Fall 2022

Integrating Current Analyses of the Breast Cancer Microbiome

Sidra Sohail

LOYOLA UNIVERSITY CHICAGO

INTEGRATING CURRENT ANALYSES OF THE BREAST CANCER MICROBIOME

A THESIS SUBMITTED TO

THE FACULTY OF THE GRADUATE SCHOOL

IN CANDIDACY FOR THE DEGREE OF

MASTER OF SCIENCE

PROGRAM IN BIOINFORMATICS

BY

SIDRA SOHAIL

CHICAGO, IL

AUGUST 2022

ACKNOWLEDGMENTS

I would like to thank Dr. Michael Burns for giving me the opportunity to join the Burns Lab during my sophomore year at Loyola and introducing me to the field of Bioinformatics. Dr. Burns's constant support and guidance has deepened my understanding of metagenomics, microbiome, statistical, and phylogenetic analyses, and has shaped me to be a dedicated and passionate scientist. I would like to thank Dr. Catherine Putonti and Dr. Xiang Gao for being on my thesis committee, and for their guidance and expertise on microbiome and statistical analyses. I would also like to thank my family, especially my parents, for supporting me, my ambitions, and for encouraging and motivating me to pursue my passion for bioinformatics.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

16S rRNA	16S ribosomal Ribonucleic acid
RDP	Ribosomal Database Project
BLAST	Basic Local Alignment Search Tool
ASV	Amplicon Sequence Variant
OTU	Operational Taxonomic Unit
PCoA	Principal Coordinate Analysis
NAF	Nipple Aspirate Fluid
NS	Nipple Skin
PBS	Post-Betadine Skin

ABSTRACT

Breast cancer is the leading cause of cancer death for women in the US. Many cancer types have significant associations with their resident microbial communities - emerging evidence suggests that breast cancers also interact with the local tissue-associated microbiota. Studies have examined the relationship between breast cancer and its microbiome, however, the studies varied in their approaches used to evaluate these relationships. Microbiome research advances rapidly and analysis pipelines and databases are updated frequently. This dynamic environment makes inter-study comparisons and superficial evaluations challenging as no two studies are using the same standards for evaluation.

Researchers have observed the microbiota of tumor tissue, surrounding normal sites, and healthy breast tissue from non-cancer individuals (Hieken et al., 2016; Urbaniak et al., 2016; Xuan et al., 2014), but they have not been able to translate their findings into information that can be used for breast cancer treatment or detection nor address what affect studying different variable regions has in their analysis. Within the majority of these studies, comparisons of the tumor tissue with adjacent normal tissue has revealed differences. This project will integrate all available studies related to breast cancer and the mammary microbiome to 1 reassess the original findings in light of advances in this rapidly progressing field and 2 incorporate all the data available as a large meta-analysis to identify general trends and specific differences across patient cohorts and studies.

CHAPTER ONE

INTRODUCTION

Breast cancer is the leading cause of cancer death for women in the US (*American Cancer Society: About Breast Cancer*, n.d.). Many cancer types have significant associations with their resident microbial communities - emerging evidence suggests that breast cancers also interact with the local tissue-associated microbiota (Chan et al., 2016; Hieken et al., 2016; Urbaniak et al., 2016; Xuan et al., 2014). Studies have examined the relationship between breast cancer and its microbiome, however, the studies varied in their approaches used to evaluate these relationships. Microbiome research advances rapidly and analysis pipelines and databases are updated frequently. This dynamic environment makes inter-study comparisons and superficial evaluations challenging as no two studies are using the same standards for evaluation.

Microbiome research widely uses 16S rRNA analysis to taxonomically identify bacterial and archaeal strains (Bukin et al., 2019). The 16S approach is widely used in microbiome research and has many advantages over shotgun sequencing. The 16S approach is cost effective, bioinformatic pipelines are available for data analysis for 16S data, and archived data are available for reference (Ranjan et al., 2016). Shotgun sequencing is able to identify more bacterial genera than 16S (Durazzi et al., 2021); however taxonomic classification and reference databases in shotgun metagenomics are not well-studied (Ye et al., 2019). It is known that taxonomic classifiers in shotgun analyses report large numbers of low-abundance taxa that are

false positives which lowers the precision of those estimates (Ye et al., 2019). The 16S amplicon approach is well-studied and cost effective where data analysis can be performed through known bioinformatic pipelines and there are established and well-studied taxonomic reference databases available for taxonomic classification such as SILVA (Pruesse et al., 2007).

The 16S rRNA gene includes nine hypervariable regions which range from V1 to V9, and these have considerable sequence diversity across different bacterial communities (Chakravorty et al., 2007). However, each variable region exhibits different degrees of sequence diversity, so a single variable region cannot distinguish between all known bacteria (Chakravorty et al., 2007). A study analyzed breast tissue microbiota of 16 patients from Italy across eight variable regions from V2 to V9 where the V3 region accounted for 45% of the reads and the V4 region accounted for 21% of the reads (Costantini et al., 2018). However, the optimal variable region to target for the breast microbiome is still of debate and more research is required to further understand the breast microbiome. There are different taxonomic reference databases available to taxonomically identify bacterial and archaeal strains, such as the Greengenes (McDonald et al., 2012) and SILVA (Pruesse et al., 2007) reference databases. There are also differences between the Greengenes, SILVA, and RDP (Cole et al., 2005) taxonomic databases and of the three, SILVA is the largest 16S taxonomic database in size and Greengenes is the smallest with RDP in between (Balvočiūtė & Huson, 2017). When comparing the Greengenes and SILVA databases using different bioinformatic tools , the SILVA database performed better, predicted more accurate predictions of taxa in simulated datasets, and had higher sensitivity than the Greengenes database (Almeida et al., 2018).

Researchers have observed the microbiota of tumor tissue, surrounding normal sites, and healthy breast tissue from non-cancer individuals (Hieken et al., 2016; Urbaniak et al., 2016; Xuan et al., 2014), but they have not been able to translate their findings into information that can be used for breast cancer treatment or detection nor address what affect studying different variable regions has in their analysis. Within the majority of these studies, comparisons of the tumor tissue with adjacent normal tissue has revealed differences. The microbiota differ drastically with the malignant tissue showing an increased abundance of pro-inflammatory genera and a decrease in bacterial community diversity and bacterial load (Hieken et al., 2016; Xuan et al., 2014). The depleted bacterial diversity in the malignant tissue can be potentially explained by the hypoxic, inflammatory microenvironment of tumor tissue. There is a decrease in the bacterial load in advanced tumors, and also a reduction in the antibacterial response in the breast tumor tissue where more severe tumors have a lower abundance of innate immune receptors in breast tissue (Xuan et al., 2014). However, when evaluating the microbiota of other cancer types, the cancers harbor more diverse communities (Burns et al., 2015; Mira-Pascual et al., 2015).

In comparisons of healthy and malignant breast tissue (Hieken et al., 2016), *Proteobacteria* and *Firmicutes* show increased abundance in tumor tissue (Urbaniak et al., 2016; Xuan et al., 2014). **However, there is not a functional or clear mechanistic explanation of these differences nor any inkling of how this translates to potential treatment or improvements in diagnosis for breast cancer.** Also, each study uses different bioinformatic methods, data formats, and variable regions further blurring the applicability of the results of each study to our understanding of the disease generally. The available studies (Chan et al.,

2016; Hieken et al., 2016; Urbaniak et al., 2016) fail to address the question of what the microbial composition in the mammary microbiome entails for mammary health, and whether looking at the different variable regions in their 16S rRNA analysis has any major effects on their results. It is necessary to combine the findings from these studies with respect to the different variable regions and patient cohorts. Our study addresses these questions by conducting a meta-analysis of all available studies on breast tissue and the microbiome starting from the original raw sequencing data.

CHAPTER TWO

INDIVIDUAL STUDY ANALYSES

Introduction

The Microbiome of Aseptically Collected Human Breast Tissue in Benign and Malignant Disease study by Hieken et al. (2016) targets the V3-V5 variable region of their 16S rRNA data. Their data is comprised of 98 forward and 98 reverse fastq files with a patient sample size of 33 patients (Fig. 1A), of which there were 16 women with benign disease and 17 women with malignant disease. They analyzed their 16S rRNA paired-end data by implementing the IM-TORNADO bioinformatics pipeline (Jeraldo, 2020) using the Greengenes version 13.5 reference database (McDonald et al., 2012). They implemented two alpha diversity measures, two beta diversity measures, and performed differential abundance analysis to identify significant taxa. The tissues of interest are buccal swab, skin swab, breast tissue, and breast skin tissue, which is also referred to as skin tissue in the paper. This paper did not perform multiple correction and reported unadjusted p-values.

Characterization of the microbiome of nipple aspirate fluid of breast cancer survivors study by Chan et al. (2016) targets the V4 variable region of their 16S rRNA data. Their data is comprised of 137 forward and 137 reverse fastq files with a patient sample size of 48 patients (Fig. 1B), of which there were 23 healthy control women and 25 women with a history of cancer samples. They analyzed their 16S rRNA paired-end data by implementing the Schloss MiSeq standard operating procedure in mothur (Schloss et al., 2009) using the SILVA version 119

reference database (Pruesse et al., 2007). They implemented an alpha diversity measure, a beta diversity measure, and performed differential abundance analysis to identify significant taxa. There are healthy control and women with a history of breast cancer samples and tissues of interest are NAF, NS, and PBS in this paper. The women with a history of breast cancer are also referred to as cancer samples in this section. This paper did not apply false discovery correction when comparing the OTU relative abundances, but they did apply multiple test correction when performing functional prediction with PICRUST (Langille et al., 2013) and KEGG (Kanehisa & Goto, 2000). The OTU table was rarefied for calculating dissimilarity measures and PCoA to account for any bias from uneven sequencing depth.

The Microbiota of Breast Tissue and Its Association with Breast Cancer study by Urbaniak et al. (2016) targets the V6 variable region of their 16S rRNA data. Their data is comprised of 68 merged fastq files with a patient sample size of 71 patients (Fig. 1C), of which there were 13 benign, 45 cancerous, and 23 healthy samples. They analyzed their 16S rRNA paired-end data by clustering the reads into OTUs by 97% identity using Uclust of USEARCH version 7 (Edgar, 2010) where taxonomy for each OTU was assigned using the SILVA reference database (Pruesse et al., 2007), manually verified using the RDP Seqmatch Tool (*Sequence Match*, n.d.), and further verified by using BLAST (Altschul et al., 1990) against the Greengenes database (McDonald et al., 2012) where the highest percent identity and coverage hit was assigned as taxonomy. They measured a weighted UniFrac distance, implemented unsupervised K-means clustering of CLR-transformed data, and identified significantly increased abundant taxa through the ALDEx2 package in R (Gloor et al., 2022). The samples of interest are malignant, benign, and healthy samples from breast tissue in this paper, where the samples are

normal adjacent tissue from women with breast cancer which is either benign or malignant and tissue from healthy controls. This paper applied the Benjamini-Hochberg correction and reported the Benjamini-Hochberg corrected p-values of the Wilcoxon rank test when running ALDEx2.

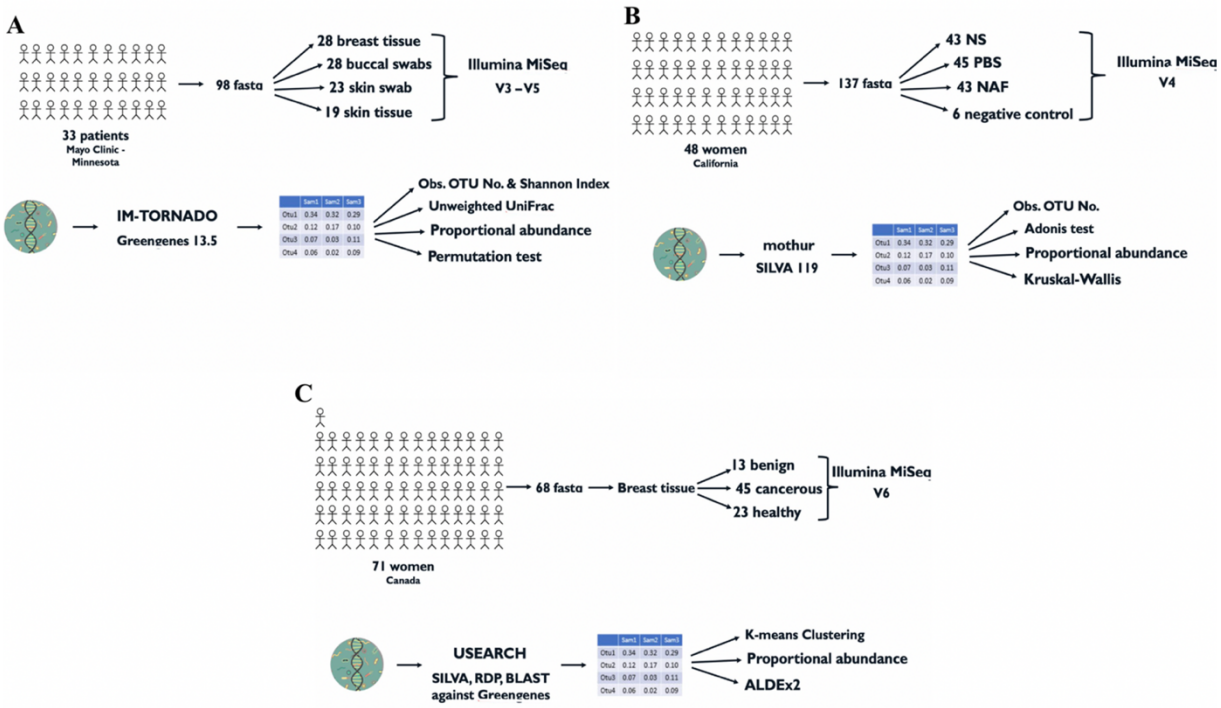

Figure 1. Overview of the studies of interest.
(A) Hieken study overview outlining their patient sample size, number of samples, sequencing technology and variable region, bioinformatics pipeline and taxonomic reference database, and downstream analyses. (B) Chan study overview outlining their patient sample size, number of samples, sequencing technology and variable region, bioinformatics pipeline and taxonomic reference database, and downstream analyses. (C) Urbaniak study overview outlining their patient sample size, number of samples, sequencing technology and variable region, bioinformatics pipeline and taxonomic reference database, and downstream analyses.

Methods

Hieken

Dataset overview. The study *The Microbiome of Aseptically Collected Human Breast Tissue in Benign and Malignant Disease* by Hieken et al. in August 2016, compares the microbial communities between malignant and benign tumors of patients with breast cancer, using normal adjacent tissues as controls. The Hieken et al. breast microbial data comprises 98 forward fastq files and 98 reverse fastq files for a total of 196 fastq files of which there are 98 samples across 33 distinct patients. The sequencing data for this study is available from the NCBI SRA (*Sequence Read Archive (SRA)*, n.d.) and was downloaded from the SRA with accession number PRJNA335375, using the SRA explore interface at sra-explorer.info (Ewels et al., n.d.). The samples were collected from 33 patients with either cancer or benign disease. In this study, the IM-TORNADO bioinformatics pipeline (Jeraldo, 2020) was implemented to conduct analyses on the paired-end data along with the Greengenes reference database version 13.5 (McDonald et al., 2012) for assigning taxonomy. Furthermore, there were two alpha diversity measures implemented, one was the observed OTU number which reflects the species richness and second the Shannon index which reflects the species evenness. There were two beta diversity measures implemented as well which were the unweighted and weighted UniFrac distances (Lozupone & Knight, 2005), and these diversity measures were calculated using an OTU table and phylogenetic tree. The unweighted UniFrac (Lozupone et al., 2011) measures differences in community presence such as whether or not there is an OTU present, and the weighted UniFrac measures differences in community abundance.

Preprocessing. Prior to beginning denoising, the primers used to amplify the V3-V5 regions of the 16S rRNA gene were trimmed from the unzipped paired end fastq files using cutadapt version 1.15 (Martin, 2011). The following shell commands were used to trim the primers 357F and 926R found in the Hieken study. The 357F primer used was AATGATACGGCGACCACCGAGATCTACACTATGGTAATTGTCCTACGGGAGGCAGC AG and the 926R primer used was CAAGCAGAAGACGGCATACGAGATGCCGCATTCGATXXXXXXXXXXXXCCGTCAAT TCMTTTRAGT.

After trimming the primers, FastQC version 0.11.5 (*Babraham Bioinformatics - FastQC A Quality Control Tool for High Throughput Sequence Data*, n.d.) was used to assess the Phred quality score of each forward and reverse fastq file as well as to check for potential artifactual sequences in the dataset.

DADA2 Pipeline Analysis. The trimmed fastq files were denoised by implementing the DADA2 version 1.16 pipeline (*DADA2 1.16 Pipeline*, n.d.) using the dada2 R package (B. Callahan et al., 2022). The trimmed reverse fastq files were of very low quality and few of them made it through the quality control steps within the pipeline; therefore, only forward fastq files were included in the analysis, as these reads met our quality control standards. The trimmed forward fastq files were analyzed using default parameters in the DADA2 1.16 pipeline (*DADA2 1.16 Pipeline*, n.d.), with changes in the filterAndTrim command, makeSequenceTable command, and the mergers command. In the filterAndTrim command, the reverse file input and truncLen parameter were removed and the maxEE value was adjusted from the default maxEE = c(2,2) to maxEE = 3. These changes were implemented to allow for flexibility and to relax the

stringency in our analysis so to prevent massive removal of sample reads which would make interpretation of the results impossible. Also, the multithread parameter was set to equal two to avoid memory issues. In the makeSequenceTable command, the dadaFs variable was used to make the seqtab variable instead of the default mergers variable. The mergers command was removed as the analysis was performed with only the forward fastq files as input. Additionally, the SILVA version 138 reference database (Pruesse et al., 2007) compatible with DADA2 (B. J. Callahan et al., 2016) was downloaded from the Zenodo site (McLaren & Callahan, 2021). The tax_glom and transform_sample_counts functions in phyloseq were implemented to make the proportional abundance plots. Additionally, the microshades package was used to make colorblind friendly proportional abundance plots and the code can be found at the microshades website (*A Custom Color Palette for Improving Data Visualization*, n.d.). In order to make the microshades plot, the phyloseq object is used as input to the prep_mdf function. After making the proportional abundance plots, phylogenetic analysis was performed and UniFrac distances (Lozupone & Knight, 2005) were calculated using GUniFrac (Chen et al., 2022).

Phylogenetic Analysis. The phylogenetic code was adapted from the compbiocore github site (*Computational Biology Core - Brown University*, n.d.), specifically commands from sequences<-getSequences(seqtab.nochim) up to dm <- dist.ml(phang.align) were implemented. The tree was generated using dm as input to the upgma command from the phangorn R package (Schliep et al., 2021). The ASV table and taxonomy table outputs from DADA2 along with the metadata and phylogenetic tree files were used as input to the MicrobiomeAnalyst web interface (Dhariwal et al., 2017) for visual exploratory data analysis.

Statistical Analysis. In order to calculate the UniFrac distances (Lozupone & Knight, 2005), the GUniFrac function from the GUniFrac R package (Chen et al., 2022) was implemented where the input was the ASV table output from the DADA2 analysis. The unweighted UniFrac (Lozupone et al., 2011) beta diversity ordination plots were generated through principal coordinate analysis using the cmdscale function from the vegan package (Jari Oksanen et al., 2022). Then, in order to identify differentially abundant taxa, the taxa with prevalence of less than 10% and relative abundance of less than 0.2% were filtered out. The prevalence filtering step was performed by filtering out ASVs that have more than 90% of their sample values as zeros. The relative abundance filtering step was performed by first modifying the ASV table so that it has proportional counts, and this was completed through total sum scaling of the ASV table. Then, ASVs that had a maximum proportional count of less than 0.002 (0.2%) were filtered out, and after this filtering step this taxa proportional data was square-root transformed for the permutation test. The ASV table was transposed so that the ASVs were the row names, and p values were calculated for the ASVs through the permutation test. The permutation test was performed using the linda function in the GUniFrac library (Chen et al., 2022), and the proportional, square-root transformed, and transposed ASV table was used as input. The significant ASVs were then graphed in the form of bar plots.

Urbaniak

Dataset overview. The study *The Microbiota of Breast Tissue and Its Association with Breast Cancer* by Urbaniak et al. in August 2016, compares the microbial composition between breast tissue from healthy control women and normal adjacent breast tissue from women with breast cancer, and they also compare normal adjacent breast tissue and breast tumor tissue. The

Urbaniak et al. breast microbial data comprises 68 merged fastq files across 71 women where 13 were benign tumors, 45 were cancerous tumors, and 23 were healthy control women. The sequencing data for this study is available from the NCBI SRA (*Sequence Read Archive (SRA)*, n.d.) and was downloaded from the SRA with accession number SRP076038, using the SRA explore interface at sra-explorer.info (Ewels et al., n.d.).

In this study, the Uclust algorithm within USEARCH version 7 (Edgar, 2010) was used to cluster reads by 97% identity into OTUs. Taxonomic assignment of each OTU was made by extracting best hits from SILVA database (Pruesse et al., 2007), manually verifying using Ribosomal Database Project (RDP) SeqMatch tool (Cole et al., 2005), and then using BLAST (Altschul et al., 1990) against the Greengenes database (McDonald et al., 2012), where the hits with the highest percent identity and coverage were used to assign taxonomy. OTU sequences were aligned using MUSCLE (Edgar, 2004) and were inputted to FASTTREE (Price et al., 2009) to build a tree of OTU sequences where PCoA plots of weighted UniFrac distances (Lozupone et al., 2011) were made in QIIME (Caporaso et al., 2010) using this tree of OTU sequences. Additionally, because microbiome data are compositional, k-means clustering was performed using euclidean distances of center-log ratio (CLR) transformed data. The ALDEx2 R package (Gloor et al., 2022) was used to measure the relative abundances of genera where Benjamini-Hochberg corrected p-value of the Wilcoxon rank test was used to test for significance. Furthermore, to test for differences between microbiota a microbiome regression-based kernel association test (mirkat) was performed in R using the MiRKAT package (Zhao et al., 2015) where a kernel metric was built using UniFrac distances and Bray-Curtis dissimilarity metric. UniFrac distances and Bray-Curtis dissimilarity metric were both used in the kernel metric

simultaneously as MiRKAT allows for multiple distance or dissimilarity metrics to be used at the same time.

Preprocessing. Upon downloading the fastq files from the SRA (*Sequence Read Archive (SRA)*, n.d.), the files were already merged and pre-processed.

DADA2 Pipeline Analysis. The trimmed fastq files were denoised by implementing the DADA2 version 1.16 pipeline (*DADA2 1.16 Pipeline*, n.d.) using the dada2 R package (B. Callahan et al., 2022). The merged fastq files were analyzed using default parameters in the DADA2 1.16 pipeline (*DADA2 1.16 Pipeline*, n.d.), with changes in the filterAndTrim command, makeSequenceTable command, and the mergers command. In the filterAndTrim command, the reverse file input and truncLen parameter were removed and the maxEE value was adjusted from the default maxEE = c(2,2) to maxEE = 3. These changes were implemented to allow for flexibility and to relax the stringency in our analysis so to prevent massive removal of sample reads which would make interpretation of the results impossible. In the makeSequenceTable command, the dadaFs variable was used to make the seqtab variable instead of the default mergers variable. The mergers command was removed as the analysis was performed with merged fastq files as input. Additionally, the SILVA version 138 reference database compatible with DADA2 was downloaded from the Zenodo site (McLaren & Callahan, 2021). The tax_glom and transform_sample_counts functions in phyloseq were implemented to make the proportional abundance plots. Additionally, the microshades package was used to make colorblind friendly proportional abundance plots and the code can be found at the microshades website (*A Custom Color Palette for Improving Data Visualization*, n.d.). In order to make the microshades plot, the phyloseq object is used as input to the prep_mdf function. After making

the proportional abundance plots, phylogenetic analysis was performed and UniFrac distances (Lozupone & Knight, 2005) were calculated using GUniFrac (Chen et al., 2022).

