Mining Public Microbiome Datasets to Identify Specific Microbial Taxa Associated with Anxiety-Related Disorders

by
Austin Martin

A THESIS

submitted to

Oregon State University

Honors College

in partial fulfillment of
the requirements for the
degree of

Honors Baccalaureate of Science in Microbiology
(Honors Scholar)

Presented May 26, 2019
Commencement June 2020

AN ABSTRACT OF THE THESIS OF

Austin Martin for the degree of Honors Baccalaureate of Science in Microbiology presented on May 26, 2020. Title: Mining Public Microbiome Datasets to Identify Specific Microbial Taxa Associated with Anxiety-Related Disorders.

Abstract approved:_____

Maude David

Over the last ten years, clinical, pre-clinical and animal studies have shown associations between the microbiota and neurological functions. Recent work by the scientific community on the gut-microbiome-brain axis have revealed that gut dysbiosis and specific microbial taxa are associated with a myriad of neurological conditions, such as autism spectrum disorder (ASD), anxiety and depression. Many neurological conditions are associated with symptoms of anxiety and stress or are associated with comorbid disorders to anxiety which also correlate with alterations in gut-microbiota composition. In order to study the gut microbiota of an individual, the scientific community widely uses 16S amplicon sequencing of stool samples and comparison of these sequences to a database for taxonomic classification. While 16S amplicon analysis can be limited as it only probes for a small portion of the genomes of microbiota (and subsequently taxonomic classification is usually limited to the genus level), 16S analysis is a widely used means of classifying taxa within the gut microbiota, and a large amount of this data has been made available in public databases.

In this project, we leverage available data from multiple studies to determine a common set of 16S amplicons associated with anxiety. To perform this meta-analysis, we used new methods of 16S analysis that have been developed in recent years to extract exact amplicon variants, allowing better comparison of markers across studies. In addition, as many studies in the realm of the gut-brain axis suffer from extremely small samples size, we address these

issues with a meta-analysis of 1266 samples containing individuals with anxiety, depression, autism, and ADHD from three studies conducted using a DADA2 pipeline with DESeq2, Metagenomeseq, and ANCOM as a means of differential analysis between individuals with anxiety-related conditions and the neurotypical. Eight different amplicon sequence variants (ASVs) were significant in Metagenomeseq and ANCOM across all datasets. 32 other variants were significant when random subsets of the data were analyzed using similar means. When these both of these ASV groups were used as predictors in a random forest model (using 10 fold cross validation), these ASVs allowed the model to perform better than random at 56% for the eight ASVs found across datasets and greater than 63% when using predictor ASVs found among the random subsets. This study reveals the potential significance of microbial biomarkers of anxiety identified across several studies, identifies significant taxa with the benefits of meta-analyses, and demonstrates the effects these taxa have classification in random forest models.

Key Words: Gut-microbiome, anxiety disorder, 16S, bioinformatics

Corresponding e-mail address: martiaus@oregonstate.edu

Mining Public Microbiome Datasets to Identify Specific Microbial Taxa Associated with Anxiety-Related Disorders

by
Austin Martin

A THESIS

submitted to

Oregon State University

Honors College

in partial fulfillment of
the requirements for the
degree of

Honors Baccalaureate of Science in Microbiology
(Honors Scholar)

Presented May 26, 2020
Commencement June 2020

Honors Baccalaureate of Science in Microbiology project of Austin Martin presented on May 26, 2020.

APPROVED:

_____

Maude David, Mentor, representing Microbiology

_____

Lloyd Walter Ream, Committee Member, representing Microbiology

_____

Megan MacDonald, Committee Member, representing Public Health and Human Sciences

_____

Toni Doolen, Dean, Oregon State University Honors College

I understand that my project will become part of the permanent collection of Oregon State University, Honors College. My signature below authorizes release of my project to any reader upon request.

_____

Austin Martin, Author

INTRODUCTION

The gut-brain axis, defined as the bidirectional means of communication between the enteric and central nervous systems, is largely impacted by microbiota with the gut-microbiome (Mayer, Tillisch, and Gupta 2015). Multiple potential mechanisms have been proposed as to how this phenomenon occurs through neural, humoral, and immunal links (Carabottia et al. 2015). The immune system, tryptophan metabolism, the vagus nerve that connects the lumen of the gut to the brain, and metabolite-enteric system interactions have all been proposed as possible routes of communication for this phenomenon (Cryan et al. 2019). Some hypothesize that gut microbiota may affect signals sent through the vagus nerve that could impact regulation in the hypothalamic-pituitary-adrenal (HPA) axis responsible for regulating anxiety and stress responses in the body (Carabottia et al. 2015) Other potential mechanisms involve microbiota influencing intestinal permeability or causing mucosal immune activation, which can trigger inflammation as well as stress responses from the HPA axis (Carabottia et al. 2015).

