiome viewed across age and geography. *Nature*, *486*(7402), 222–227. <https://doi.org/10.1038/nature11053>
- Yu, J. C., Khodadadi, H., & Baban, B. (2019). Innate immunity and oral microbiome: a personalized, predictive, and preventive approach to the management of oral diseases. *The EPMA Journal*, *10*(1), 43–50.
- Yuan, X., Han, L., Qian, S., Xu, G., & Yan, H. (2019). Singular value decomposition based recommendation using imputed data. *Knowledge-Based Systems*, *163*, 485–494.
- Zaborin, A., Penalver Bernabe, B., Keskey, R., Sangwan, N., Hyoju, S., Gottel, N., Gilbert, J. A., Zaborina, O., & Alverdy, J. C. (2020). Spatial compartmentalization of the microbiome between the lumen and crypts is lost in the Murine cecum following the process of surgery, including overnight fasting and exposure to antibiotics. *MSystems*, *5*(3). <https://doi.org/10.1128/mSystems.00377-20>
- Zhang, X., Mallick, H., Tang, Z., Zhang, L., Cui, X., Benson, A. K., & Yi, N. (2017). Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinformatics*, *18*(1), 4.

Zhang, Y., Parmigiani, G., & Johnson, W. E. (2020). ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genomics and Bioinformatics*, 2(3), lqaa078.

Zhou, Y., Gao, H., Mihindukulasuriya, K. A., La Rosa, P. S., Wylie, K. M., Vishnivetskaya, T., Podar, M., Warner, B., Tarr, P. I., Nelson, D. E., Fortenberry, J. D., Holland, M. J., Burr, S. E., Shannon, W. D., Sodergren, E., & Weinstock, G. M. (2013). Biogeography of the ecosystems of the healthy human body. *Genome Biology*, 14(1), R1.

Chapter 7: VITA