Phylogenetic Analysis. The phylogenetic code is adapted from the compbiocore github site (*Computational Biology Core - Brown University*, n.d.), specifically commands from sequences<-getSequences(seqtab.nochim) upto dm <- dist.ml(phang.align) are implemented. The tree is generated using dm as input to the upgma command from the phangorn R package (Schliep et al., 2021). The ASV table and taxonomy table outputs from DADA2 along with the metadata and phylogenetic tree files were used as input to the MicrobiomeAnalyst web interface (Dhariwal et al., 2017) for visual exploratory data analysis.

Statistical Analysis. In order to calculate the UniFrac distances (Lozupone & Knight, 2005), the GUniFrac function from the GUniFrac R package (Chen et al., 2022) was implemented where the input was the ASV table output from the DADA2 analysis. The unweighted and weighted UniFrac (Lozupone et al., 2011) beta diversity cluster plots were generated through the pam and clusplot functions from the cluster package (Rousseeuw et al., 2022). The Bray-Curtis beta diversity distance was calculated through the vegdist function from the vegan package (Jari Oksanen et al., 2022), and the beta diversity cluster plots were generated through the pam and clusplot functions from the cluster package (Rousseeuw et al., 2022). The pam function was used to perform K-Means Clustering and the clusplot function was used to visualize the results. The ALDEx2 R package (Gloor et al., 2022) was used to measure the relative abundances of statistically significant genera where ASVs in the ALDEx2 output were used to link associated genera to the output. Each genus in the output was visualized in the form of a box plot.

Chan

Dataset overview. The study *Characterization of the microbiome of nipple aspirate fluid of breast cancer survivors* by Chan et al. in June 2016, investigates the microbial community present in nipple aspirate fluid and their potential association with breast cancer by comparing NAF between healthy control samples and women with history of breast cancer samples. The Chan et al. microbial data comprises 137 forward fastq files and 137 reverse fastq files for a total of 274 fastq files of which there are 137 samples across 48 distinct patients. The sequencing data for this study is available from the NCBI SRA (*Sequence Read Archive (SRA)*, n.d.) and was downloaded from the SRA with accession number PRJNA314877, using the SRA explore interface at sra-explorer.info (Ewels et al., n.d.). The samples were collected from 48 patients with 23 healthy control women and 25 women with a history of breast cancer.

In this study, the Schloss MiSeq SOP (*MiSeq SOP*, n.d.) was implemented using mothur (Schloss et al., 2009) to conduct sequence processing on the paired-end data along with SILVA reference database version 119 (Pruesse et al., 2007) for assigning taxonomy. Additionally, OTUs detected in empty control Eppendorf tubes were removed to account for contaminating microbial 16S rDNA sequences from extraction kits and reagents. Furthermore, Bray-Curtis dissimilarity was implemented for PCoA using rarefied OTU abundances. Analysis of variance (Adonis) was used to measure community composition differences and a nonparametric Kruskal-Wallis test was used to test whether the OTU relative abundances were statistically significant between the healthy control and breast cancer samples for NAF, NS, and PBS samples. A paired Wilcoxon signed-rank test was implemented to compare the OTU abundances between NAF and NS samples from the same patients. Additionally, a paired two-tailed t-test was used to compare

the microbial composition between paired NAF and nipple skin samples. The statistical analyses used a p-value cutoff of 0.05 and false discovery rate correction was not implemented for OTU relative abundance comparison.

Preprocessing. Prior to beginning denoising, the primers used to amplify the V4 region of the 16S rRNA gene were trimmed from the unzipped paired end fastq files using cutadapt version 1.15 (Martin, 2011). The following shell commands were used to trim the primers F515 and R806. The F515 primer used was AATGATACGGCGACCACCGAGACGTACGTACGGTGTGCCAGCMGCCGCGGTAA and the R806 primer used was CAAGCAGAAGACGGCATACGAGATXXXXXXXXXXXXACGTACGTACCGGATACHVGGGTWTCTAAT. After trimming the primers, FastQC version 0.11.5 (*Babraham Bioinformatics - FastQC A Quality Control Tool for High Throughput Sequence Data*, n.d.) was used to assess the Phred quality score of each forward and reverse fastq file as well as to check for potential artifactual sequences in the dataset.

DADA2 Pipeline Analysis. The trimmed fastq files were denoised by implementing the DADA2 version 1.16 pipeline (*DADA2 1.16 Pipeline*, n.d.) using the dada2 R package (B. Callahan et al., 2022). The trimmed reverse fastq files were of very low quality and few of them made it through the quality control steps within the pipeline; therefore, only forward fastq files were included in the analysis, as these reads met our quality control standards. The trimmed forward fastq files were analyzed using default parameters in the DADA2 1.16 pipeline (*DADA2 1.16 Pipeline*, n.d.), with changes in the filterAndTrim command, makeSequenceTable command, and the mergers command. In the filterAndTrim command, the reverse file input and

truncLen parameter were removed and the maxEE value was adjusted from the default maxEE = c(2,2) to maxEE = 3. These changes were implemented to allow for flexibility and to relax the stringency in our analysis so as to prevent massive removal of sample reads which would make interpretation of the results impossible. Also, the multithread parameter was set to equal two to avoid memory issues. In the makeSequenceTable command, the dadaFs variable was used to make the seqtab variable instead of the default mergers variable. The mergers command was removed as the analysis was performed with only the forward fastq files as input. Additionally, the SILVA version 138 reference database (Pruesse et al., 2007) compatible with DADA2 was downloaded from the Zenodo site (McLaren & Callahan, 2021). The tax_glom and transform_sample_counts functions in phyloseq were implemented to make the proportional abundance plots. Additionally, the microshades package was used to make colorblind friendly proportional abundance plots and the code can be found at the microshades website (*A Custom Color Palette for Improving Data Visualization*, n.d.). In order to make the microshades plot, the phyloseq object was used as input to the prep_mdf function. After making the proportional abundance plots, phylogenetic analysis was performed and UniFrac distances (Lozupone & Knight, 2005) were calculated.

Phylogenetic Analysis. The phylogenetic code is adapted from the compbiocore github site (*Computational Biology Core - Brown University*, n.d.), specifically commands from sequences <- getSequences(seqtab.nochim) upto dm <- dist.ml(phang.align) are implemented. The tree is generated using dm as input to the upgma command from the phangorn R package (Schliep et al., 2021). The ASV table and taxonomy table outputs from DADA2 along with the

metadata and phylogenetic tree files were used as input to the MicrobiomeAnalyst web interface (Dhariwal et al., 2017) for visual exploratory data analysis.

Statistical Analysis. The Adonis test was run to measure community composition differences and a nonparametric Kruskal-Wallis test was run to test whether the OTU relative abundances were statistically significant between the healthy control and breast cancer samples for NAF, NS, and PBS samples. The p-values reported for the Adonis and Kruskal-Wallis tests are unadjusted p-values as the Chan study also reported unadjusted p-values. In order to perform the Adonis test, the distance function in phyloseq (McMurdie & Holmes, 2013) was run to calculate Bray-Curtis distance with a proportionally rarefied phyloseq object as input. The calculated distance and its associated sample metadata were used as input to the adonis function from the vegan package (Jari Oksanen et al., 2022). The Kruskal-Wallis test was then run for each ASV from a rarefied phyloseq object, and ASVs that are not significant are filtered out. The significant ASVs were then linked with their associated taxa and the abundance values for each taxon was visualized through a dot plot.

Results

Hieken et al. Original Study Results

The figures reported in this section are from the original Hieken et al. paper, unless specified otherwise.

Alpha Diversity. In the alpha diversity analysis, they used observed OTU number which shows species richness and the Shannon index which shows species evenness. When comparing breast and skin tissue, the observed OTU number revealed that there were more bacterial species richness in breast tissue than skin tissue (Fig. 2A), whereas the species evenness through the

Shannon Index were similar between the breast and skin tissue (Fig. 2B). To further explore the increased species richness of the breast tissue, a heat map visualizes species richness between the skin and breast tissue where more OTUs are observed for the breast tissue and these OTUs in breast tissue are mostly of low abundance (Fig. 2C).

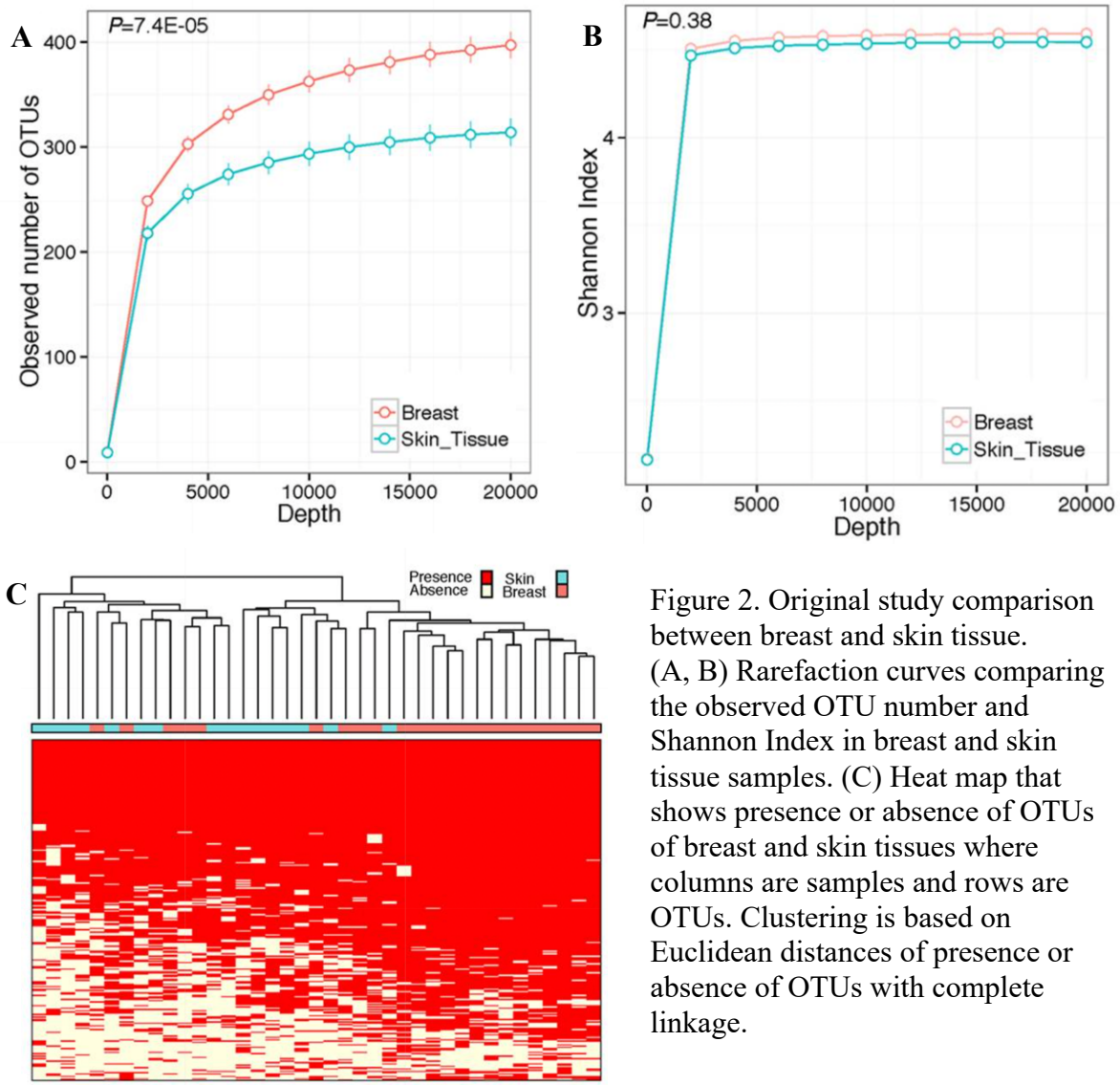

Figure 2. Original study comparison between breast and skin tissue. (A, B) Rarefaction curves comparing the observed OTU number and Shannon Index in breast and skin tissue samples. (C) Heat map that shows presence or absence of OTUs of breast and skin tissues where columns are samples and rows are OTUs. Clustering is based on Euclidean distances of presence or absence of OTUs with complete linkage.

Beta Diversity. In the beta diversity analysis, they used weighted and unweighted UniFrac distances which were calculated through the GUniFrac package in R (Chen et al., 2022) with the OTU table and phylogenetic tree as input. Prior to alpha and beta diversity analyses, the OTU table was rarefied to a sequencing depth of 20,000 sequences to reduce confounding effects. They performed MiRKAT (Zhao et al., 2015) and calculated p-values to identify the association between the two beta diversity measures. Beta diversity analysis was performed to compare breast and skin tissue microbiota, where the unweighted UniFrac distance showed a significant difference in microbial community between breast and skin tissue (Fig. 3A) and reported a MiRKAT p-value of 0.0001. However, the weighted UniFrac distance did not show a significant difference between breast and skin tissue (Fig. 3B) and reported a MiRKAT p-value of 0.14. To ascertain that the significant difference was not due to differential sequencing depth between the breast and skin tissue, they performed rarefaction and adjusted the sequencing depth

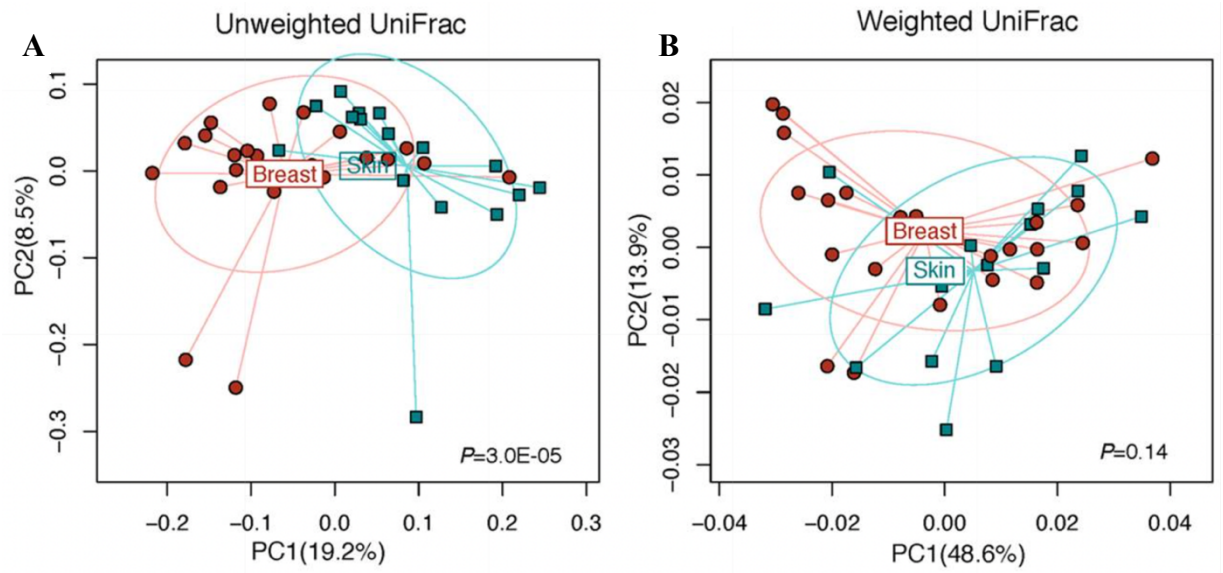

Figure 3. Original study comparison of microbiota of breast and skin tissue.
(A) Unweighted UniFrac ordination plot showing clustering of the two tissues.
(B) Weighted UniFrac ordination plot showing clustering of the two tissues.

in the model. Rarefaction and adjustment to sequencing depth did not change the statistical significance found in the alpha and beta diversity results.

Also, they compared the microbial community between breast tissue adjacent to invasive cancer disease and breast tissue adjacent to benign disease using the unweighted and weighted UniFrac distances and reported the associated p-values from MiRKAT. The unweighted UniFrac distance showed that the microbial community differed significantly between breast tissue adjacent to cancer disease state and breast tissue adjacent to benign disease state (Fig. 4) with a MiRKAT p-value of 0.009. However, the weighted UniFrac distance showed that the difference is not significant in the microbial community between breast tissue adjacent to cancer disease state and breast tissue adjacent to benign disease state. They state that the differences found in

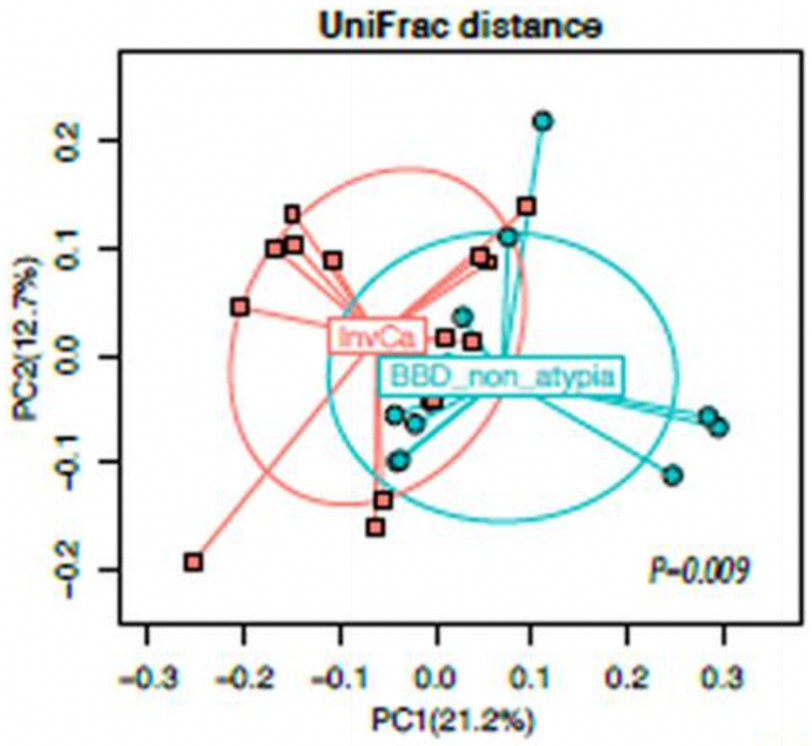

Figure 4. Original study ordination plot of unweighted UniFrac distance metric. Unweighted UniFrac distance ordination plot showing the clustering of breast tissue microbiota between BBD_non_atypia and InvCa samples.

the unweighted UniFrac analysis are mainly in the rare or less abundant lineages. Additionally, they perform PCoA on the unweighted UniFrac distances and showed that the microbiome of the different tissue types – buccal swab, skin swab, breast tissue, and skin tissue – cluster distinctly from one another. The buccal swab and skin swab microbiota clearly separate from each other and the other two tissues, and the skin and breast tissue microbiota are closer in space but are also separate from each other.

Proportional Abundance Analysis. In the proportional abundance analysis, they assessed the taxonomic composition of breast and skin tissue microbiota at phylum, family, and genus levels. They report that these taxonomic plots show similar abundance in the major taxa between the breast and skin tissue (Fig. 7A) and the associated taxa are shown in the supplementary material. Additionally, they assess the taxonomic composition of breast tissue microbiota between benign and malignant disease states (Fig. 8A) and of buccal and skin swab microbiota at the phylum, family, and genus levels (Fig. 9A). These plots are located in the supplementary materials. The plots in the supplementary material are provided here for reference.

Differential Abundance Analysis. In the differential abundance analysis, they filtered out taxa with prevalence of less than 10% and relative abundance of less than 0.2%. In order to identify the differentially abundant taxa, they implemented a permutation test where they fitted a regular linear model and taxa proportion data was the outcome variable. To assess statistical significance, they used 1,000 permutations and F-stat as the test statistic and reported unadjusted p-values. Based on this permutation test, they identified differentially abundant taxa between the breast and skin tissue from the following phyla groups Firmicutes, Actinobacteria, Bacteroidetes,

and Proteobacteria. The low-abundant differential taxa were also identified from the permutation test with an unadjusted p-value less than 0.05 where they used a permutation test to assess the differential taxa in breast tissue microbiota between benign and malignant disease states. Through this permutation test, they found that there is an increased relative abundance of low-abundant genera in the invasive cancer breast tissue. The low-abundant genera were *Fusobacterium, Atopobium, Hydrogenophaga, Gluconacetobacter*, and *Lactobacillus* with an unadjusted p-value less than 0.05. Barplots further confirm the abundances of the five differential genera between the benign and malignant disease states in breast tissue (Fig. 5).

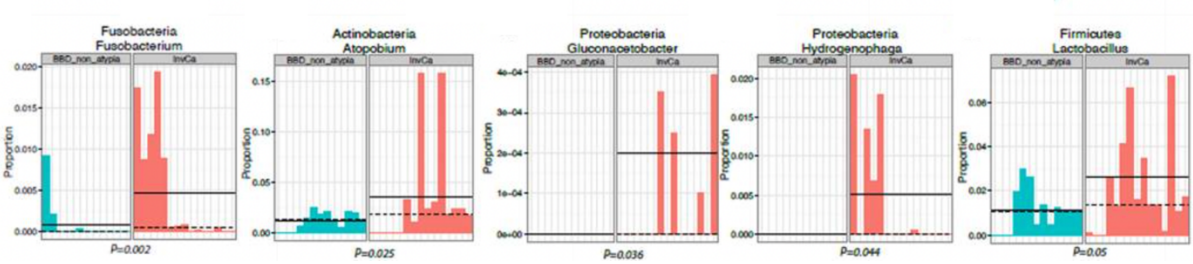

Figure 5. Original Study Differentially abundant genera present in breast tissue microbiota. These genera are abundant in InvCa in breast tissue and are low-abundant genera. P-values are unadjusted.

Hieken et al. Re-analysis Results

The figures reported in this section are from our results of the re-analysis of the Hieken et al. paper.

Alpha Diversity. In the alpha diversity analysis, we have used the observed OTU number and Shannon Index to look at breast and skin tissue microbiota. The observed OTU number shows difference between the two tissues with an increased abundance of breast tissue OTUs than skin tissue OTUs as shown through the observed OTU number scatter plot and box plot (Fig. 6A). However, the Shannon Index does not show this difference as shown through the

Shannon scatter plot and box plot (Fig. 6B), and the heatmap also does not show a difference between the breast and skin OTU abundance (Fig. 6C).

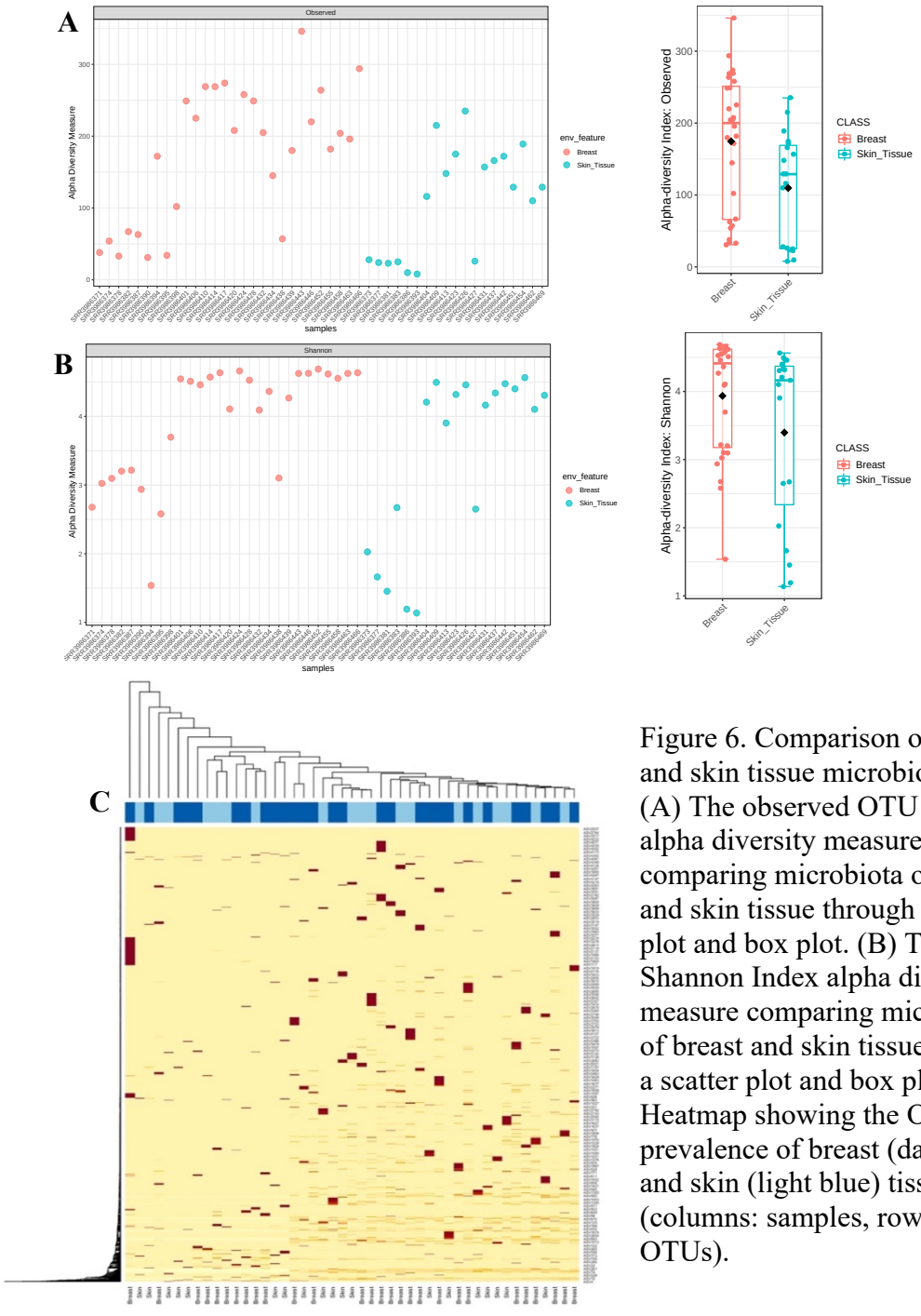

Figure 6. Comparison of breast and skin tissue microbiota. (A) The observed OTU number alpha diversity measure comparing microbiota of breast and skin tissue through a scatter plot and box plot. (B) The Shannon Index alpha diversity measure comparing microbiota of breast and skin tissue through a scatter plot and box plot. (C) Heatmap showing the OTU prevalence of breast (dark blue) and skin (light blue) tissues (columns: samples, rows: OTUs).