Intestinal microbiota also influence levels of neurotransmitters such as acetylcholine, serotonin, and GABA or levels of neurotransmitter precursors such as tryptophan (Liu and Zhu 2018). Tryptophan, being an essential amino acid, is mainly obtained through diet and is absorbed through the intestinal epithelium leaving it subject to influence from the gut-microbiome (Gao et al. 2019). It is also used to synthesize serotonin, a key neurotransmitter with anxiety and depression (Albert and Benkelfat 2013). The gut alone produces 90% of the serotonin within the body, and sporulating bacteria influence this production as well as the levels of tryptophan present (Yano et al. 2015). Overall,

neurotransmitters, their precursors, or other compounds such as short-chain fatty acids can affect the activity of cells in the gut such as enterochromaffin cells which can send signals to the brain via the vagus nerve (Martin et al. 2018).

While the exact mechanisms of gut-microbiome and gut-brain axis interactions have not been fully defined, many animal studies have shown association with gut microbiota dysbiosis with anxiety disorders and conditions often comorbid with anxiety. Germ-free mice for example have reduced anxiety-like behavior and neurochemical changes when compared to mice with regular gut-microbiota (Neufeld et al. 2011). Certain probiotic treatments consisting of *Bifidobacterium* and *Lactobacillus* strains can alter levels of anxiety within mice (Martin et al. 2018). Similar findings are beginning to be discovered in humans as well across a variety of neurological conditions. Neurological disorders such as depression, anxiety, and autism have been shown to be associated with gut-microbiome community changes or probiotic treatments in both human and animal models (Foster and McVey Neufeld 2013; Vuong and Hsiao 2017).

Our meta-analysis involves individuals with anxiety, and comorbid disorders to anxiety such as depression, ADHD, and Autism Spectrum Disorder (ASD) from three different studies (Hill-Burns et al. 2017; McDonald et al. 2018; Kang et al. 2017). These conditions and disorders were grouped since they are all commonly comorbid with anxiety, and have shown association with gut microbiota dysbiosis. People with ASD have a higher prevalence of depressed or anxious symptoms than the general population (Strang et al. 2012). ADHD is also associated with comorbid anxiety, and there is some overlap in the diagnostic criteria for the two conditions (Bilgiç et al. 2013). Depression and anxiety are also

commonly comorbid with one another with one study showing that 56.8% of people with Major Depressive Disorder had an anxiety disorder as well (Zhou et al. 2017). We grouped all individuals with these diagnoses under one group we referred to as individuals with anxiety-related conditions.

One popular method of analyzing gut-microbiome communities is through the analysis of 16S gene amplicon within microbial genomes found in stool samples. 16S analysis is done in a variety of settings to characterize microbial communities because of its ubiquitous nature. It can be used as a reliable molecular clock even across distant prokaryotes since the ribosomal gene is evolutionarily conserved (Case et al. 2007). The gene itself is divided into highly conserved regions and hypervariable regions. There are 9 total hypervariable regions referred to as V1 through V9. Some regions vary too much to be reliable for classification while others may vary too little to distinguish taxa from one another. The 16S hypervariable regions V3, V4 and V5 have been widely used in the literature as they are the most representative of the full-length 16S rRNA, and therefore the most reliable use for taxonomic classification across bacterial phyla (Yang, Wang, and Qian 2016). Because of its wide usage, the scientific community has established multiple databases considering full length of the gene and/or variable regions, and 16S sequences can be traced against known databases in order to classify the sequence taxonomically.