Beta Diversity. In the beta diversity analysis, we have used weighted and unweighted UniFrac distances which were calculated through the GUniFrac package in R (Chen et al., 2022) with the OTU table and phylogenetic tree as input. Prior to alpha and beta diversity analyses, the OTU table was rarefied to reduce confounding effects. Beta diversity analysis was performed to compare breast and skin tissue microbiota, where the unweighted UniFrac distance and weighted UniFrac distance did not show a significant difference in microbial community between breast and skin tissue and the MiRKAT p-values for both unweighted and weighted UniFrac were not significant ($p > 0.05$) (Fig. 7A; Fig. 7B). The microbial community between breast tissue adjacent to invasive cancer disease and breast tissue adjacent to benign disease were also compared using the unweighted and weighted UniFrac distances (Fig. 7C). The unweighted UniFrac distance and weighted UniFrac distance did not show a significant difference in the microbial community between breast tissue adjacent to cancer disease state and breast tissue adjacent to benign disease state and the MiRKAT p-values for both unweighted and weighted UniFrac were not significant ($p > 0.05$). Additionally, PCoA was performed on the unweighted UniFrac distances and showed that the microbiome of the different tissue types – buccal swab, skin swab, breast tissue, and skin tissue – cluster separately from one another (Fig. 7D). The buccal swab and skin swab microbiota clearly separate from each other and the other two tissues, and the skin and breast tissue microbiota are closer in space but also cluster separate from each other.

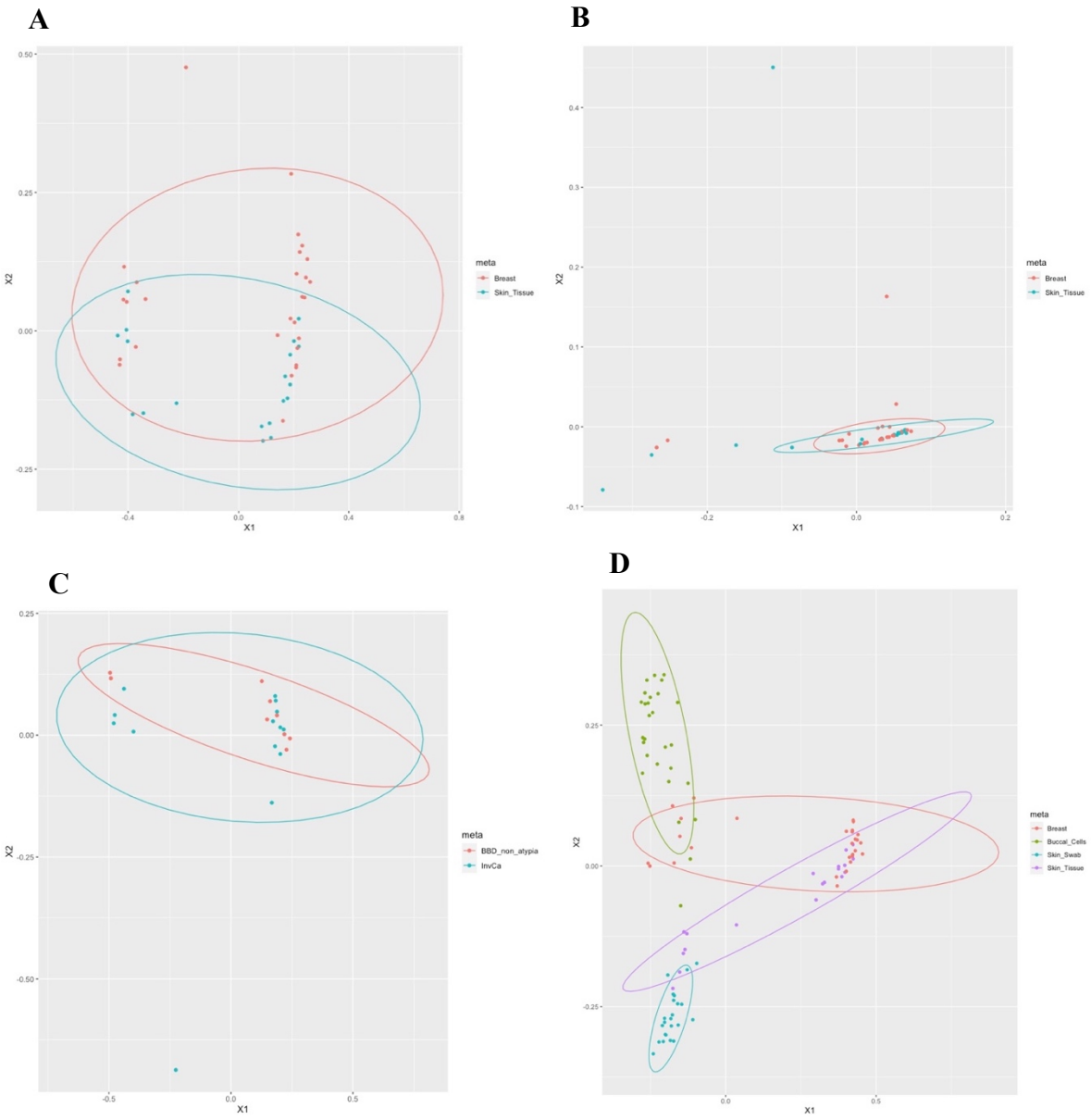

Figure 7. Comparing microbiota of tissues and types of cancers.
(A) Unweighted UniFrac distance plot showing clustering of breast (red) and skin (blue) tissues. (B) Weighted UniFrac distance plot showing clustering of breast (red) and skin (blue) tissues. (C) Unweighted UniFrac distance plot showing clustering of breast tissue microbiota between benign (BBD_non_atypia) and cancer (InvCa) states. (D) PCoA plot of unweighted UniFrac distance showing clustering of all 4 tissues where the color scheme is breast tissue (red), skin tissue (purple), skin swab (blue), and buccal swab/cells (green).

Proportional Abundance Analysis. In the proportional abundance analysis, we have assessed the taxonomic composition of breast and skin tissue microbiota, breast tissue microbiota in benign, also known as BBD_non_atypia, and invasive cancer, also known as InvCa, disease states, and buccal and skin swab microbiota at phylum, family, and genus levels. The proportional abundance plots include the top 100 sequences in order to clearly show the taxa present at the phylum, family, and genus levels. The breast and skin tissue show similar abundances of major taxa from phyla Actinobacteriota, Bacteroidota, Firmicutes, and Proteobacteria; however, the skin tissue also shows taxa from the phylum Fusobacteriota (Fig. 8B). The breast tissue microbiota in benign and invasive cancer disease states show similar abundances of taxa from phyla Actinobacteriota, Bacteroidota, Firmicutes, and Proteobacteria; however, the invasive cancer disease state shows a low abundance of phylum Spirochaetota (Fig. 9B). The buccal and skin swab show similar abundances of taxa from phyla Actinobacteriota, Bacteroidota, Firmicutes, and Proteobacteria with Fusobacteriota found only in the skin swab (Fig. 10B). At the genus level, there are clearer differences in taxonomic composition between the buccal and skin swab microbiota. The buccal swab microbiota shows a greater abundance of

Veillonella and *Streptococcus,* whereas the skin swab microbiota shows a greater abundance of *Staphylococcus* and *Escherichia-Shigella*.

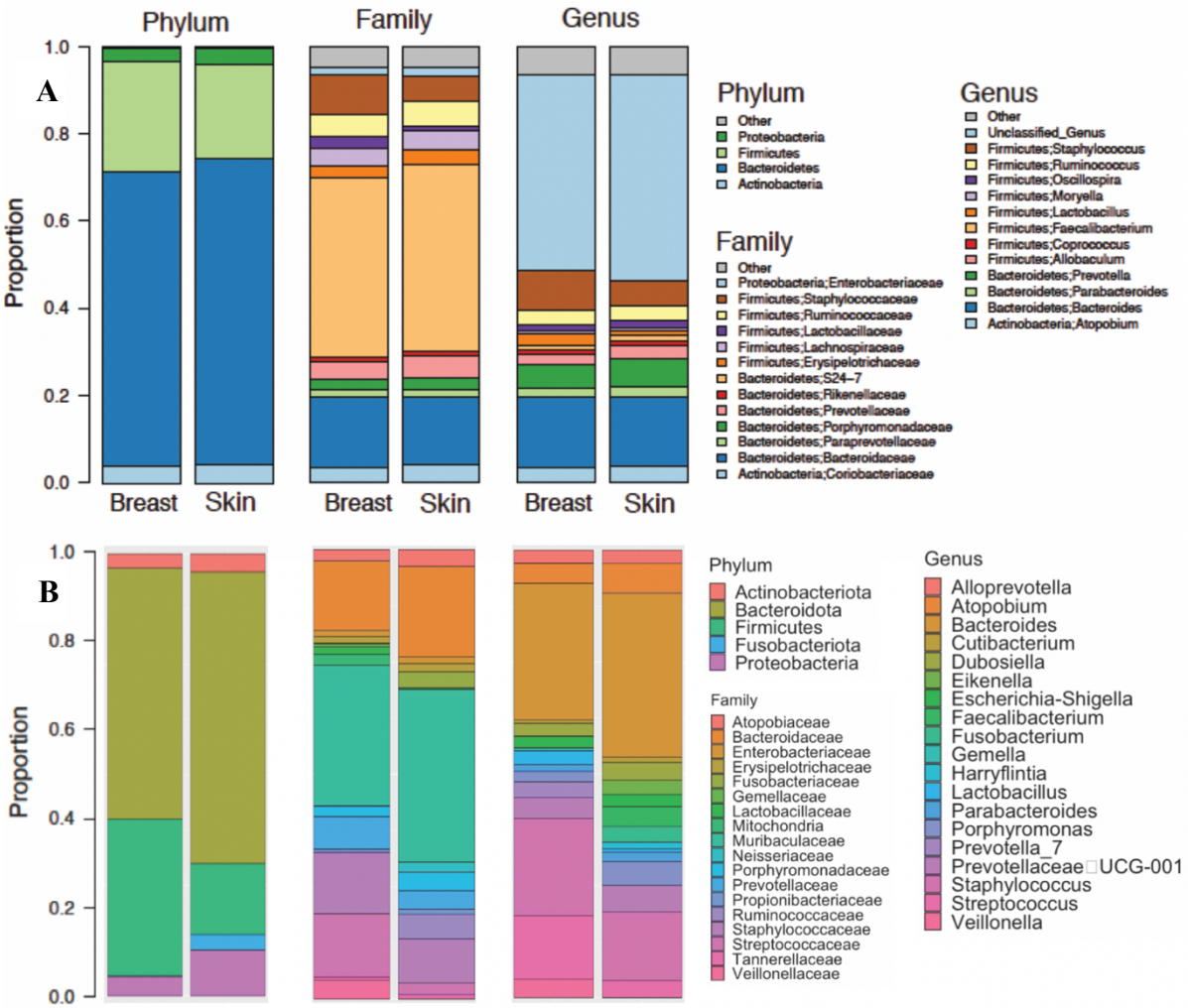

Figure 8. Proportional abundance barplots of taxonomic profiles of breast and skin tissue microbiota.
(A) Proportional abundance barplots from the original Hieken study showing taxonomic composition of breast and skin tissue microbiota at the phylum, family, and genus levels.
(B) Proportional abundance barplots generated from our DADA2 analysis of the Hieken data showing taxonomic composition of breast and skin tissue microbiota at the phylum, family, and genus levels.

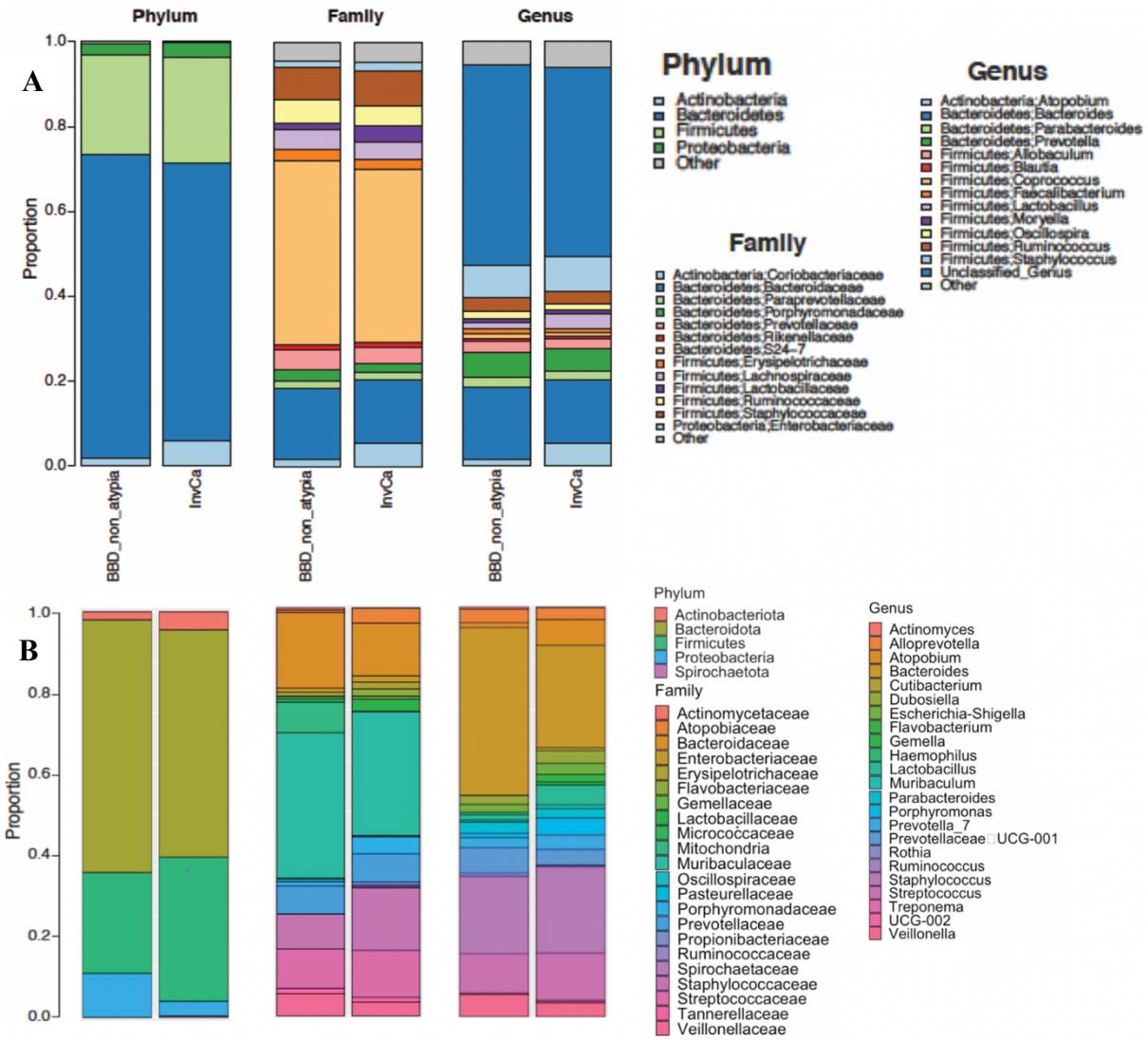

Figure 9. Proportional abundance barplots of taxonomic profiles of breast tissue microbiota between BBD_non_atypia and InvCa samples.
(A) Proportional abundance barplots from the original Hieken study showing taxonomic composition of BBD_non_atypia and InvCa samples from breast tissue at the phylum, family, and genus levels. (B) Proportional abundance barplots generated from our DADA2 analysis of the Hieken data showing taxonomic composition of BBD_non_atypia and InvCa samples from breast tissue at the phylum, family, and genus levels.

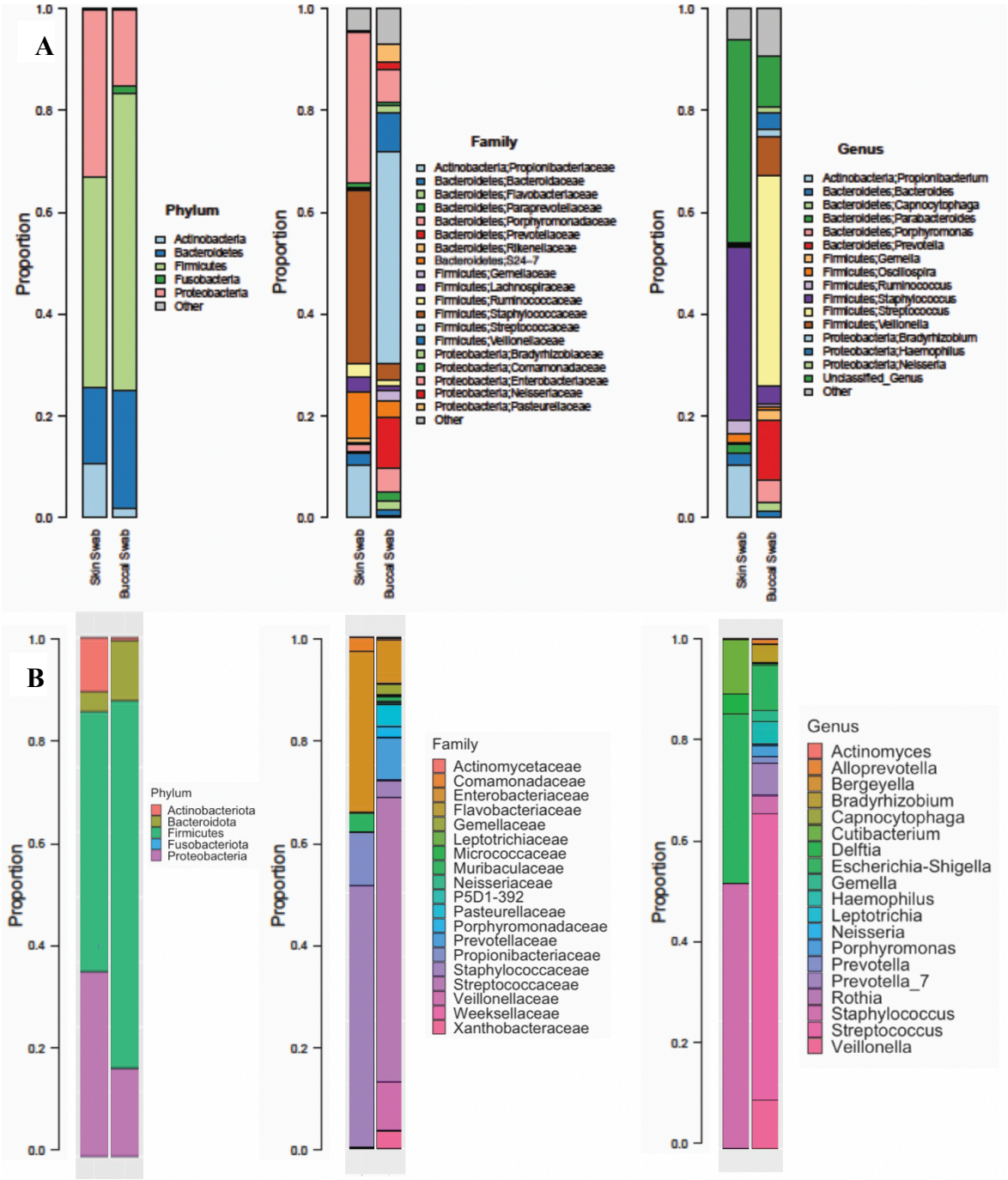

Figure 10. Proportional abundance barplots of taxonomic profiles of buccal swab and skin swab microbiota.
(A) Proportional abundance barplots from the original Hieken study showing taxonomic composition of buccal swab and skin swab microbiota at the phylum, family, and genus levels. (B) Proportional abundance barplots generated from our DADA2 analysis of the Hieken data showing taxonomic composition of buccal swab and skin swab microbiota at the phylum, family, and genus levels.

Differential Abundance Analysis. In the differential abundance analysis, taxa with prevalence of less than 10% and relative abundance of less than 0.2% were filtered out. In order to identify the differentially abundant taxa, we implemented a linear (lin) model for differential abundance (da) called linda which fits linear regression models on high dimensional data (Zhou et al., 2022) and the linda tool is in the MicrobiomeStat package in R (Zhang et al., 2022). Based on this permutation test, there were twelve significant differentially abundant taxa identified in breast tissue microbiota in benign and invasive cancer disease states. These twelve significant differential taxa were *Eubacterium xylanophilum group, ASF356, Atopobium, Bacteroides, Clostridia UCG-014, Colidextribacter, Escherichia-Shigella, Lachnospiraceae NK4A136 group, Rikenellaceae RC9 gut group, Staphylococcus, Subdoligranulum,* and *Veillonella* where the reported p-values were unadjusted for false discovery correction. The barplots further confirm the abundances of the twelve differential taxa between the benign and malignant disease states in breast tissue (Fig. 11).

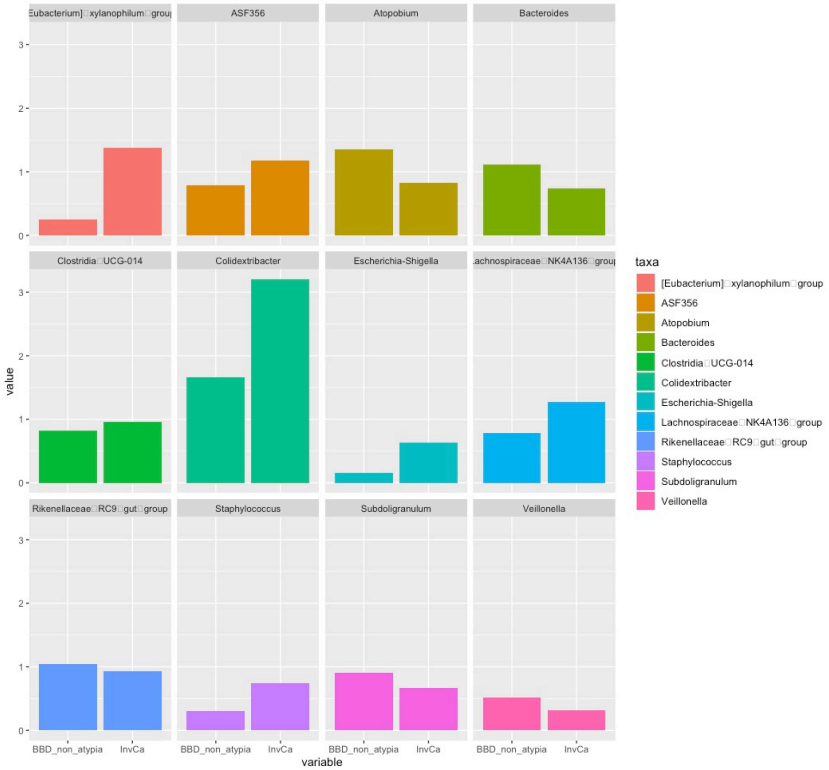

Figure 11. Differentially abundant taxa in breast tissue microbiota of benign and malignant disease states. Differential taxa in breast tissue microbiota of benign (BBD_non_atypia) and malignant disease states based on the linda model.

Chan et al. Original Study Results

The figures reported in this section are from the original Chan et al. paper.

Alpha Diversity. In the alpha diversity analysis, they used observed OTU numbers with ten permutations of random sampling at each sequencing depth. They compared the healthy control and cancer samples from NS, NAF, and PBS samples, and performed non-parametric t-tests. When comparing the healthy control and cancer samples from the NS microbiome, the observed OTUs as a function of sequencing depth were assessed by diversity rarefaction curves and the difference in diversity assessed by the non-parametric t-test was not significant with an unadjusted p-value of 0.929 (Fig. 12A). When comparing the healthy control and cancer samples from the NAF microbiota, the observed OTUs as a function of sequencing depth were assessed by diversity rarefaction curves and the difference in diversity assessed by the non-parametric t-test was not significant with an unadjusted p-value of 0.65 (Fig. 12B). When comparing the healthy control and cancer samples from the PBS microbiota, the observed OTUs as a function of sequencing depth were assessed by diversity rarefaction curves and the difference in diversity assessed by the non-parametric t-test was not significant with an unadjusted p-value of 0.151 (Fig. 12C).

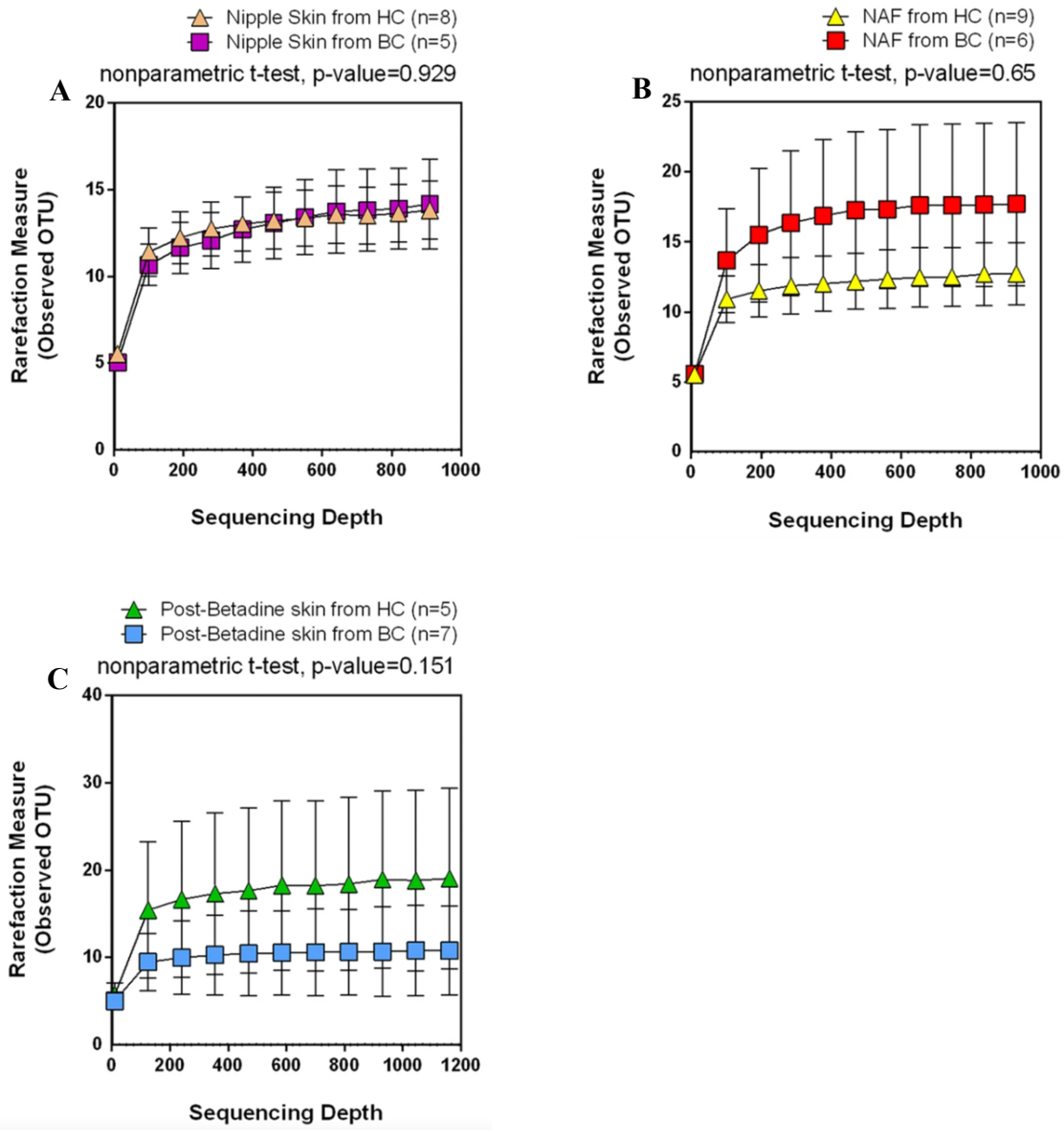

Figure 12. Original Study Alpha Diversity: Observed OTU number.
Alpha diversity rarefaction curves where number of OTUs observed is on x-axis and sequencing depth is on y-axis. (A) NS microbiota composition. (B) NAF microbiota composition. (C) PBS microbiota composition.

Beta Diversity. In the beta diversity analysis, they implemented a Bray-Curtis dissimilarity metric and performed PCoA using the rarefied OTU abundances as input where the genus-level OTUs were used in PCoA. The Adonis test was used to test for compositional differences, and the adonis function in the vegan package (Oksanen et al., 2022) from R was used to implement this test. The healthy control and cancer samples from the NS microbiota did not form separate clusters and when running the Adonis test there were no significant differences found in the bacterial composition with an unadjusted p-value of 0.945 (Fig. 13B). The healthy control and cancer samples from the NAF microbiota formed separate clusters and when running the Adonis test there was a significant difference in the bacterial composition with an unadjusted p-value of 0.002 (Fig. 13A). They deduced that the Adonis test indicates that having had a history of breast cancer significantly affects the NAF microbiota and explains 13.5% of variability between the samples. The healthy control and cancer samples from the PBS microbiota did not form separate clusters and when running the Adonis test there were no significant differences found in the bacterial composition with an unadjusted p-value of 0.478 (Fig. 13C).

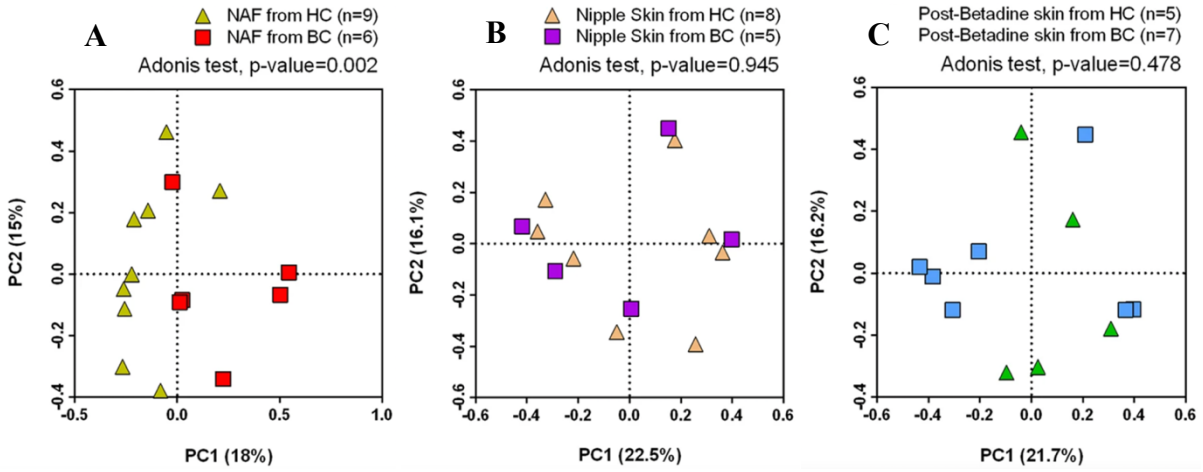

Figure 13. Original Study Adonis test: compositional differences.
(A) NAF microbiota. (B) NS microbiota. (C) PBS microbiota.

Proportional Abundance Analysis. In the proportional abundance analysis, they assessed the taxonomic composition of NS, NAF, and PBS microbiota at the phylum level. The NS microbial composition was predominantly comprised of the phyla Proteobacteria, Firmicutes, and Bacteroidetes (Fig. 14A). At the genus level, they reported that *Alistipes* was the most abundant OTU in the nipple skin samples. The PBS microbial composition was predominantly comprised of the phyla Proteobacteria, Firmicutes, and Bacteroidetes (Fig. 14B). The NAF microbial composition was predominantly comprised of the phyla Firmicutes, Proteobacteria, and Bacteroidetes (Fig. 14C).

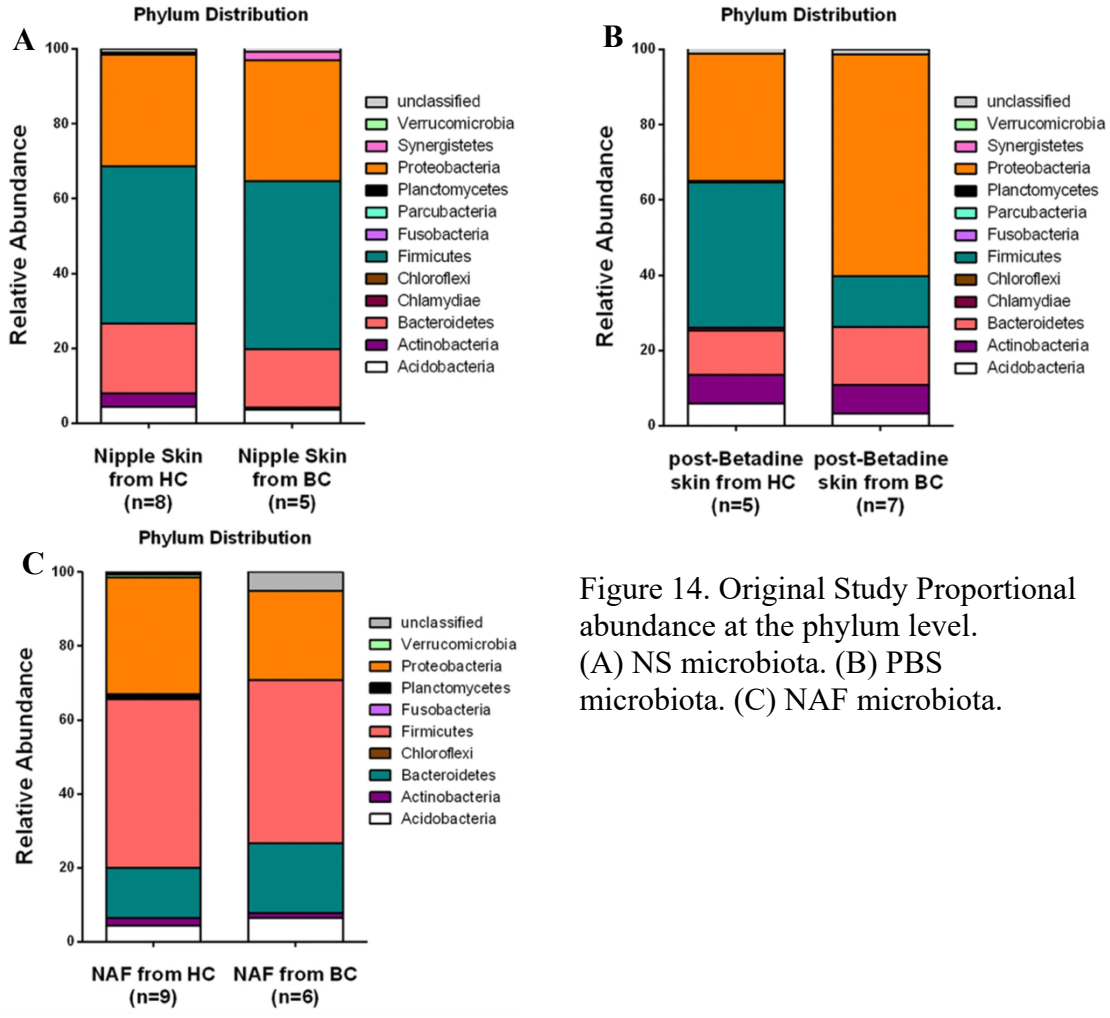

Figure 14. Original Study Proportional abundance at the phylum level. (A) NS microbiota. (B) PBS microbiota. (C) NAF microbiota.

Differential Abundance Analysis. In the differential abundance analysis, the Kruskal-Wallis test (Kruskal and Wallis, 1952) was performed on the healthy and cancer samples from NS, NAF, and PBS OTUs. The NS OTUs were not significantly different when comparing the healthy and cancer samples through the Kruskal-Wallis test. There were two NAF OTUs identified to be significantly different in relative abundance when comparing the healthy and cancer samples through the Kruskal-Wallis test, and they were *Alistipes* at the genus level present in only cancer NAF samples and *Sphingomonadaceae* at the family level present in both healthy and cancer NAF samples (Fig. 15). The PBS OTUs were not significantly different when comparing the healthy and cancer samples through the Kruskal-Wallis test.

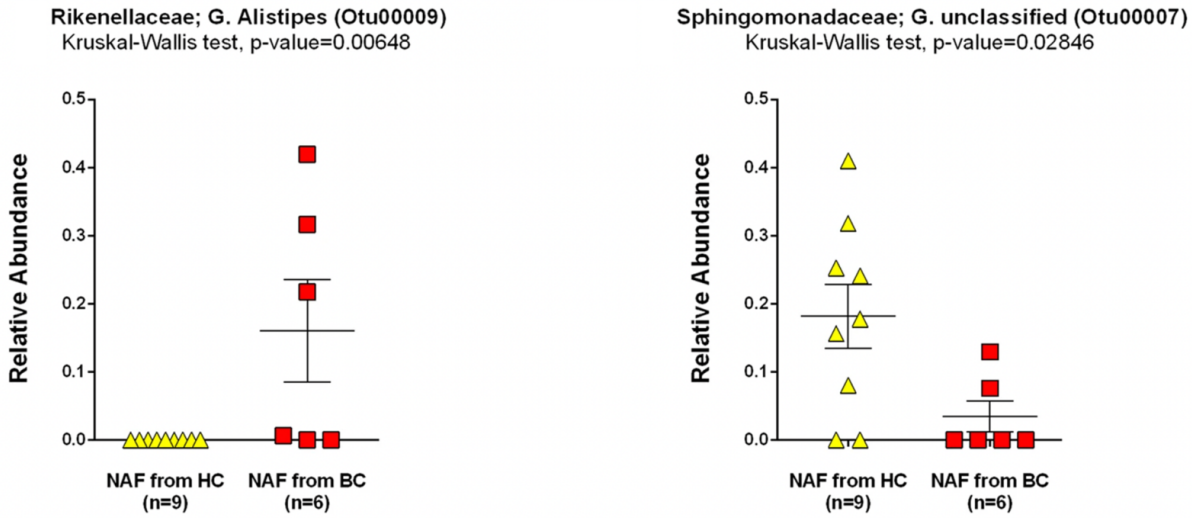

Figure 15. Original Study Differentially abundant taxa present in NAF.
Kruskal-Wallis test identifying differentially abundant taxa.

Chan et al. Re-analysis Results

The figures reported in this section are our results from the re-analysis of the Chan et al. paper.

Alpha Diversity. In the alpha diversity analysis, we have implemented the observed OTU number along with a nonparametric t-test on NS, NAF, and PBS microbiota. The observed OTU number shows the number of OTUs observed for the NS, NAF, and PBS tissues. The t-test assesses whether the microbial diversity is significantly different between the healthy control and cancer samples from NS, NAF, and PBS environments. The p-values are not significant for NS, NAF, and PBS samples (Fig. 16A; Fig. 16B; Fig. 16C).

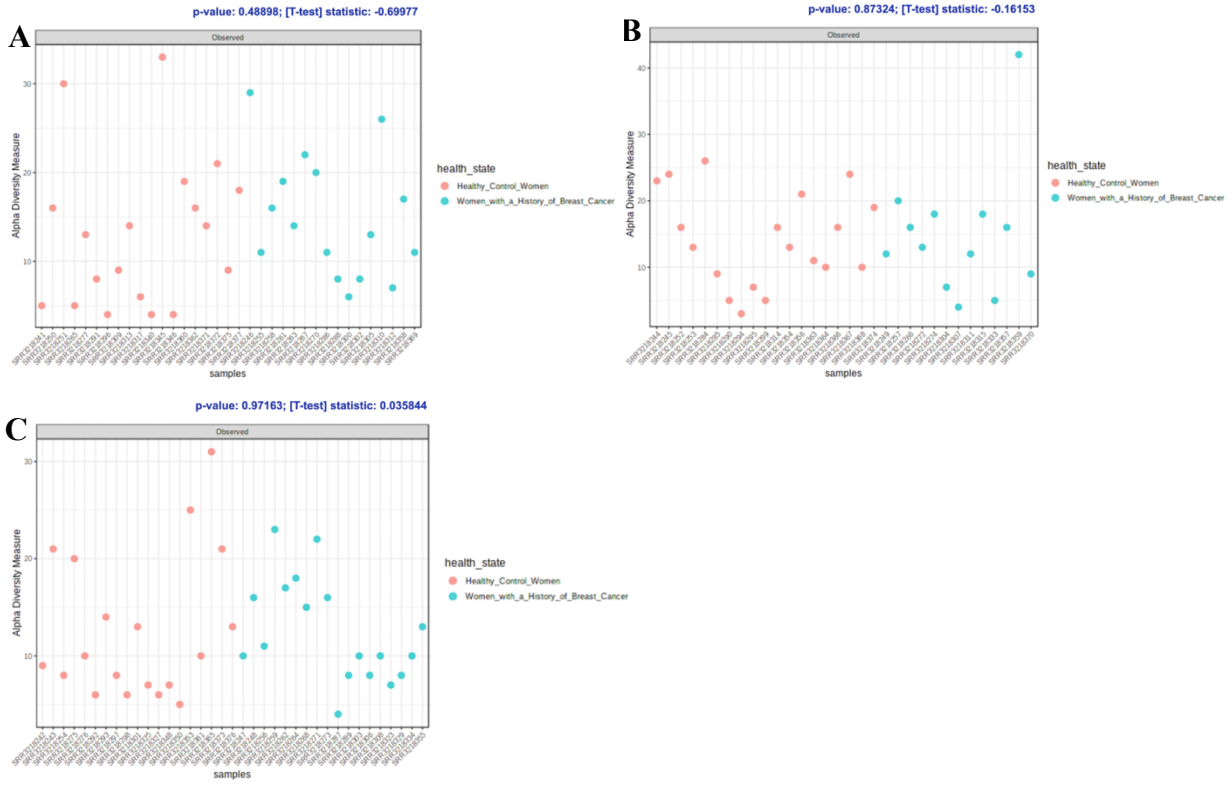

Figure 16. Alpha Diversity metric: Observed OTU number.
(A) NS Observed OTU metric with p-value of 0.489. (B) NAF Observed OTU metric with p-value of 0.873. (C) PBS Observed OTU metric with p-value of 0.972.

Beta Diversity. In the beta diversity analysis, we have implemented a Bray-Curtis dissimilarity metric and performed PCoA using the rarefied OTU abundances as input where the genus-level OTUs were used in PCoA. The Adonis test was used to test for compositional differences, and the adonis function in the vegan package (Oksanen et al., 2022) from R was used to implement this test. The healthy control and cancer samples from the NS microbiota appear to separate into clusters, but when running the Adonis test there were no significant differences found in the bacterial composition with an unadjusted p-value of 0.262 (Fig. 17A). The healthy control and cancer samples from the NAF microbiota appear to separate into clusters, but when running the Adonis test there were no significant differences found in the bacterial composition with an unadjusted p-value of 0.254 (Fig. 17B). The healthy control and cancer samples from the PBS microbiota appear to separate into clusters, but when running the Adonis test there were no significant differences found in the bacterial composition with an unadjusted p-value of 0.785 (Fig. 17C).

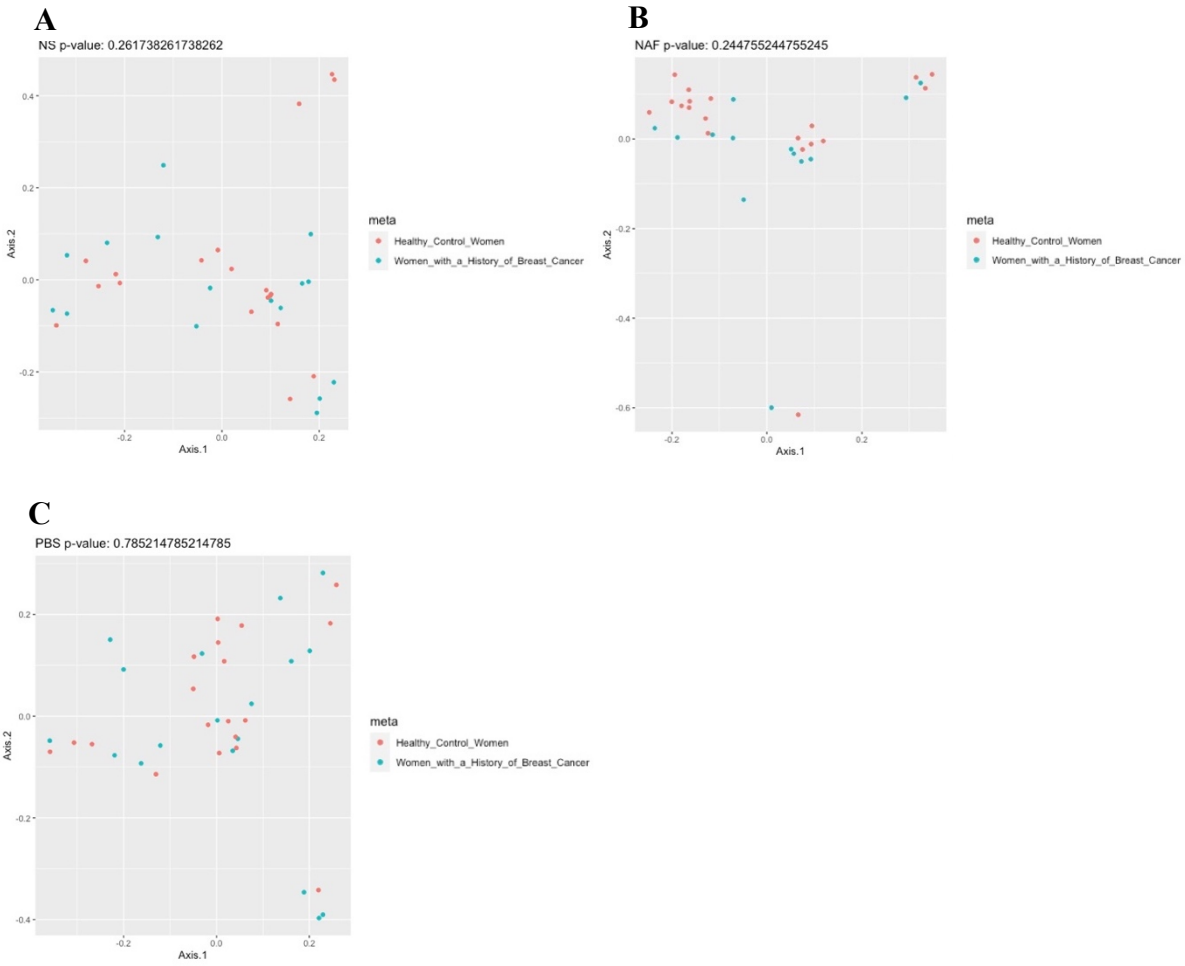

Figure 17. Adonis test: Bacterial Diversity.
Comparing healthy control women and women with a history of breast cancer across (A) NS, (B) NAF, and (C) PBS samples.