While these methods are widely used, studies on the gut-microbiome and neurological conditions are often limited in their findings due to small sample sizes or their means of 16S analysis. Technology within high-throughput sequencing has changed drastically in recent years, as well as the way in which we analyze and handle these data. The

original method of 16S analysis involves clustering, which groups sequences into sequences that are 97% similar to one another (also called Operational Taxonomic unit, or OTUs). Many pipelines such as Qiime or Mothur employ clustering strategies based on this level of percent similarity (97%) (Kuczynski et al. 2005; Schloss et al. 2009) The ultimate goal of this clustering of similar sequences is to create groups of sequences that are almost identical, likely due to actual taxonomic variation, but also allow approximate sequencing errors by grouping the majority of the sequences into a cluster and considering only the consensus sequences of this cluster. Instead of clustering, newer methods involve directly taking count of sequences on an individual basis without grouping by similarity and the sequences are organized into sequence amplicon variants (i.e. considering sequences 100% similar to each other). These new methods employ denoising algorithms and models that correct for sequencing errors. Two newer methods of processing and denoising sequence data that use quality data to correct for sequencing errors are DADA2 and deblur (Benjamin J. Callahan et al. 2015; Amir et al. 2017). For this meta-analysis, we chose DADA2. DADA2 is different from clustering strategies in that it uses the quality data contained in fastq files to estimate particular error rates within each Illumina run (Benjamin J. Callahan et al. 2015). This allows the pipeline to identify and correct sequencing errors caused by the sequencing instrument. Therefore, DADA2 does not group sequences into clusters by 97% similarity, but instead allows each sequence to be counted individually as an ASV (Amplicon Sequence Variant) due to DADA2's increased capability to determine if two closely similar sequences are different due to sequence error or are actually two distinct sequences. By being able to correct the sequences and determine these ASVs, reproducible variants can be found across

different studies, which can not be done with clusters since clustering can change between studies.

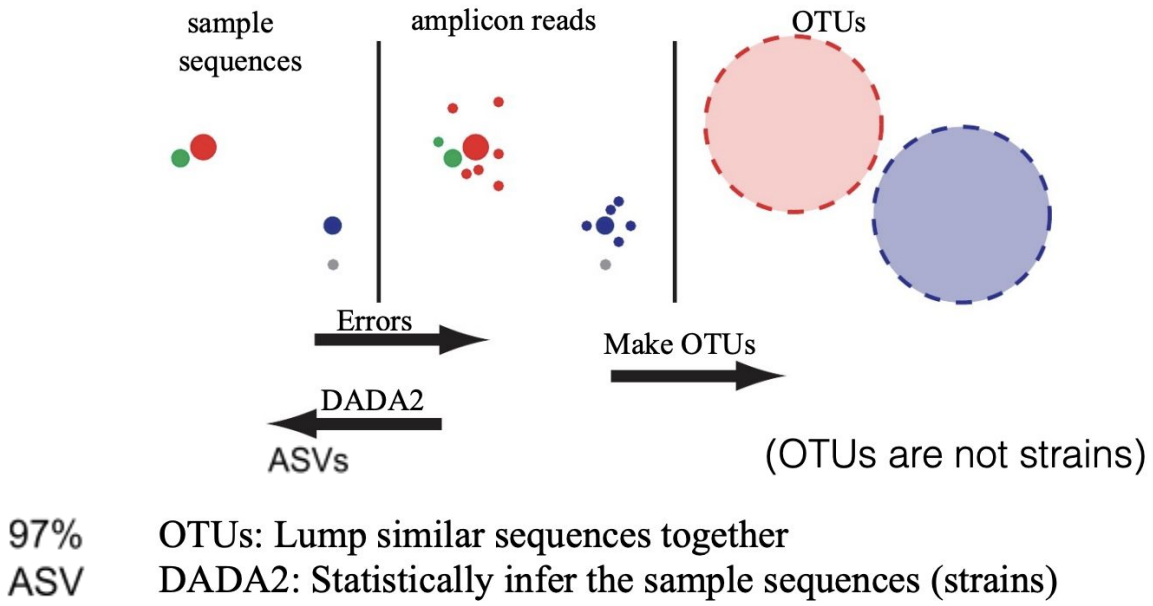


Diagram 1 Amplicon Sequence Variants (ASVs) vs. Operational Taxonomic Units (OTUs) showing how sequences are inferred from noisy reads

Figure made available by Susan Holmes at Stanford University, teaching website. As sequences from samples get sequences and organized into reads, errors in the sequence are generated due to sequence machinery. DADA2 attempts to correct for those errors to have sequence data closer to the true sample sequences. Making OTUs, or clustering, as seen above takes reads containing sequencing errors and groups them by similarity, so that slight errors are insignificant within the group