Proportional Abundance Analysis. In the proportional abundance analysis, we have assessed the taxonomic composition of NS, NAF, and PBS microbiota at the phylum level. The proportional abundance plots include the top 300 sequences in order to clearly show the taxa present at the phylum level. The NS microbial composition was predominantly comprised of the phyla Proteobacteria, Firmicutes, and Bacteroidota (Fig. 18A). The PBS microbial composition was predominantly comprised of the phyla Proteobacteria, Firmicutes, Bacteroidota, and Actinobacteriota (Fig. 18C). The NAF microbial composition was predominantly comprised of the phyla Proteobacteria and Firmicutes (Fig. 18B).

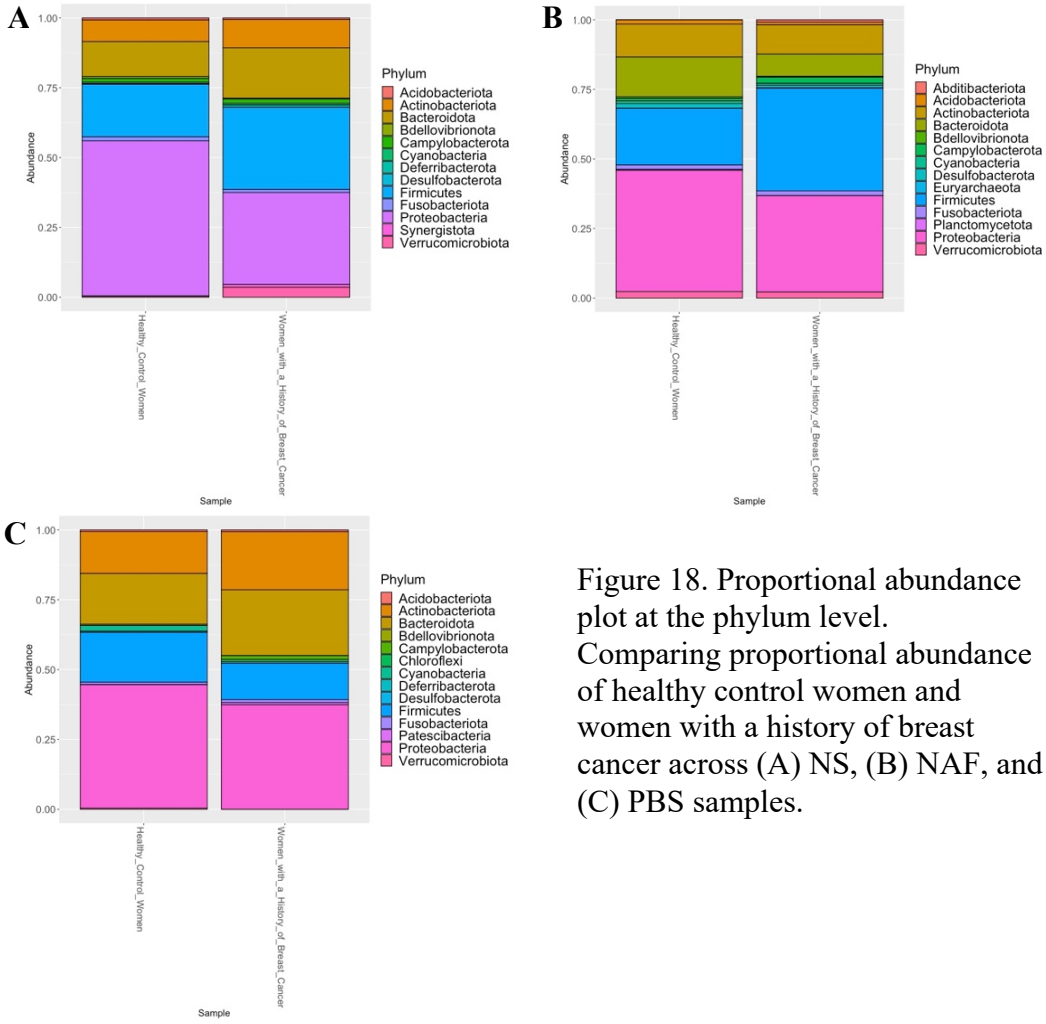

Figure 18. Proportional abundance plot at the phylum level. Comparing proportional abundance of healthy control women and women with a history of breast cancer across (A) NS, (B) NAF, and (C) PBS samples.

Differential Abundance Analysis. In the differential abundance analysis, the Kruskal-Wallis test was performed on the healthy and cancer samples from NS, NAF, and PBS OTUs and this test was performed through the kruskal.test function in the stats package (Bolar, 2019) in R. The NS OTUs were not significantly different when comparing the healthy and cancer samples through the Kruskal-Wallis test. There was one PBS OTU identified to be significantly different in relative abundance when comparing the healthy and cancer samples through the Kruskal-Wallis test, and it was *Bacteroides* at the genus level present in only cancer PBS samples with an unadjusted p-value of 0.032 (Fig. 19B). There were two NAF OTUs identified to be significantly different in relative abundance when comparing the healthy and cancer samples through the Kruskal-Wallis test, and they were *Alistipes* and *Acinetobacter* at the genus level (Fig. 19A) present in both healthy and cancer NAF samples with an unadjusted p-value of 0.045 for *Acinetobacter* and 0.011 for *Alistipes*.

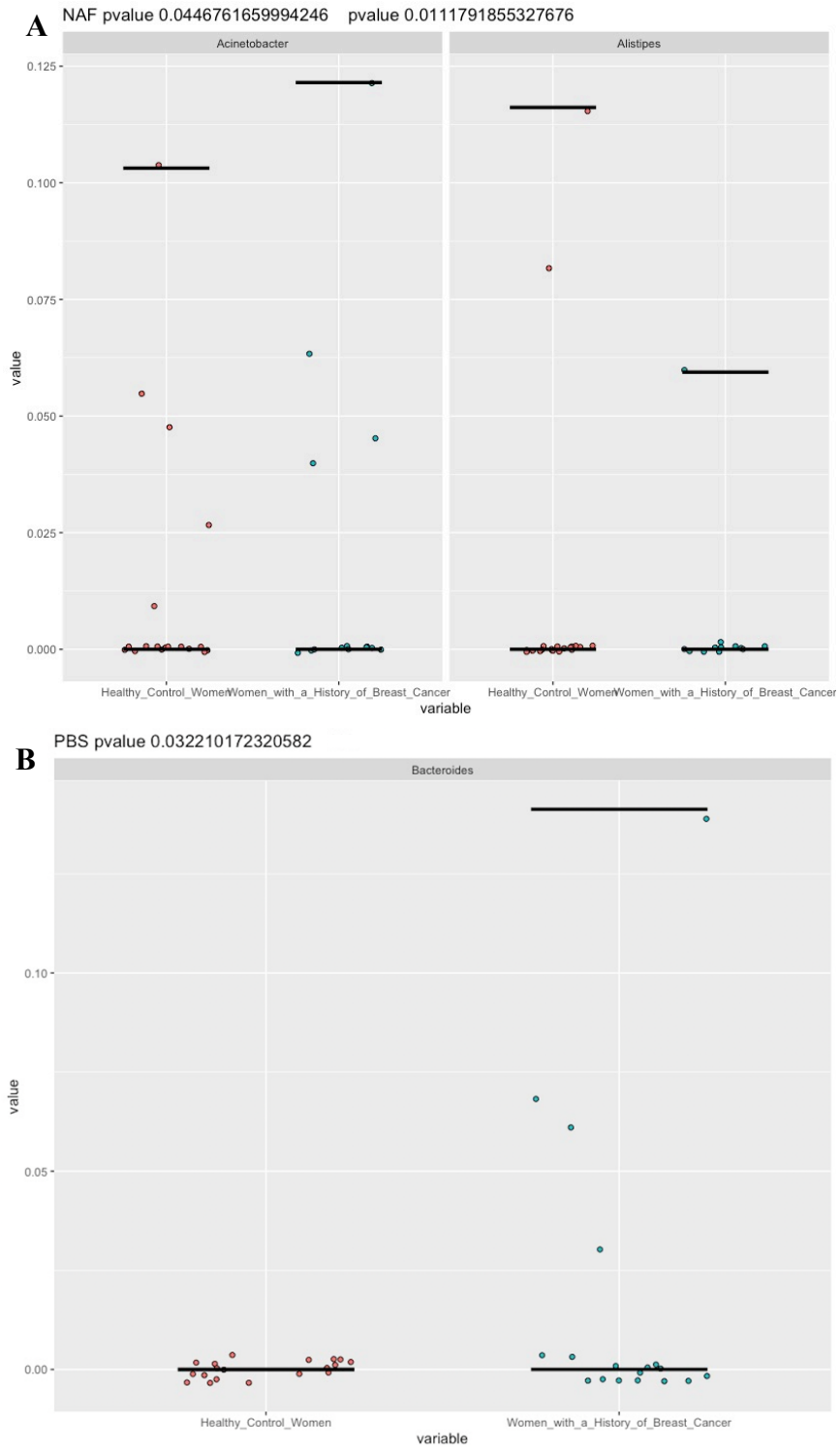

Figure 19. Kruskal-Wallis: significant OTUs.
(A) NAF: Kruskal-Wallis test result of differentially abundant genera, *Acinetobacter* and *Alistipes*. (B) PBS: Kruskal-Wallis test result of differentially abundant genus *Bacteroides*.

Urbaniak et al. Original Study Results

The figures reported in this section are from the original Urbaniak et al. paper.

Beta Diversity. In the beta diversity analysis, they used weighted UniFrac distance and performed PCoA to show the different bacterial microbiota between the cancer and healthy samples. They also performed unsupervised K-means clustering of CLR-transformed data which was followed by PCoA. They found that the weighted UniFrac PCoA plot showed different bacterial profiles between the cancer and healthy samples (Fig. 20A). PERMANOVA was then performed on the weighted UniFrac distance which reported a p-value of 0.01, and it was deduced that the differences observed are statistically significant. The unsupervised K-means clustering was performed using Euclidean distances on the CLR-transformed data and this was followed by PCoA (Fig. 20B). The unsupervised K-means clusterplot clearly indicated two separate clusters and a clear separation between the healthy and cancer samples, and these results were further confirmed through MiRKAT of which the reported p-values were significant.

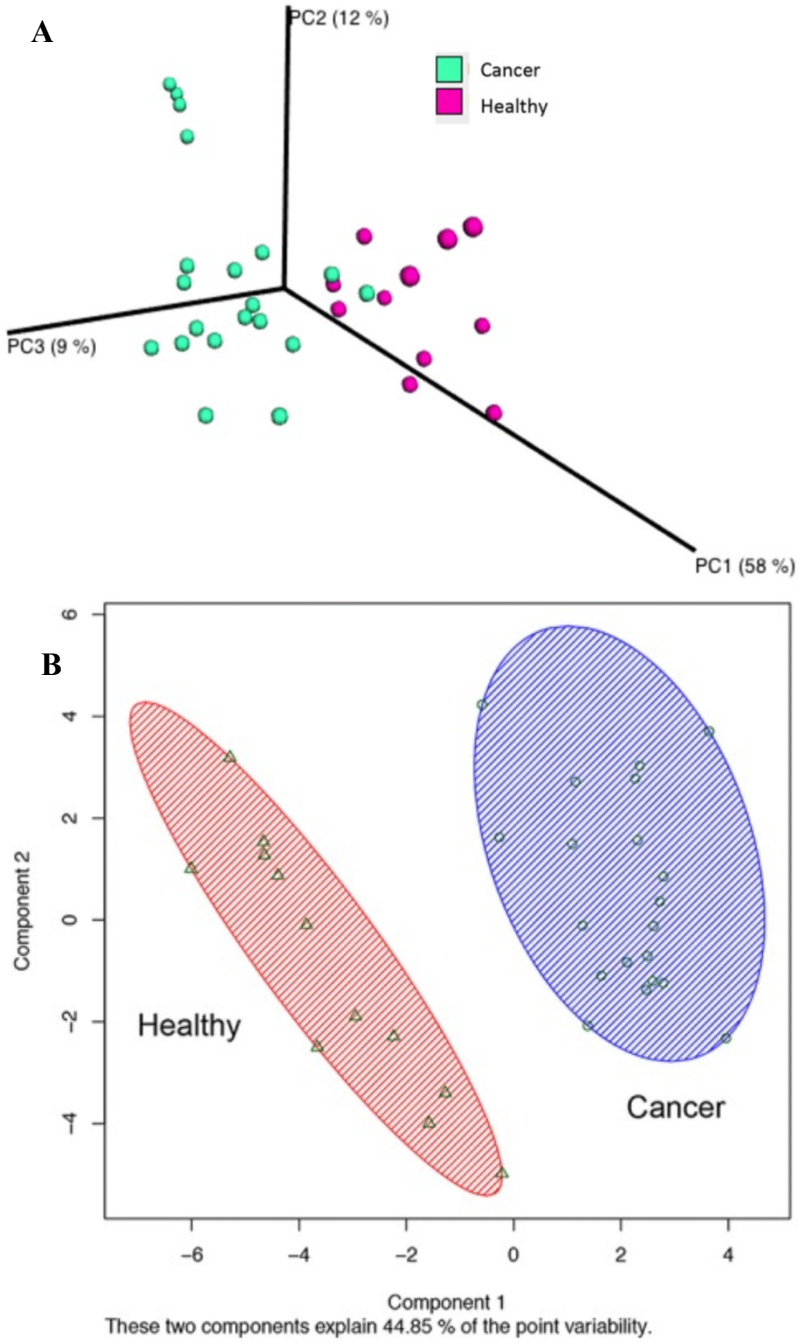

Figure 20. Original Study Comparison of healthy and control samples. (A) Weighted UniFrac PCoA plot. (B) K-means cluster plot of clr-transformed data.

Proportional Abundance Analysis. In the proportional abundance analysis, they assessed the taxonomic composition of healthy, cancer, and benign samples at the genus level. The genus level proportional abundance plot showed a diverse population of bacteria containing 28 genera and 61 OTUs (Fig. 21). The proportional abundance plot predominantly consisted of the phyla Proteobacteria and Firmicutes.

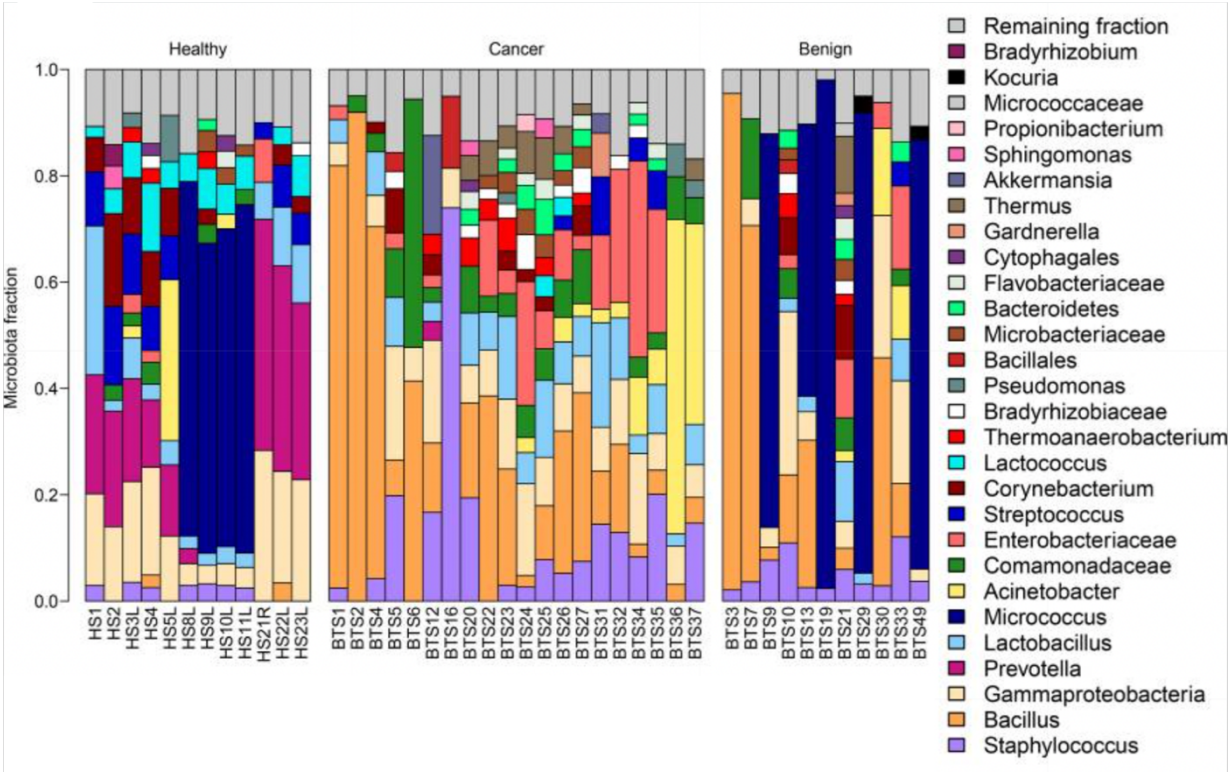

Figure 21. Original Study Proportional abundance plot at the genus level.

Differential Abundance Analysis. In the differential abundance analysis, the ALDEx R package version 2 (Gloor et al., 2022) was used to compare relative abundances of taxa at the genus level of CLR-transformed data. The reported p-values from ALDEx2 are Benjamini-Hochberg corrected p-values of the Wilcoxon rank test. The ALDEx2 output was visualized through boxplots and their results showed significantly higher abundances of the following genera in healthy samples *Prevotella*, *Lactococcus*, *Streptococcus*, *Corynebacterium*, and *Micrococcus*. There were significantly higher abundances of the following taxa in cancer samples *Bacillus*, *Staphylococcus*, *Enterobacteriaceae* (unclassified genus), *Comamonadaceae* (unclassified genus), and *Bacteroidetes* (unclassified genus) (Fig. 22).

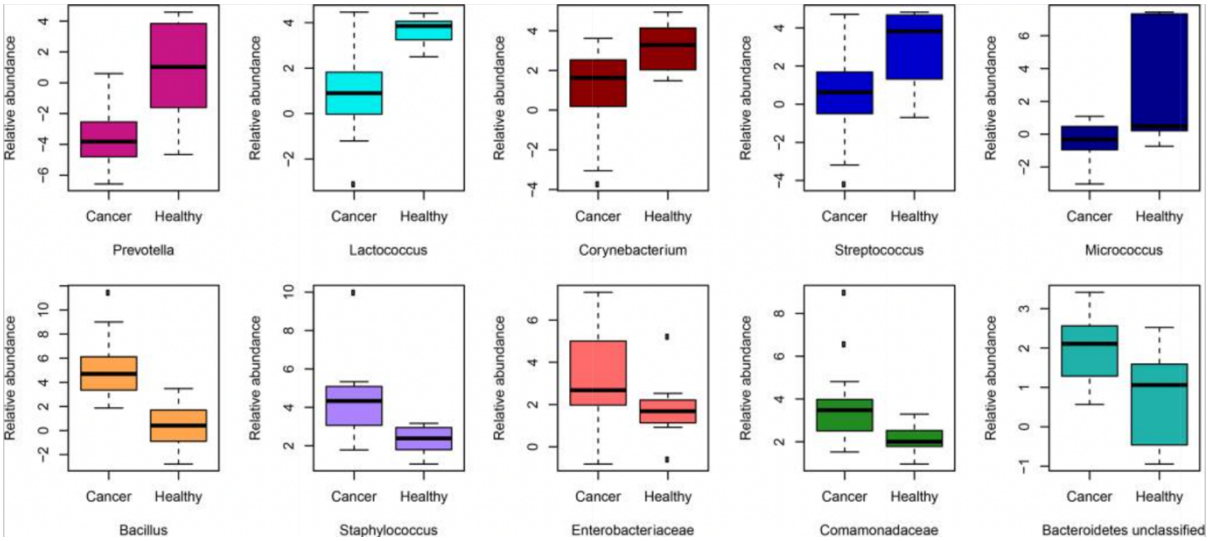

Figure 22. Original Study ALDEx2 output.
Differentially abundant taxa in healthy samples: *Prevotella, Lactococcus, Streptococcus, Corynebacterium,* and *Micrococcus.*
Differentially abundant taxa in cancer samples: *Bacillus*, *Staphylococcus*, *Enterobacteriaceae* (unclassified genus), *Comamonadaceae* (unclassified genus), and *Bacteroidetes* (unclassified genus).

Urbaniak et al. Re-analysis Results

The figures reported in this section are our results from the re-analysis of the Urbaniak et al. paper.

Beta Diversity. In the beta diversity analysis, we performed unsupervised k-means clustering on CLR-transformed data. The clr function in the R package compositions (Boogart et al., 2022) was used to CLR-transform the data and the pam function in the R package cluster (Rousseeuw et al., 2022) was used to perform the unsupervised k-means clustering. There is a clear separation between the breast tumor (BT) cancer samples and the healthy (H) samples as shown in the clusterplot (Fig. 23) where the two components explain 67.18% of the variability.

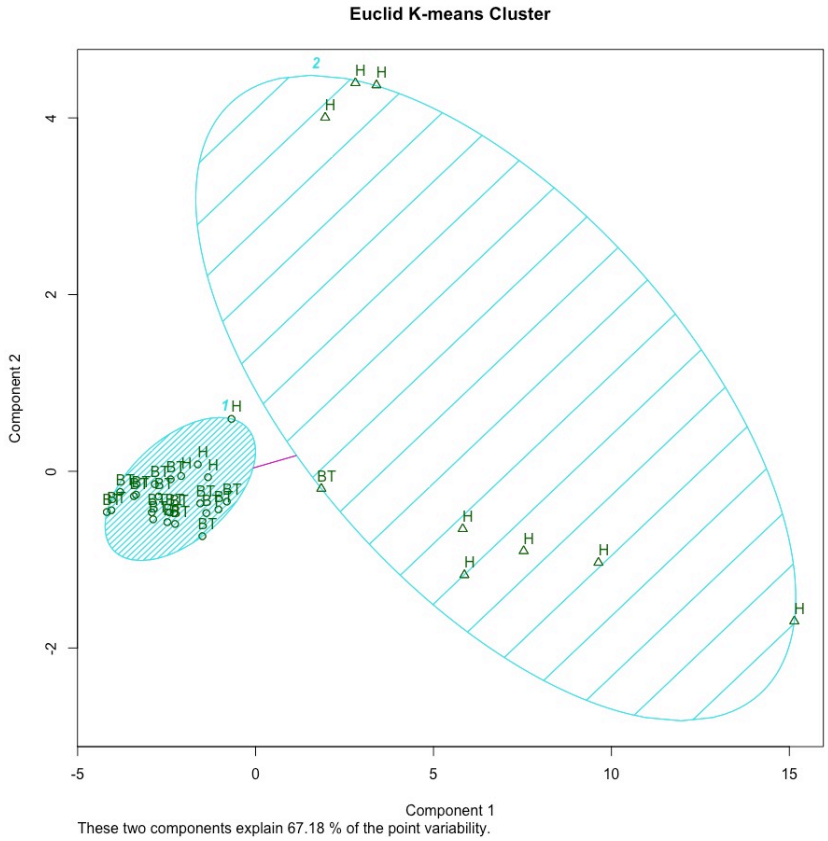

Figure 23. K-means cluster plot comparing healthy and control samples.
K-means clustering plot of CLR-transformed ASV table output from the DADA2 analysis.

Proportional Abundance Analysis. In the proportional abundance analysis, we assessed the taxonomic composition of healthy, cancer, and benign samples at the genus level. The genus level proportional abundance plot showed a diverse population of bacteria containing 65 genera and top 100 OTUs (Fig. 24). The proportional abundance plot consists of *Pseudomonas* and *Bacillus* across the benign, cancer, and healthy samples, where *Escherichia-Shigella* was shown to be prevalent in the cancer samples.

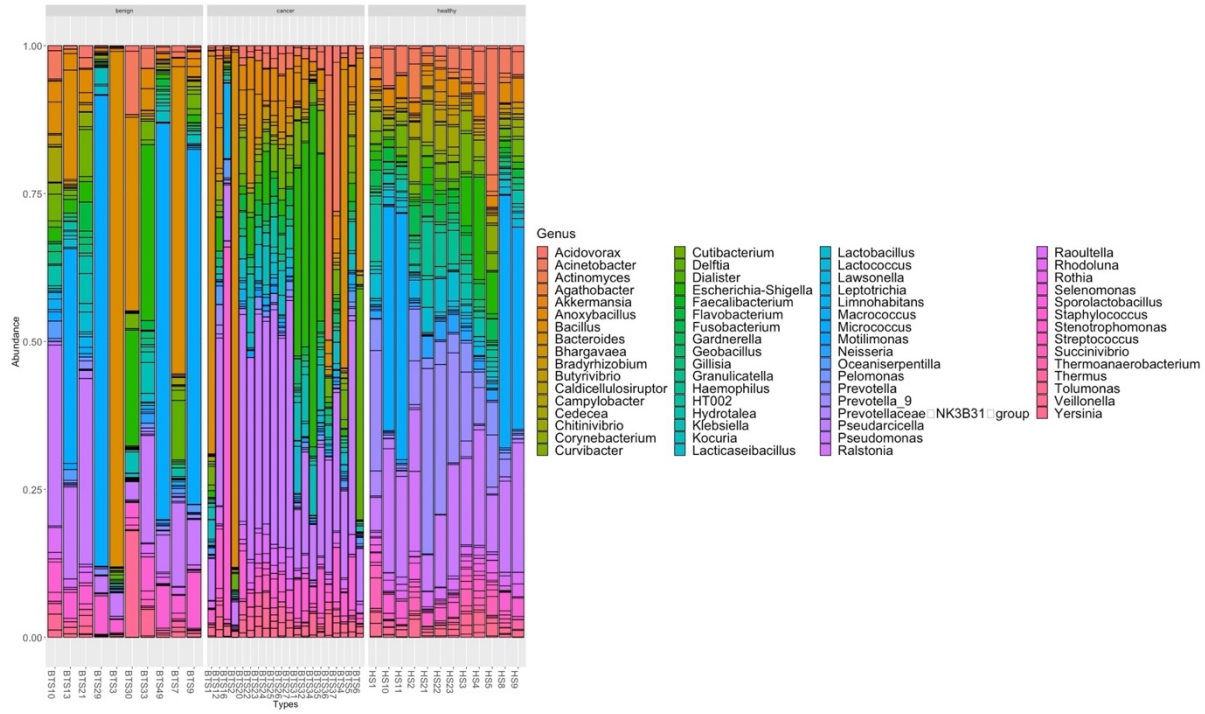

Figure 24. Proportional abundance plot at the genus level.
Proportional abundance plot made using the ASV table output from the DADA2 analysis and taxonomic assignment based on SILVA version 138, with benign (left), cancer (middle), and healthy (right) samples.

Differential Abundance Analysis. In the differential abundance analysis, the ALDEx R package version 2 (Gloor et al., 2022) was used to compare relative abundances of taxa at the genus level of CLR-transformed data. The reported p-values from ALDEx2 are Benjamini-Hochberg corrected p-values of the Wilcoxon rank test. The ALDEx2 output was visualized through boxplots (Fig. 25) and there were 20 statistically significant OTUs and 16 statistically significant taxa of which the following genera were significantly higher in abundance in healthy samples *Acinetobacter*, *Prevotella_9*, *Lactobacillus*, *Corynebacterium*, *Lactococcus*, *Fusobacterium*, *Prevotella*, *Actinomyces*, *Agathobacter*, *Campylobacter*, *Butyrivibrio*, *Rothia*, *Veillonella*, *Faecalibacterium*, and *Micrococcus*. There was a significantly higher abundance of the following genus in breast tumor cancer samples *Bacillus*.

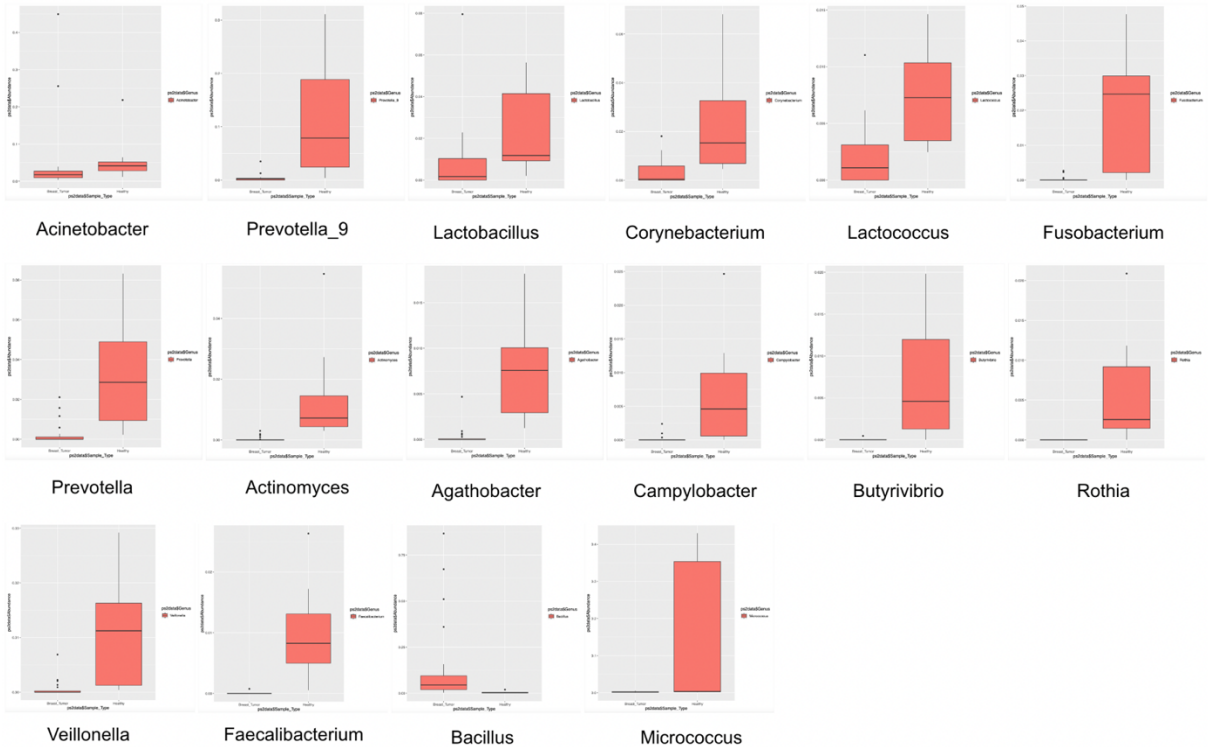

Figure 25. ALDEx2 output.
The 16 statistically significant taxa visualized through boxplots where on x-axis is healthy (right) and breast_tumor (left) and y-axis is the abundance.

Discussion and Conclusion

Hieken et al. study

In the original study's results, they found differences between the breast and skin tissue microbiota in rare lineages through the observed OTU number and heatmap. However, our results from using the ASV table output from DADA2 does not show differences between the breast and skin tissue microbiota as shown through the heatmap and Shannon Index, where the observed OTU number shows some difference between the two tissues.

In the original study, they also found a significant difference between breast and skin tissue microbiota from performing unweighted UniFrac distance metric but did not find significant difference in weighted UniFrac distance. They also found a significant difference in microbiota between breast tissue adjacent to invasive cancer and women with benign disease from performing unweighted UniFrac distance metric but did not find significant difference in weighted UniFrac distance. Additionally, they showed that the microbiome of the different tissue types – buccal swab, skin swab, breast tissue, and skin tissue – cluster distinctly from one another through a PCoA plot of the unweighted UniFrac distance.

In our analysis, we performed unweighted and weighted UniFrac distance metrics to compare breast and skin tissue microbiota and there was no significant difference found between the two tissue microbiota and MiRKAT p-values were not significant with p-values greater than 0.05. We also performed unweighted and weighted UniFrac distance metrics to compare microbiota between breast tissue adjacent to invasive cancer and women with benign disease. There was no significant difference found and MiRKAT p-values were not significant with p-values greater than 0.05. Through a PCoA plot of the unweighted UniFrac distance, we also

showed that the microbiome of buccal swab, skin swab, breast tissue, and skin tissue cluster separately from one another. Similar to the original study's results, there is a clear separation between the buccal and skin swabs; however, the separation between the breast and skin tissue is not as clear as that of the original study's PCoA plot as the breast tissue overlaps with the tail of the buccal swab ellipses and overlaps across the skin tissue.

Additionally, the original study assessed the taxonomic composition between breast and skin tissue, benign and malignant disease states in breast tissue, and buccal and skin swab microbiota. There were unclassified taxa frequently observed in their proportional abundance plots. However, our results from using the ASV table output from DADA2 does not show any unclassified taxa at any taxonomic level in any of the proportional abundance plots, and there are more genera identified in our proportional abundance plots as compared to the original study's results.

In original study's differential abundance analysis, they observed five differentially low-abundant genera to be significant and are primarily located in the invasive cancer disease state in breast tissue. The five differentially abundant genera prevalent in the invasive cancer disease state are *Fusobacterium, Atopobium, Gluconacetobacter, Hydrogenophaga,* and *Lactobacillus*. However, our results show twelve different differentially abundant taxa that are primarily located in the invasive cancer disease state in breast tissue through a more recent method called linda. Of the twelve differentially abundant taxa, *Eubacterium xylanophilum group, ASF356, Colicodextribacter, Lachnospiraceae NK4A136 group, Escherichia-Shigella,* and *Staphylococcus* are primarily located in the invasive cancer disease state, *Atopobium, Bacteroides,* and *Subdoligranulum* are primarily in the benign disease state, and *Rikenellaceae*

RC9 gut group, Veillonella, and Clostridia UCG-014 are prevalent in both the invasive cancer and benign disease states.

Chan et al. study

In the original study's results, they implemented observed OTU number on NS, NAF, and PBS microbiota and did not find any differences in diversity as the reported p-values were not significant. Similarly, our results showed that there were no differences in diversity as the p-values were not significant.

In the original study, they implemented a Bray-Curtis dissimilarity metric and the Adonis test on NS, NAF, and PBS samples where difference in bacterial diversity was found in the NAF samples as the unadjusted p-value was significant for NAF microbiota. In our analysis, through the Adonis test, there were no significant differences found across the NS, NAF, and PBS samples.

In comparison of the proportional abundance plots between the original study's results and our results, there were more phyla identified by our ASV table from the DADA2 analysis than what was reported in the original study through implementation of updated methods and reference database such as DADA2 and SILVA.

In the original study, they identified two significant OTUs through the Kruskal-Wallis test, and those two significant OTUs were present in only NAF samples and the two OTUs were the genus *Alistipes* and family *Sphingomonadaceae*. In our analysis, there were three significant ASVs identified of which two were present in NAF and one in PBS and they were genera *Alistipes* and *Acinetobacter* in NAF and genus *Bacteroides* in PBS.

Urbaniak et al. study

In the original study's results, their K-means clustering plot shows a clear separation between the cancer and healthy samples; however, our K-means clustering plot shows some overlap between the healthy and cancer samples. The difference between the two states is not as clear as it is shown in the original study's results.

The proportional abundance plot at the genus level from the original study shows a prevalence of *Micrococcus* in the benign and healthy samples and a prevalence of *Bacillus* in the cancer samples. However, our results show prevalence of *Pseudomonas* and *Bacillus* across the benign, healthy, and cancer samples, and prevalence of *Escherichia-Shigella* in cancer samples. Additionally, our results have identified more genera as compared to the original study's results.

Through performing ALDEx2, the original study reports the following taxa to be significantly higher in abundance in healthy samples: *Prevotella*, *Lactococcus*, *Streptococcus*, *Corynebacterium*, and *Micrococcus* and in cancer samples: *Bacillus*, *Staphylococcus*, *Enterobacteriaceae* (unclassified genus), *Comamondaceae* (unclassified genus), *and Bacteroidetes* (unclassified genus) were significantly higher in abundance. In our results when we performed ALDEx2 using the ASV table as input, *Prevotella*, *Corynebacterium*, *Micrococcus*, and *Lactococcus* were also found in higher abundance in healthy samples along with *Acinetobacter*, *Lactobacillus*, *Fusobacterium*, *Prevotella_9*, *Actinomyces*, *Agathobacter*, *Campylobacter*, *Butyrivibrio*, *Rothia*, *Veillonella*, and *Faecalibacterium.* In the cancer samples, only *Bacillus* was found as being significantly higher in abundance.

Summary

There are overlaps between the original study's results and results from our analysis; however, there are differences where results from our analysis do not identify taxa that the original study identifies as significant. There are differences in the proportional abundance as results from our analyses are able to identify more taxa and also identify differentially abundant taxa that the original studies are not able to identify. The differences between the results from our reanalysis and results from the original study could be due to the differences in the bioinformatic pipeline that was implemented or in the reference database used to assign taxonomy. The differences in the reference database, as alluded to in the introduction of this chapter, can contribute to differences in taxa that are identified in both the proportional differential abundance analyses between the original study's results and the re-analysis results. The SILVA database is well-maintained and frequently updated whereas the Greengenes database is deprecated; however, in preliminary data, not shown here, we performed downstream analyses using the ASV table output from DADA2 and the Greengenes reference database and we found that our results from these analyses did not mimic the original studies' results. These analyses were performed for the Hieken et al. dataset and Urbaniak et al. dataset, and our results tell us that the DADA2 pipeline compared to the original pipeline also plays a large role in contributing to the differences found between the re-analysis results and original studies' results. There is further research required to understand the role of the bioinformatic pipeline in microbiome analyses.

The results found from our re-analysis offer new insight into the breast tissue microbial community in healthy, benign, and malignant disease states across different variable regions and

patient cohorts. The findings from our re-analysis are able to identify and define more taxa than what was previously reported in each study. There are some overlaps between the findings from our re-analysis, such as *Escherichia-Shigella* is found to be differentially abundant in invasive cancer breast tissue in the Hieken et al. re-analysis and it was also found to be prevalent in the cancer samples in the proportional abundance plot of the Urbaniak et al. re-analysis. Also, *Veillonella* is found to be differentially abundant in benign breast tissue in the Hieken et al. re-analysis and is also found to be differentially abundant in healthy breast tissue in the Urbaniak et al. re-analysis, and, similarly, *Acinetobacter* is found to be differentially abundant in the NAF cancer disease state in the Chan et al. re-analysis and in healthy breast tissue in the Urbaniak et al. re-analysis. There are similarities across the re-analysis results for each study, especially when the tissue and disease states are similar as shown between the Urbaniak et al. re-analysis and Hieken et al. re-analysis, and these similarities may shine light on the microbial communities that could be present in the breast microbiome across disease states.

CHAPTER THREE

META-ANALYSIS

Introduction

The meta-analysis combines the following three studies *The Microbiome of Aseptically Collected Human Breast Tissue in Benign and Malignant Disease* study by Hieken et al., *Characterization of the microbiome of nipple aspirate fluid of breast cancer survivors* study by Chan et al., and *The Microbiota of Breast Tissue and Its Association with Breast Cancer* study by Urbaniak et al. The Hieken et al. study targets the V3-V5 variable region of their 16S rRNA data and look at benign and malignant samples across buccal swab, skin swab, breast, and skin tissues. The Chan et al. study targets the V4 variable region of their 16S rRNA data and look at women with a history of breast cancer and healthy control samples across NS, NAF, and PBS tissues. The Urbaniak et al. study targets the V6 variable region of their 16S rRNA data and look at benign, malignant, and healthy samples across breast tissue. In the meta-analysis, the output ASV tables from the three studies were combined into one large ASV table, the metadata tables for each study were combined into one large metadata table, and the output taxonomy tables for each study were combined into one large taxonomy table. These combined files are in pre-normalization format and post-normalization format, where normalization was performed through the percentile_norm python script from the *Correcting for batch effects in case-control microbiome studies* by Gibbons et al. (Gibbons et al., 2018).

Methods

Normalization

Prior to combining the ASV tables, metadata tables, and taxonomy tables, the ASV tables were normalized. Each individual ASV table was first total sum scaled and then normalized by running the percentile_norm python script (Gibbons et al., 2018) where the ASV table, control sample names, and cancer sample names were input. This python script normalizes the data and then converts the normalized values to percentages. After performing normalization, the ASV tables, metadata tables, and taxonomy tables were combined into each respective combined table for the pre-normalized and normalized datasets.

Alpha Diversity

In order to perform alpha diversity analysis, we implemented observed OTU number and the Shannon Index on the pre-normalized and normalized data through the MicrobiomeAnalyst web interface (Dhariwal et al., 2017). The combined ASV table, metadata table, and taxonomy table were input to the MicrobiomeAnalyst web interface for the pre-normalized and normalized datasets.

Beta Diversity Analysis

In order to perform beta diversity analysis, we implemented a robust Aitchison PCoA plot comparing the microbiota of tissue across all studies, and a robust Aitchison PCoA plot comparing tumor vs healthy samples across all studies. Prior to calculating the robust Aitchison distance, the combined ASV table was sorted by decreasing abundance with the most abundant ASV first and least abundant ASV last through an abundance filter. Thereafter, this ASV table was input to phyloseq and this phyloseq object was passed through a prevalence filter through

the phyloseq_filter_prevalence function in the metagMisc package in R (Mikryukov, 2018), where a prevalence filter of 10% was applied to the pre-normalized ASV table and prevalence filter of 25% was applied to the normalized ASV table. After applying the abundance and prevalence filters, the ASV table was input to the vegdist function in the vegan package in R (Jari Oksanen et al., 2022) to calculate the robust Aitchison distance and this distance was input to the pcoa function in the ape package in R (Emmanuel Paradis et al., 2022).

Proportional Abundance Analysis

In order to perform proportional abundance analysis, the combined ASV table, metadata table, and taxonomy table for both pre-normalized and normalized data were input to the phyloseq function to make the proportional abundance plots. The prune_taxa, tax_glom, transform_sample_counts functions in phyloseq were implemented to make the proportional abundance plots. The proportional abundance plots were made at the phylum, family, and genus levels for both the pre-normalized and normalized data.

Differential Abundance Analysis

In order to perform differential abundance analysis, Kruskal-Wallis was performed and p-values were adjusted with the p.adjust function in the stats package in R (Bolar, 2019). The combined ASV table, metadata table, and taxonomy table for each dataset, pre-normalized and normalized, were passed through the abundance and prevalence filter as stated in this section's beta diversity analysis methods. The otu table from the phyloseq object from the phyloseq_filter_prevalence function was input to the kruskal.test function from the stats package, and then p.adjust was run to adjust the p-values for correction. The taxa with significant p-values were visualized in R in form of dot plots.

Furthermore, the combined normalized data was pooled based on different tissue types and the modified ASV, metadata, and taxa tables were input to the MicrobiomeAnalyst web interface (Dhariwal et al., 2017). In order to compare breast tissue, the benign and malignant samples were pooled from the Hieken and Urbaniak datasets as these studies compared breast tissue between benign and malignant disease states and comparisons were made by disease states. To pool the benign and malignant samples of breast tissue from Hieken and Urbaniak datasets, the benign samples from the Hieken data were combined with the benign samples from the Urbaniak data and same was applied for combining the malignant samples. Any samples that were not breast tissue were removed from both the ASV and metadata tables. The modified ASV, metadata, and taxonomy tables were input to MicrobiomeAnalyst and Classical Univariate, metagenomeSeq, edgeR, and LEfSe analyses were performed.

In order to compare skin swabs, the benign, malignant or cancer, and healthy samples were pooled from the Hieken and Chan datasets as these studies compared skin swabs between benign, cancer, and healthy states. To pool the benign, cancer, and healthy samples of skin swabs from Hieken and Chan datasets, the benign samples from the Hieken data were combined with the healthy samples from the Chan data as the Chan data did not have benign samples and Hieken data did not have healthy samples. The malignant samples from the Hieken data were combined with the cancer samples from the Chan data. Any samples that were not skin swabs were removed from both the ASV and metadata tables. The modified ASV, metadata, and taxonomy tables were input to MicrobiomeAnalyst and Classical Univariate, metagenomeSeq, edgeR, and LEfSe analyses were performed.

Results

Alpha Diversity

In the alpha diversity analysis, the observed OTU number and Shannon Index were implemented for the pre-normalized and normalized data. The observed OTU number and Shannon Index alpha diversity metrics for the pre-normalized data revealed that the data clusters by study which is clearly indicated by the scatter and box plots for the two alpha diversity metrics (Fig. 26). The observed OTU number and Shannon Index alpha diversity metrics for the normalized data revealed that the data clearly clusters by study through the scatter and box plots for the two alpha diversity metrics (Fig. 27).

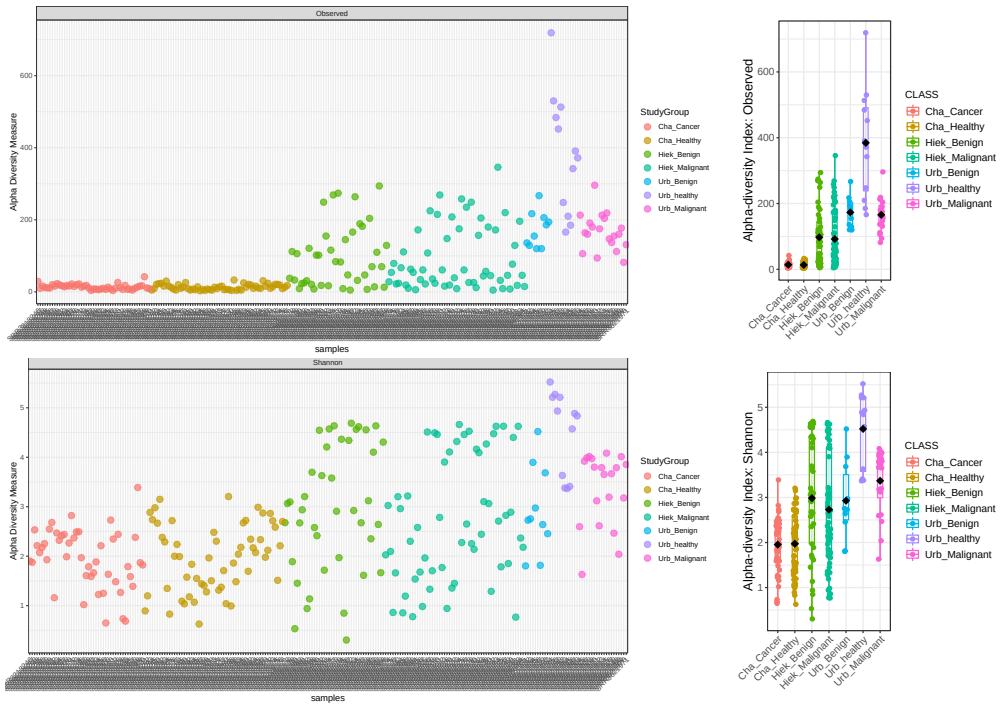

Figure 26. Alpha Diversity metrics for pre-normalized data: Observed OTU number and Shannon Index.
Observed OTU number scatter and box plots (top) and Shannon Index scatter and box plots (bottom) show the pre-normalized combined data clusters by study.

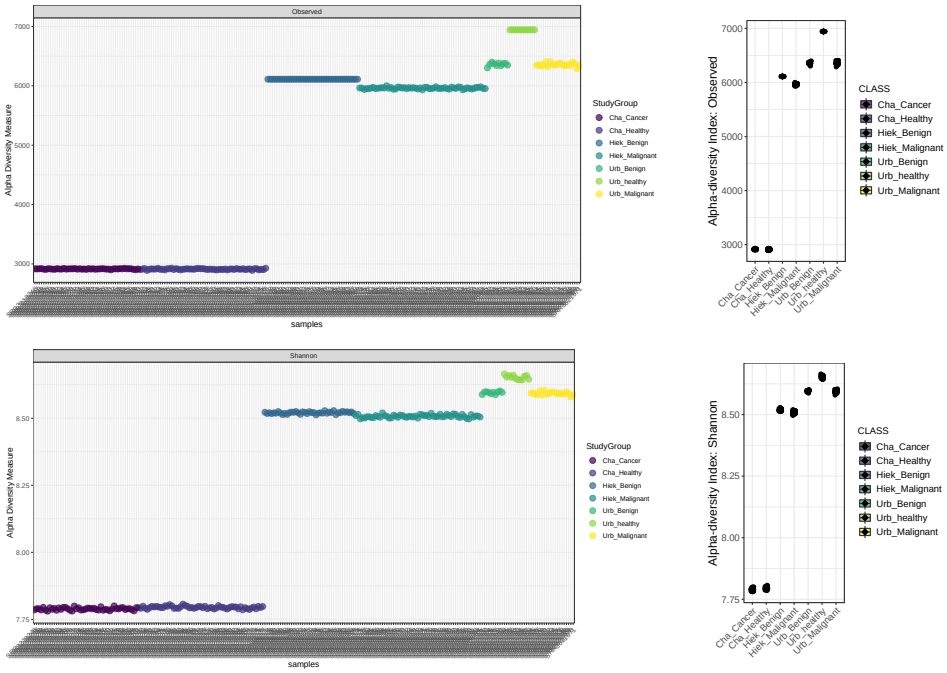

Figure 27. Alpha Diversity metrics for normalized data: Observed OTU number and Shannon Index.
Observed OTU number scatter and box plots (top) and Shannon Index scatter and box plots (bottom) show the normalized combined data clusters by study.

Beta Diversity

In the beta diversity analysis, the robust Aitchison distance metric was performed for the pre-normalized and normalized datasets comparing tissue and tumor vs control samples. The robust Aitchison distance metric for the pre-normalized data showed that when comparing tissue types (Fig. 28A), the samples cluster by study where the Chan samples are clustered separately from the Hieken and Urbaniak samples and Hieken samples are clustered separately from the Urbaniak samples and a similar trend is observed when comparing the tumor vs control samples (Fig. 28B). However, the robust Aitchison distance metric for the normalized data showed that when comparing tissue types, the samples mostly cluster by tissue type (Fig. 28C). Similarly, when comparing tumor vs control samples, the robust Aitchison distance metric for the

normalized data showed that the samples cluster by tumor vs control samples where the tumor samples cluster to one side and healthy samples overlap with each other across studies (Fig. 28D).

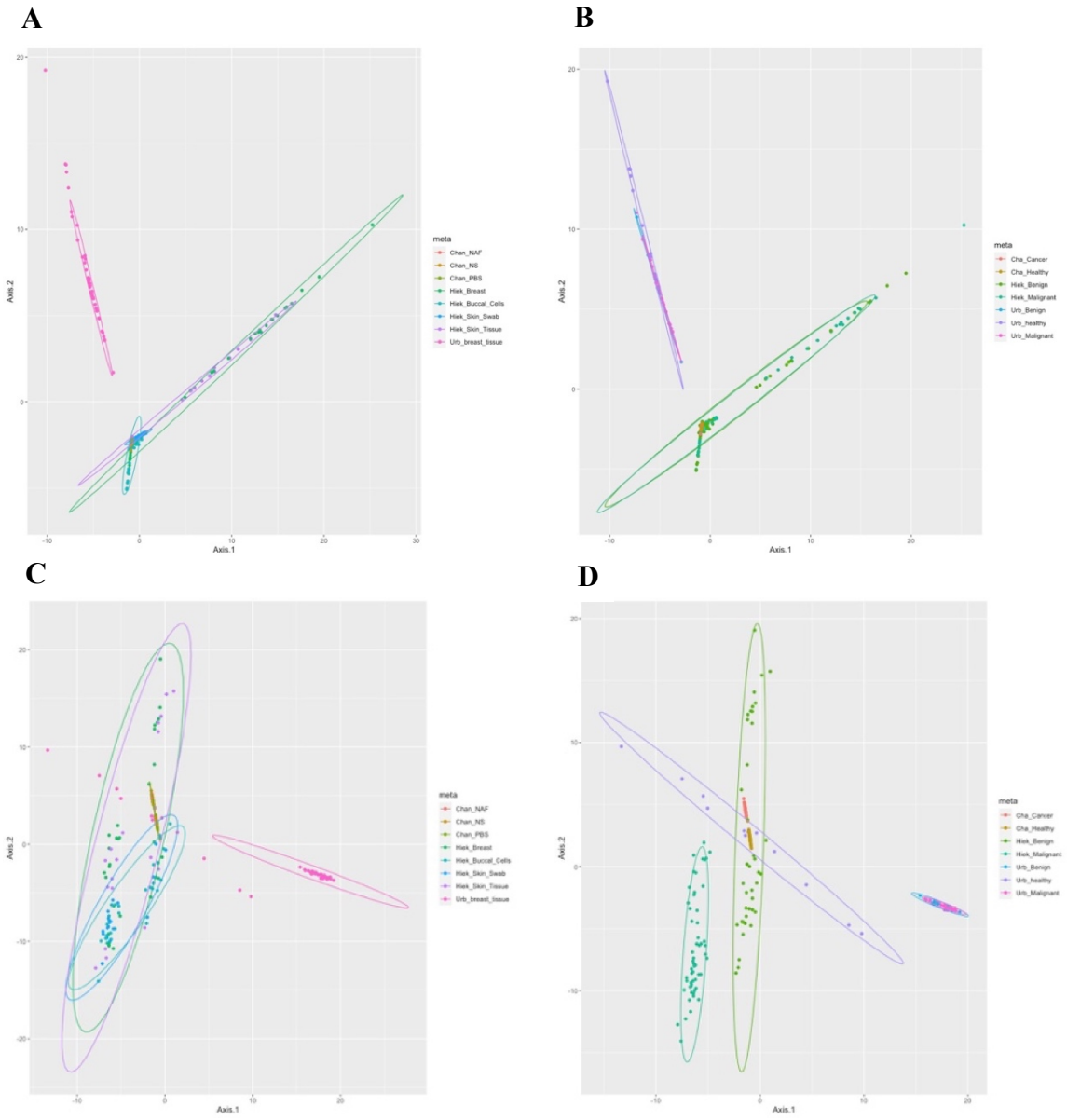

Figure 28. Robust Aitchison PCoA.
(A) PCoA plot comparing tissue types for pre-normalized data. (B) PCoA plot comparing tumor vs control for pre-normalized data. (C) PCoA plot comparing tissue types for normalized data. (D) PCoA plot comparing tumor vs control for normalized data.

Proportional Abundance Analysis

In the proportional abundance analysis, the proportional abundance plots at the phylum, family, and genus levels were made to compare the taxonomic composition across studies for the pre-normalized and normalized data. The proportional abundance plots compared the taxonomic composition across studies for tissue samples and tumor vs control samples. The proportional abundance plot for the pre-normalized data when comparing tissue microbiota across studies shows some similarity in taxonomic composition (Fig. 29). There is an abundance of the phyla Bacteroidota and Firmicutes across Hieken breast tissue, Hieken skin tissue, and Urbaniak breast tissue. Similarly, there is an abundance of the genus *Staphylococcus* across all tissue samples. The proportional abundance plot for the pre-normalized data when comparing tumor vs control samples across studies shows some similarity in taxonomic composition at the phylum level as there is a prevalence of the phyla Firmicutes and Proteobacteria across all samples (Fig. 30). At the genus level, there is an abundance of *Escherichia-Shigella* across Hieken benign, Hieken malignant, Urbaniak benign, and Urbaniak malignant samples. However, the proportional abundance plots at the genus level for the normalized data do not show overlaps between the samples across studies. The proportional abundance plots for the normalized data show that the taxonomic composition varies by study across all samples, and this is observed in both the tissue and tumor vs control samples (Fig. 31; Fig. 32).

The proportional abundance plot for the normalized data when comparing tissue disease states in the pooled samples shows similarity across all the samples (Fig. 33). However, the samples still appear to group together by study as the buccal_cells and skin_tissue samples from the Hieken study in benign and malignant disease states are similar to each other in abundance.

The NAF and PBS samples from the Chan study for the healthy and malignant disease states also appear to group together. The benign_breast and malignant_breast samples appear similar to each other as those are samples that have the Hieken and Urbaniak benign and malignant disease state data combined. The skin samples have the Hieken and Chan healthy and malignant disease state data combined, and it appears to be in similar abundance to the benign and malignant breast tissue samples. The proportional abundance plot for the pre-normalized data when comparing tissue disease states in the pooled samples does not show similarity across all the samples (Fig. 34). The different tissue samples from the Hieken data do not have similar abundances to each other, rather the benign and malignant disease state of a tissue share similar abundances as the buccal_cells in the benign and malignant disease state look similar to each other and similarly for the skin tissue samples. However, the Chan samples show similar abundances between disease states such as the healthy NAF and PBS samples show similar abundances and are different than the malignant NAF and PBS samples. The healthy breast tissue also shows similar abundances to that of the healthy skin, NAF, and PBS samples. Similar to the normalized proportional abundance plot (Fig. 33), the benign and malignant breast tissue samples show similar abundance and are also similar to the benign and malignant skin tissue proportional abundances.

Differential Abundance Analysis

In the differential abundance analysis, the significant taxa were identified based on adjusted p-values from the Kruskal-Wallis test. In the pre-normalized data, there were 121 OTUs and 61 genera identified as significant from the Kruskal-Wallis test after applying the false discovery correction. These significant taxa were present when comparing the tumor vs control samples in the pre-normalized data (Fig. 35). In the normalized data, significant taxa were not

identified, through the Kruskal-Wallis test, after applying the false discovery correction to the p-values.

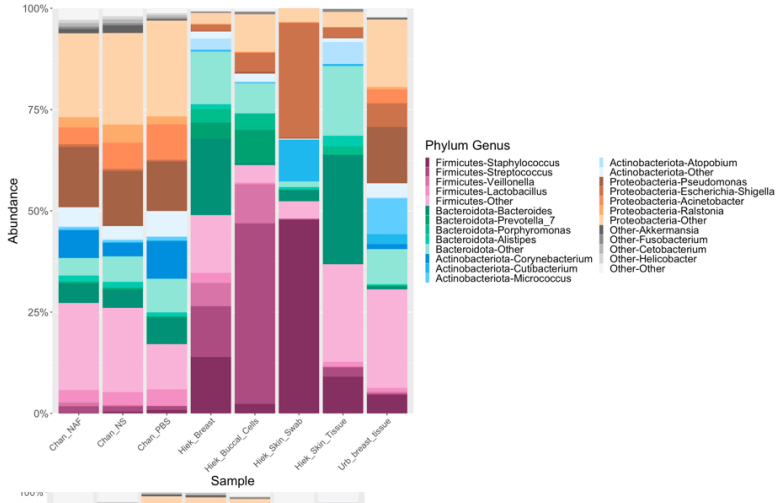

Figure 29. Proportional abundance plot at the phylum and genus levels of pre-normalized data – tissues. Comparing tissue samples across all studies.

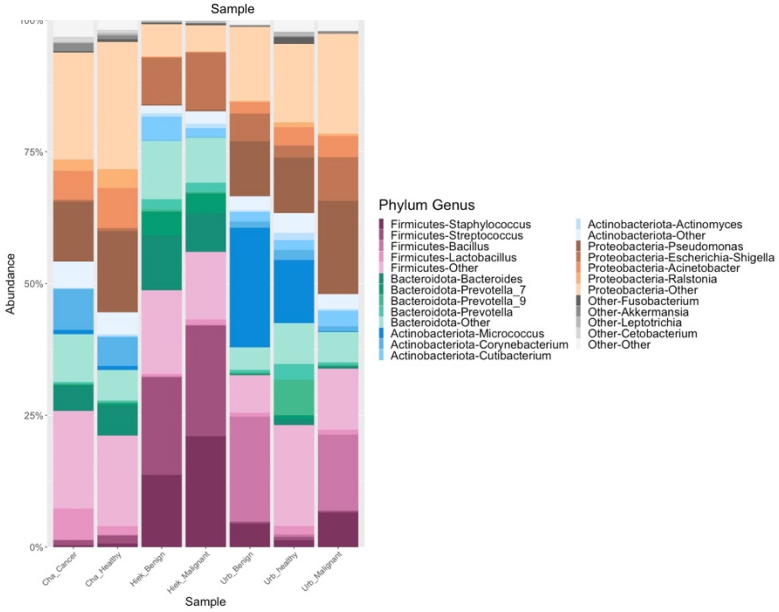

Figure 30. Proportional abundance plot at the phylum and genus levels of pre-normalized data – type. Comparing tumor vs control samples across all studies.

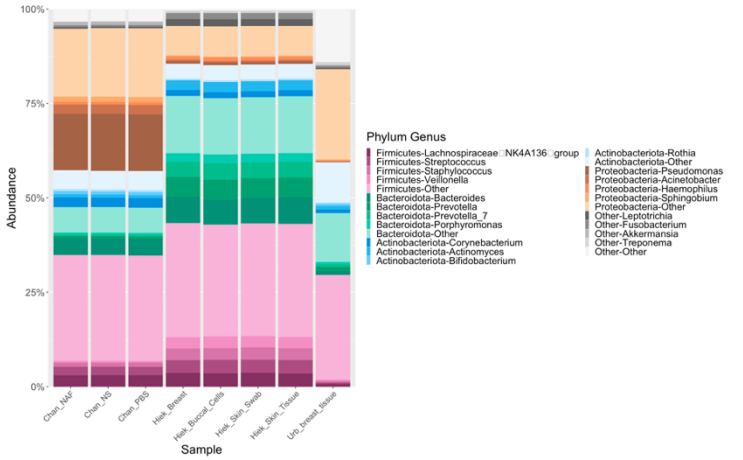

Figure 31. Proportional abundance plot at the phylum and genus levels of normalized data – tissues. Comparing tissues across all studies.

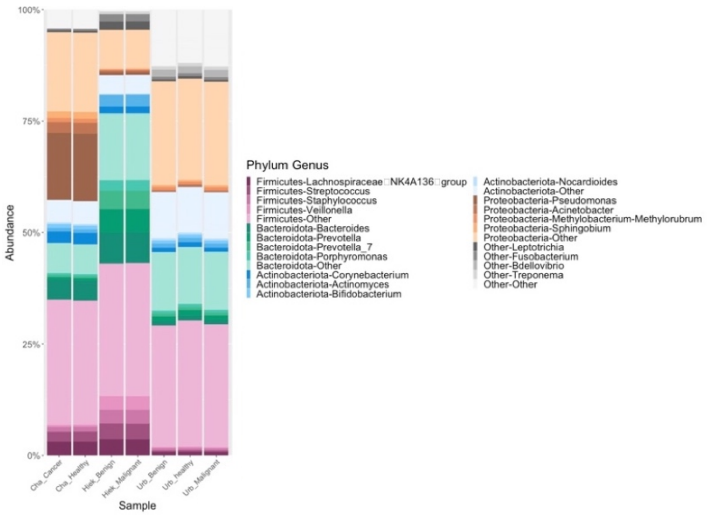

Figure 32. Proportional abundance plot at the phylum and genus levels of normalized data – type. Comparing tumor vs control samples across all studies.

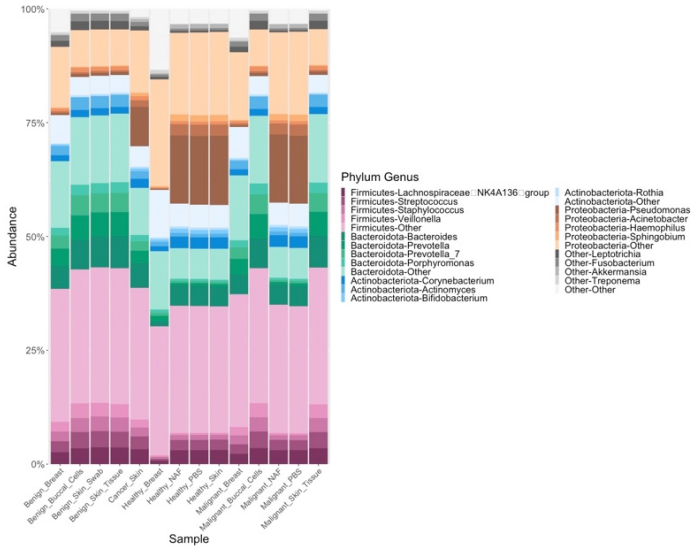

Figure 33. Proportional abundance plot at the phylum and genus levels of normalized data – type and tissue. Comparing pooled samples across all studies. The benign_breast samples consist of Hieken benign breast tissue samples combined with Urbaniak benign breast tissue and similar with the malignant breast tissue. The healthy_skin samples consist of Hieken benign skin swabs and Chan healthy skin swabs, and the cancer skin swabs are also combined from both studies for the cancer_skin samples.

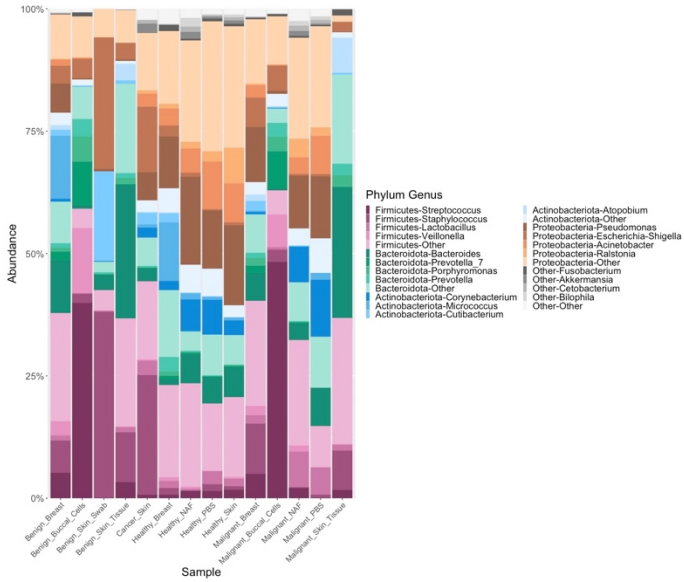

Figure 34. Proportional abundance plot at the phylum and genus levels of pre-normalized data – type and tissue. Comparing pooled samples across all studies. The benign_breast samples consist of Hieken benign breast tissue samples combined with Urbaniak benign breast tissue and similar with the malignant breast tissue. The healthy_skin samples consist of Hieken benign skin swabs and Chan healthy skin swabs, and the cancer skin swabs are also combined from both studies for the cancer_skin samples.

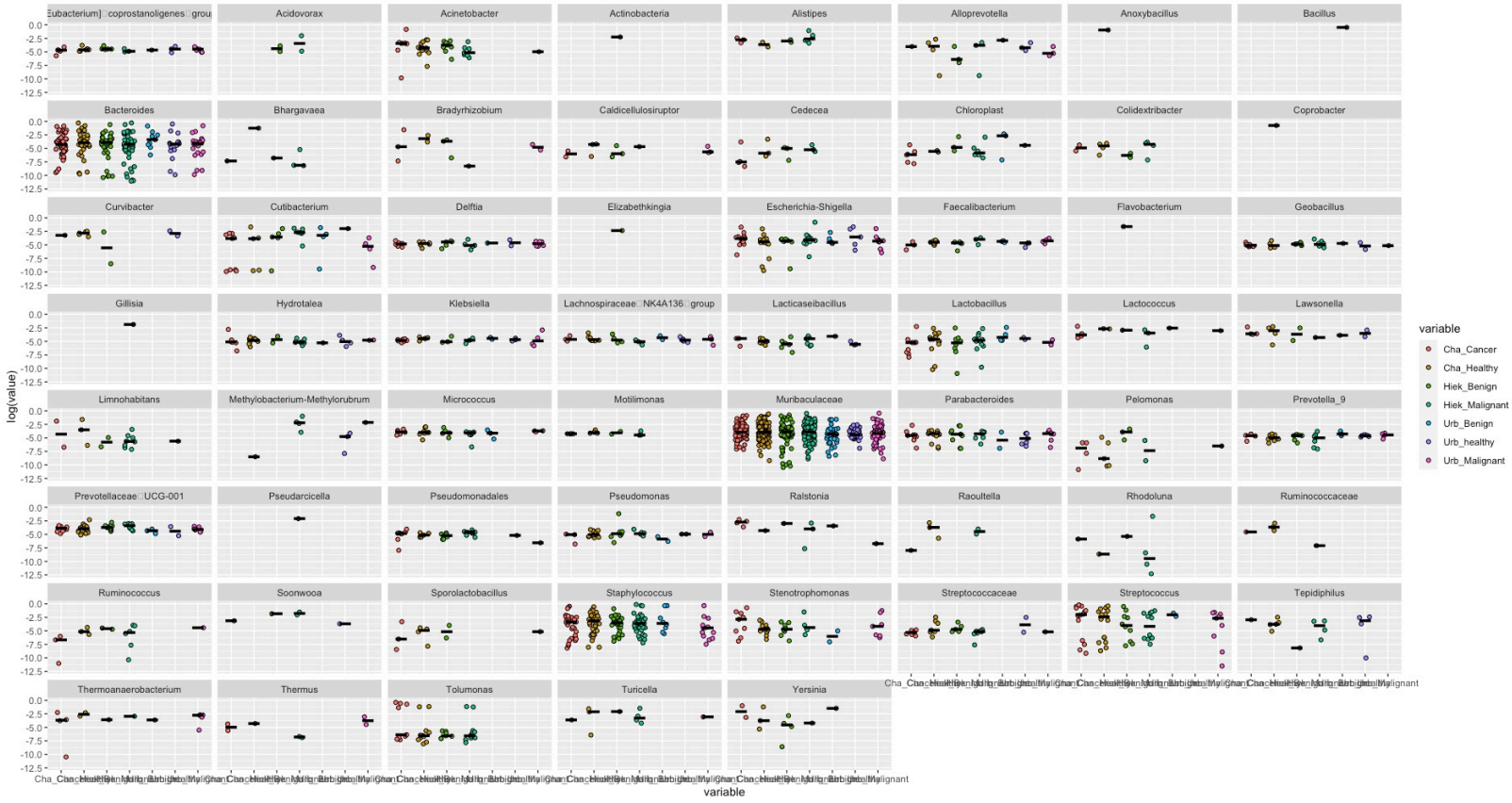

Figure 35. Kruskal-Wallis: significant taxa for pre-normalized data.
Dot plots of the significant taxa found in the pre-normalized data where the x-axis are the samples and y-axis are the logscale of the abundance data. The most abundant taxa are shown to be *Muribaculaceae* and *Bacteroides* across the studies based on the dot plots.

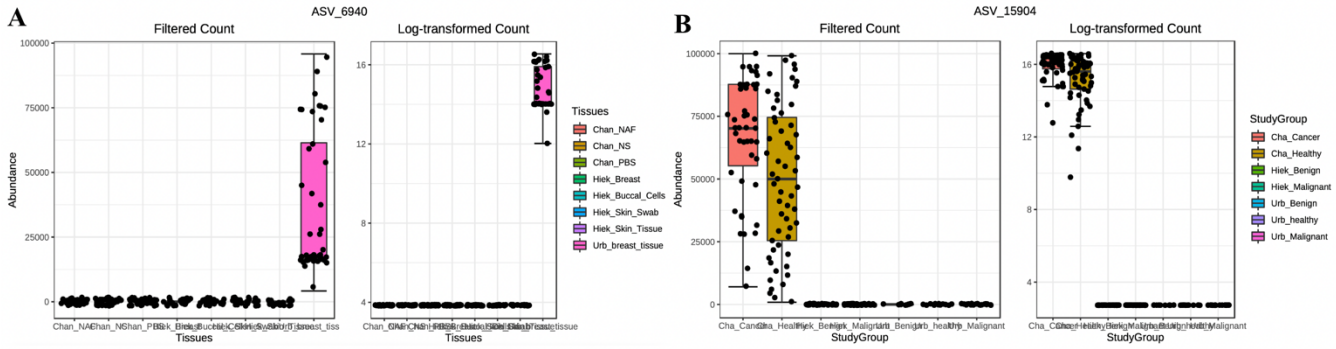

Figure 36. Comparison of all samples across the three studies. (A) Comparing tissues across the three studies. (B) Comparing disease states across the three studies.

In order to further look at differentially abundant taxa in the normalized data, the significant taxa were identified based on adjusted p-values from Classical Univariate, metagenomeSeq, edgeR, and LEfSe analyses in MicrobiomeAnalyst (Dhariwal et al., 2017) for comparing across all tissue types and disease states. The significant taxa were visualized through box plots and they show the significance being separated by study when comparing across all samples from all three studies (Fig. 36). When comparing all samples from the three studies at once, there were thousands of taxa identified to be significant; hence, we then compared only skin swab samples and breast tissue samples for specific studies. The significant taxa were also identified based on adjusted p-values from Classical Univariate, metagenomeSeq, edgeR, and LEfSe analyses in MicrobiomeAnalyst (Dhariwal et al., 2017) for comparing breast tissue and skin swab microbiota. There were no significant taxa found when comparing skin swabs from Hieken and Chan datasets from any analyses in MicrobiomeAnalyst. There were five significant taxa found when comparing breast tissue from Hieken and Urbaniak datasets, of which two were from metagenomeSeq analysis and three were from edgeR analysis. In metagenomeSeq the genera *Lactococcus* and *Agathobacter* were identified as significant (Fig. 37A) and in edgeR the

genera *Lactococcus*, *Agathobacter*, and *Prevotella_9* were identified as significant (Fig. 37B). The boxplot results show that the significance is primarily driven by outliers that are located at the top of each plot (Fig. 37).

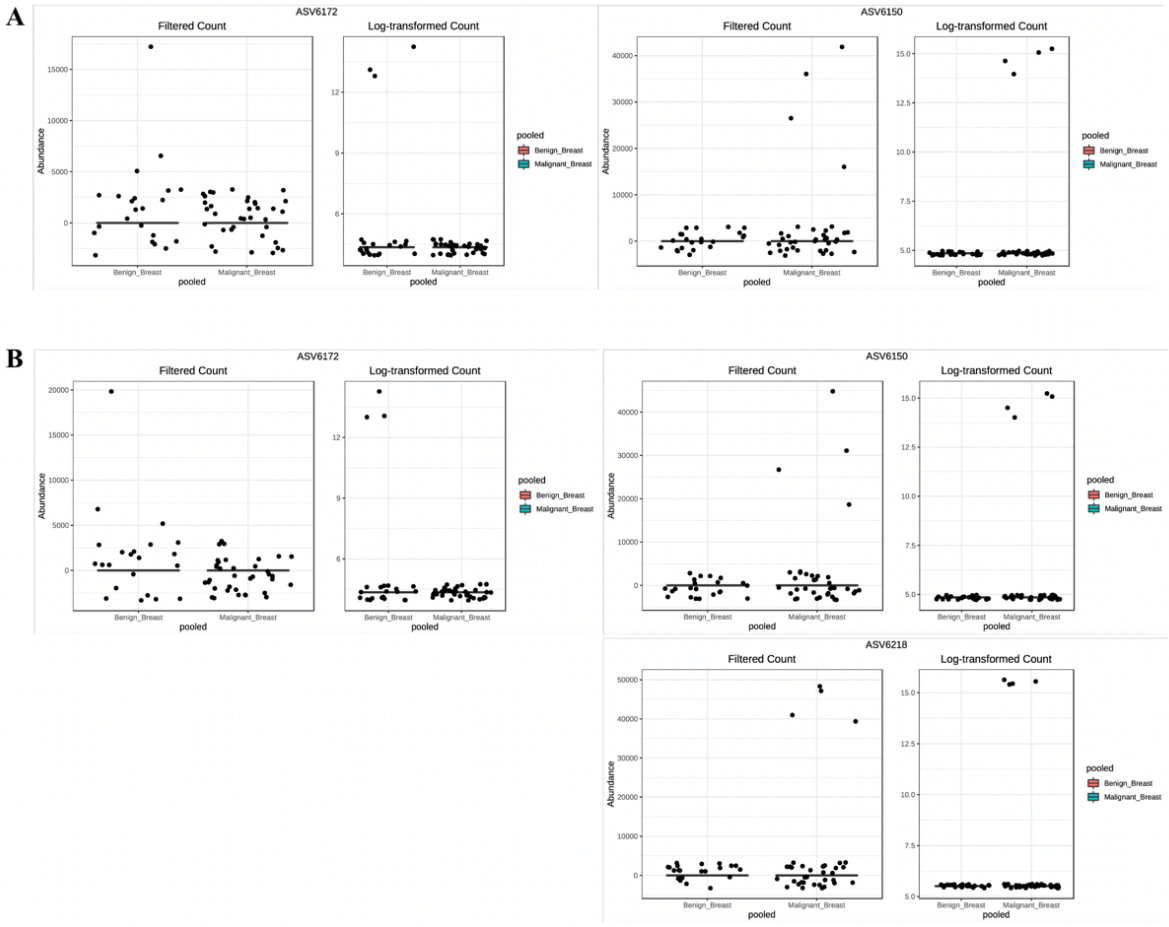

Figure 37. Differentially abundant taxa from metagenomeSeq and edgeR output. (A) The significant taxa from metagenomeSeq output are *Agathobacter* (ASV6172) and *Lactococcus* (ASV6150). (B) The significant taxa from edgeR output are *Agathobacter* (ASV6172), *Lactococcus* (ASV6150), and *Prevotella_9* (ASV6128).

Discussion and Conclusion

The different types of analyses performed on the pre-normalized and normalized data show that, prior to normalization, the results show some overlap in taxonomic composition

amongst similar types of samples across studies, such as the genus *Escherichia-Shigella* being abundant in benign and malignant samples across the Hieken and Urbaniak studies. Similarly, these different analyses also show that, prior to normalization, the results show differentially abundant significant taxa for the benign, malignant, and healthy samples across studies. However, after normalizing the data, the results show that the taxonomic composition varies by study and that there are no differentially abundant significant taxa identified from analyses that compare samples between studies. The differences between the pre-normalized and normalized data are due to the fact that normalization helps to reduce study-specific biases. While study-specific biases were not completely removed in our meta-analysis, the study-specific biases were greatly reduced. This is clear in the proportional abundance plots as there is greater variability in the pre-normalized abundance plots, and similarly, in the beta diversity analyses, the samples separate by study in the pre-normalized data whereas there is separation by sample in the normalized data.

In the breast tissue and skin swab analyses from MicrobiomeAnalyst, there were some significant taxa identified from metagenomeSeq and edgeR. In the skin swab analyses, there were no significant taxa found. This is what was found in the results from each individual study. There were no significant taxa found for skin swabs from the Hieken study's original results, and no significant taxa were found for skin swabs from the Chan study's original results. In the breast tissue analyses, there were significant taxa identified from metagenomeSeq and edgeR. This is similar to what was found in the original Urbaniak study's results. In the Urbaniak study, they implemented ALDEx2 (Gloor et al., 2022) for differential abundance analysis and identified ten taxa to be significant in breast tissue between cancer and healthy samples. The results from

metagenomeSeq and edgeR identified *Lactococcus*, *Agathobacter*, and *Prevotella_9* as significant where only edgeR identified *Prevotella_9* as significant and *Lactococcus and Agathobacter* were identified as significant by both metgenomeSeq and edgeR. These three genera are also identified as significant in the ALDEx2 output from our re-analysis of the Urbaniak data (Fig. 25).

The Urbaniak study's original results also identified *Lactococcus* as significant, and there are no significant taxa from the Hieken study's original results that overlap with results from metagenomeSeq and edgeR. A reason that the Hieken study's results may not overlap with our results from metagenomeSeq and edgeR is that the Urbaniak study's patient sample size is much larger than that of the Hieken study's patient sample size. There were 71 women in the Urbaniak study and 33 patients in the Hieken study; hence, the larger patient sample size of the Urbaniak study may be the reason that it is well-represented in our metagenomeSeq and edgeR results.

CHAPTER FOUR

DISCUSSION AND CONCLUSION

The individual studies implemented tools, pipelines, and methods that are outdated and are not currently used for bioinformatic analyses, such as IM-TORNADO and the Greengenes reference database. While we have observed results that are similar to that of each original study's results such as when distinguishing between cancer and healthy samples, there are differences present. Our results have identified more taxa than the original studies and have shown that some taxa identified to be significant in the original paper are not found to be significant when re-analyzed with more recent and up-to-date techniques and methods. As mentioned in the introduction, differences in both the bioinformatic pipeline and reference database can contribute to differences in taxonomic assignment and downstream analyses results. The methods implemented in our re-analysis are the updated and known to perform better than their older counterparts. The ASV table approach through DADA2 provides greater resolution and lowers false positives (Prodan et al., 2020), and the SILVA database is known to perform better than the Greengenes database (Almeida et al., 2018).

Additionally, the Hieken et al., Chan et al., and Urbaniak et al. studies are seminal papers in the breast cancer microbiome field, and these studies were not able to get a complete and accurate picture of the breast microbiome. Through our re-analysis, we are able to improve upon their results using modern best practices and have discovered new phenomenon that are hitherto unknown and are able to correct their mistaken findings. These findings are important as they

will help to elucidate the direction that researchers interested in the breast cancer microbiome should move towards to investigate the appropriate taxa found in the breast microbiome. Therefore, re-analyses of past results and studies are needed to provide the most accurate and reliable results in the scientific community.

In meta-analyses approach it is important to normalize each individual study prior to combining the studies. The normalization removes noise and resolves artifactual results found in the non-normalized microbiome data. Normalization is important in order to remove or reduce study-specific biases to allow for accurate downstream analyses. In the pre-normalized data, there are more overlaps and significance observed across the studies, whereas in the normalized data these overlaps and significant taxa are removed. The overlaps and significant taxa observed in the pre-normalized data are most likely due to artifacts in the microbiome data. The study-specific biases were not completely removed after normalizing the data as study effect was shown in the alpha diversity, proportional abundance, and differential abundance analyses. A reason for that may be because each study has samples from patients in different geographical regions where differences in environment, diet, and lifestyle can affect the microbial composition. The Hieken study has samples collected from patients in Minnesota, the Chan study has samples collected from patients in California, and the Urbaniak study has samples collected from patients in Canada.

Through our meta-analysis, we are combining multiple studies that investigate the breast cancer microbiome and perform statistical and differential abundance analyses to search for biomarkers or significant microbes that appear across studies in cancer or control samples. Through the meta-analysis, we are providing new findings on whether there is a distinct

microbial composition in different samples such as cancer vs control samples. These new findings include that normalization prior to implementing a cross-study meta-analysis is important to reduce study-specific biases and variability as shown through the beta diversity and proportional abundance analyses; hence, is important to remove noise and artifacts from microbiome data. By choosing not to normalize data prior to cross-study meta-analyses, it may lead to identifying superficial similarities and differentially abundant taxa that may be artifactual data identified as being significant. Lastly, through our re-analysis of each dataset, it was found that our re-analysis identified more proportional and differential abundance taxa at the genus level than the original study's results as more taxa have been defined since 2016 in the frequently updated SILVA database and improvements in bioinformatic pipeline analyses have refined results.

REFERENCE LIST

A custom color palette for improving data visualization. (n.d.). Retrieved May 17, 2022, from https://karstenslab.github.io/microshades/

Almeida, A., Mitchell, A. L., Tarkowska, A., & Finn, R. D. (2018). Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. *GigaScience*, *7*(5), giy054. https://doi.org/10.1093/gigascience/giy054

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

American Cancer Society: About Breast Cancer. (n.d.). Key Statistics for Breast Cancer. Retrieved June 21, 2022, from https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html

Babraham Bioinformatics—FastQC A Quality Control tool for High Throughput Sequence Data. (n.d.). Retrieved May 11, 2022, from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Balvočiūtė, M., & Huson, D. H. (2017). SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare? *BMC Genomics*, *18*(Suppl 2), 114. https://doi.org/10.1186/s12864-017-3501-4

Bolar, K. (2019). *stat: Interactive Document for Working with Basic Statistical Analysis*. https://cran.r-project.org/web/packages/STAT/STAT.pdf

Boogart, K. G. van den, Tolosana-Delgado, R., & Bren, M. (2022). *compositions: Compositional Data Analysis*. https://cran.r-project.org/web/packages/compositions/compositions.pdf

Bukin, Y. S., Galachyants, Y. P., Morozov, I. V., Bukin, S. V., Zakharenko, A. S., & Zemskaya, T. I. (2019). The effect of 16S rRNA region choice on bacterial community metabarcoding results. *Scientific Data*, *6*(1), 190007. https://doi.org/10.1038/sdata.2019.7

Burns, M. B., Lynch, J., Starr, T. K., Knights, D., & Blekhman, R. (2015). Virulence genes are a signature of the microbiome in the colorectal tumor microenvironment. *Genome Medicine*, *7*(1), 55. https://doi.org/10.1186/s13073-015-0177-8

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

Callahan, B., McMurdie, P., & Holmes, S. (2022). *dada2: Accurate, high-resolution sample inference from amplicon sequencing data*. https://bioconductor.org/packages/release/bioc/manuals/dada2/man/dada2.pdf

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*(5), 335–336. https://doi.org/10.1038/nmeth.f.303

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

Chan, A. A., Bashir, M., Rivas, M. N., Duvall, K., Sieling, P. A., Pieber, T. R., Vaishampayan, P. A., Love, S. M., & Lee, D. J. (2016). Characterization of the microbiome of nipple aspirate fluid of breast cancer survivors. *Scientific Reports*, *6*(1), 28061. https://doi.org/10.1038/srep28061

Chen, J., Douglass, J., Prasath, V., Neace, M., Atrchian, S., Manjili, M. H., Shokouhi, S., & Habibi, M. (2019). The microbiome and breast cancer: A review. *Breast Cancer Research and Treatment*, *178*(3), 493–496. https://doi.org/10.1007/s10549-019-05407-5

Chen, J., Zhang, X., & Yang, L. (2022). *GUniFrac: Generalized UniFrac Distances, Distance-Based Multivariate Methods and Feature-Based Univariate Methods for Microbiome Data Analysis*. https://cran.r-project.org/web/packages/GUniFrac/GUniFrac.pdf

Cole, J. R., Chai, B., Farris, R. J., Wang, Q., Kulam, S. A., McGarrell, D. M., Garrity, G. M., & Tiedje, J. M. (2005). The Ribosomal Database Project (RDP-II): Sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Research*, *33*(suppl_1), D294–D296. https://doi.org/10.1093/nar/gki038

Computational Biology Core—Brown University. (n.d.). Retrieved May 17, 2022, from https://compbiocore.github.io/metagenomics-workshop/

Costantini, L., Magno, S., Albanese, D., Donati, C., Molinari, R., Filippone, A., Masetti, R., & Merendino, N. (2018). Characterization of human breast tissue microbiota from core needle biopsies through the analysis of multi hypervariable 16S-rRNA gene regions. *Scientific Reports*, *8*(1), 16893. https://doi.org/10.1038/s41598-018-35329-z

DADA2 1.16 Pipeline. (n.d.). DADA2 Pipeline Tutorial (1.16). https://benjjneb.github.io/dada2/tutorial.html.

Dhariwal, A., Chong, J., Habib, S., King, I. L., Agellon, L. B., & Xia, J. (2017). MicrobiomeAnalyst: A web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic Acids Research*, *45*(Web Server issue), W180–W188. https://doi.org/10.1093/nar/gkx295

Durazzi, F., Sala, C., Castellani, G., Manfreda, G., Remondini, D., & De Cesare, A. (2021). Comparison between 16S rRNA and shotgun sequencing data for the taxonomic characterization of the gut microbiota. *Scientific Reports*, *11*(1), 3030. https://doi.org/10.1038/s41598-021-82726-y

Edgar, R. C. (2004). MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, *5*(1), 113. https://doi.org/10.1186/1471-2105-5-113

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

Emmanuel Paradis, Simon Blomberg, Ben Bolker, Joseph Brown, Santiago Claramunt, Julien Claude, Hoa Sien Cuong, Richard Desper, Gilles Didier, Benoit Durand, Julien Dutheil, RJ Ewing, Olivier Gascuel, Thomas Guillerme, Christoph Heibl, Anthony Ives, Bradley Jones, Franz Krah, Daniel Lawson, … Damien de Vienne. (2022). *Ape: Analyses of Phylogenetics and Evolution*.

Ewels, P., Duncan, A., & Fellows Yates, J. (n.d.). *SRA-Explorer*. SRA-Explorer. https://sra-explorer.info/

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

Gloor, G., Fernandes, A., Macklaim, J., Albert, A., Links, M., Quinn, T., Wu, J. R., Wong, R. G., & Lieng, B. (2022). *ALDEx2: Analysis Of Differential Abundance Taking Sample Variation Into Account*. https://bioconductor.org/packages/release/bioc/manuals/ALDEx2/man/ALDEx2.pdf

Hieken, T. J., Chen, J., Hoskin, T. L., Walther-Antonio, M., Johnson, S., Ramaker, S., Xiao, J., Radisky, D. C., Knutson, K. L., Kalari, K. R., Yao, J. Z., Baddour, L. M., Chia, N., & Degnim, A. C. (2016). The Microbiome of Aseptically Collected Human Breast Tissue in Benign and Malignant Disease. *Scientific Reports*, *6*. https://doi.org/10.1038/srep30751

Jari Oksanen, Gavin L. Simpson, F. Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R. Minchin, R.B. O'Hara, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs, Helene Wagner, Matt Barbour, Michael Bedward, Ben Bolker, Daniel Borcard, Gustavo Carvalho, Michael Chirico, Miquel De Caceres, Sebastien Durand, … James Weedon. (2022). *Vegan: Community Ecology Package*. https://cran.r-project.org/web/packages/vegan/vegan.pdf

Jeraldo, P. (2020). *IM-TORNADO: A pipeline for 16S reads from paired-end libraries* [Shell]. https://github.com/pjeraldo/imtornado2 (Original work published 2016)

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, *28*(1), 27–30.

Kruskal, W. H., & Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*. https://www.tandfonline.com/doi/abs/10.1080/01621459.1952.10483441

Langille, M. G. I., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., Clemente, J. C., Burkepile, D. E., Vega Thurber, R. L., Knight, R., Beiko, R. G., & Huttenhower, C. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology*, *31*(9), 814–821. https://doi.org/10.1038/nbt.2676

Lozupone, C., & Knight, R. (2005). UniFrac: A New Phylogenetic Method for Comparing Microbial Communities. *Applied and Environmental Microbiology*, *71*(12), 8228–8235. https://doi.org/10.1128/AEM.71.12.8228-8235.2005

Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J., & Knight, R. (2011). UniFrac: An effective distance metric for microbial community comparison. *The ISME Journal*, *5*(2), 169–172. https://doi.org/10.1038/ismej.2010.133

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10–12. https://doi.org/10.14806/ej.17.1.200

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., Andersen, G. L., Knight, R., & Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal*, *6*(3), 610–618. https://doi.org/10.1038/ismej.2011.139

McLaren, M. R., & Callahan, B. J. (2021). *Silva 138.1 prokaryotic SSU taxonomic training data formatted for DADA2* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.4587955

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

Mikryukov, V. (2018). *vmikk/metagMisc: V.0.0.4*. Zenodo. https://doi.org/10.5281/zenodo.1172500

Mira-Pascual, L., Cabrera-Rubio, R., Ocon, S., Costales, P., Parra, A., Suarez, A., Moris, F., Rodrigo, L., Mira, A., & Collado, M. C. (2015). Microbial mucosal colonic shifts associated with the development of colorectal cancer reveal the presence of different bacterial and archaeal biomarkers. *Journal of Gastroenterology*, *50*(2), 167–179. https://doi.org/10.1007/s00535-014-0963-x

MiSeq SOP. (n.d.). Https://Mothur.Org. Retrieved April 26, 2021, from https://mothur.org

Price, M. N., Dehal, P. S., & Arkin, A. P. (2009). FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution*, *26*(7), 1641–1650. https://doi.org/10.1093/molbev/msp077

Prodan, A., Tremaroli, V., Brolin, H., Zwinderman, A. H., Nieuwdorp, M., & Levin, E. (2020). Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *PLOS ONE*, *15*(1), e0227434. https://doi.org/10.1371/journal.pone.0227434

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

Ranjan, R., Rani, A., Metwally, A., McGee, H. S., & Perkins, D. L. (2016). Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochemical and Biophysical Research Communications*, *469*(4), 967–977. https://doi.org/10.1016/j.bbrc.2015.12.083

Rousseeuw, P., Struyf, A., & Hubert, M. (2022). *cluster: Methods for Cluster analysis*. https://cran.r-project.org/web/packages/cluster/cluster.pdf

Schliep, K., Paradis, E., Martins, L. de O., Potts, A., White, T. W., Stachniss, C., Kendall, M., Halabi, K., Bilderbeek, R., Winchell, K., Revell, L., Gilchrist, M., Beaulieu, J., O'Meara, B., & Qu, L. (2021). *phangorn: Phylogenetic Reconstruction and Analysis*. https://cran.r-project.org/web/packages/phangorn/phangorn.pdf

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Horn, D. J. V., & Weber, C. F. (2009). Introducing mothur: Open-

Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology*, *75*(23), 7537–7541. https://doi.org/10.1128/AEM.01541-09

Sequence Match. (n.d.). Retrieved June 22, 2022, from http://rdp.cme.msu.edu/seqmatch/seqmatch_intro.jsp

Sequence Read Archive (SRA). (n.d.). National Center for Biotechnology Information, National Library of Medicine (US). https://www.ncbi.nlm.nih.gov/sra/

Urbaniak, C., Gloor, G. B., Brackstone, M., Scott, L., Tangney, M., & Reid, G. (2016). The Microbiota of Breast Tissue and Its Association with Breast Cancer. *Applied and Environmental Microbiology*, *82*(16), 5039–5048. https://doi.org/10.1128/AEM.01235-16

Xuan, C., Shamonki, J. M., Chung, A., DiNome, M. L., Chung, M., Sieling, P. A., & Lee, D. J. (2014). Microbial Dysbiosis Is Associated with Human Breast Cancer. *PLOS ONE*, *9*(1), e83744. https://doi.org/10.1371/journal.pone.0083744

Ye, S. H., Siddle, K. J., Park, D. J., & Sabeti, P. C. (2019). Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell*, *178*(4), 779–794. https://doi.org/10.1016/j.cell.2019.07.010

Zhang, X., Chen, J., & Zhou, H. (2022). *MicrobiomeStat: Statistical Methods for Microbiome Compositional Data*. https://cran.r-project.org/web/packages/MicrobiomeStat/MicrobiomeStat.pdf

Zhao, N., Chen, J., Carroll, I. M., Ringel-Kulka, T., Epstein, M. P., Zhou, H., Zhou, J. J., Ringel, Y., Li, H., & Wu, M. C. (2015). Testing in Microbiome-Profiling Studies with MiRKAT, the Microbiome Regression-Based Kernel Association Test. *American Journal of Human Genetics*, *96*(5), 797–807. https://doi.org/10.1016/j.ajhg.2015.04.003

Zhou, H., He, K., Chen, J., & Zhang, X. (2022). LinDA: Linear models for differential abundance analysis of microbiome compositional data. *Genome Biology*, *23*(1), 95. https://doi.org/10.1186/s13059-022-02655-5

VITA

Sidra Sohail earned her Bachelor of Science in Molecular Biology with minor in Biostatistics, *magna cum laude* in May 2020 at Loyola University Chicago. Sohail started working on cancer microbiome research in the Burns Lab in 2017 and was twice awarded the Mulcahy Fellowship from 2018-2020 for her work in analyzing colorectal and gastric cancer datasets using bioinformatic tools and pipelines and was nominated for an Outstanding Undergraduate Student Researcher Award. During her undergraduate years, Sohail was a supplemental instructor, tutor, student leader, and a volunteer at Ann and Robert H. Lurie Children's Hospital of Chicago. She continued her research in the Burns lab while she pursued the Master's in Bioinformatics degree program at Loyola University Chicago. Moving forward, Sohail will begin her professional career as a scientist in the metagenomics and bioinformatics fields.