Along with advances in denoising pipelines, multiple means of differential analysis have emerged to allow the identification of taxa that significantly differ in abundance between cohorts. Common methods of multivariate analysis such as Principal Coordinates Analysis (PcoA) that create coordinates for plots based on abundance count differences as well as subsequent Permanova tests still remain standard tools for community

characterization as a whole (Anderson 2017). However, new methods have risen within the field that allow for analyses that help determine whether specific ASVs or OTUs are significantly different between cohorts. Metagenomeseq, for instance, is a method that addresses common issues of analysis of 16S data in human microbial communities, such as undersampling, and employs a normalization strategy that is more data-driven (Paulson et al. 2013). Other strategies such as ANCOM, a method focused on reducing false positives within microbiome differential analysis, therefore producing robust results (Mandal et al. 2015). Finally, DESEQ2, a method focused on providing a more quantitative analysis of ASV or OTU count differences, also exists as a useful tool to address differences between groups of samples (Love, Huber, and Anders 2014).

The goal of this study was to overcome the issue of low sample sizes and reproducibility across studies by conducting a meta-analysis of multiple studies using DADA2 as a means of sequence processing in order to identify specific taxa (that will be designated as Amplicon Sequence Variant, or ASVs) that are associated with anxiety-related conditions within the gut-microbiome. We expected to find multiple ASVs would be significantly different in counts between the anxious and non-anxious phenotype, and that the gut-microbiome communities would be different in structure after metadata such as age and sex were accounted for. Each of these studies we drew samples from also used OTU clustering at 97%. In addition to the benefits of a larger sample size, our aim of this meta-analysis was to use a full DADA2 pipeline (which employs an exact sequence variant calling method) to identify novel gut-microbial taxa associated with the aforementioned anxiety-related conditions using the three methods of differential analysis listed above:

DESEQ2, ANCOM, and Metagenomeseq. We also hypothesized that we would find ASV biomarkers that would be effective in distinguishing between the two cohorts. With this in mind, we hypothesized that these associated taxa could be used to help classify individuals between the anxious and neurotypical by using a random forest machine learning algorithm. Random forest models in past studies have been trained on gut-microbiome data to more accurately classify individuals with conditions such as colorectal cancer and fibrosis in fatty liver disease (Ai et al. 2019; Loomba et al. 2017). We expected that taxa found significantly different between the anxious and neurotypical phenotypes could be used in a random forest model in a similar manner for classification of anxiety-related conditions.

METHODS

1. Sample Collection

Sequence data from three different studies were downloaded through the NCBI online public database and QIITA. 1226 samples out of the 1586 samples downloaded were kept and used for analysis after filtering out samples with less than 5000 reads and after balancing the dataset so that there were an equal number of samples between the anxious cohort and the neurotypical. Data accession numbers for sequence information and samples can be found in the Data accession section. Information about whether the individuals had anxiety, depression, ADHD, or ASD and other diagnoses indicated by the paper were collected. Some studies had to be emailed to obtain their metadata. Age, sex, and antibiotics usage within the last 6 months were also used as metadata since this information was available across the three studies. Information about Parkinson's disease was collected from the study by Hill due to it's high prevalence within that dataset and it's potential association with the

gut-microbiome found by that study. All sequences were from the 16S V4 region and were sequenced with an Illumina Miseq machine. The following table, *Table 1* shows a summary of the samples used in this meta-analysis.

Table 1 Summary of Sample Data after Balancing for Phenotype

"HC" stands for Healthy Controls. Age is given in years. All other numbers are representative of the amount of samples in each respective category.

Study	Sample Count	Average Age	Sex Distribution	Diagnoses
AGP	1022	45.27	Male: 460 Female: 562	Depression: 349 ADHD: 105 ASD: 57 HC: 511
Hill	166	67.96	Male: 92 Female: 74	Depression: 17 Anxiety: 25 Anx+Dep: 41 HC: 83
Kang	38	11.08	Male: 34 Female: 4	ASD: 18 HC: 20

2. DADA2, Dataset Balancing, and Differential Analysis

Full code used in this meta-analysis can be found at Github at the following link: https://github.com/MaudeDavidLab/Meta_analysis. Samples were separated by each study for input into the DADA2 pipeline. Filtering and truncating were done using mostly default parameters, and reads with higher than two expected errors were discarded. However, sequences from the study by Kang were truncated by 1 bp in order to match the amplicon length of the other two studies (150bp). After using the DADA2 machine error learning algorithm to estimate error rates within each study, the sequences were dereplicated by grouping identical reads into unique sequences (Benjamin J. Callahan et al. 2015). The

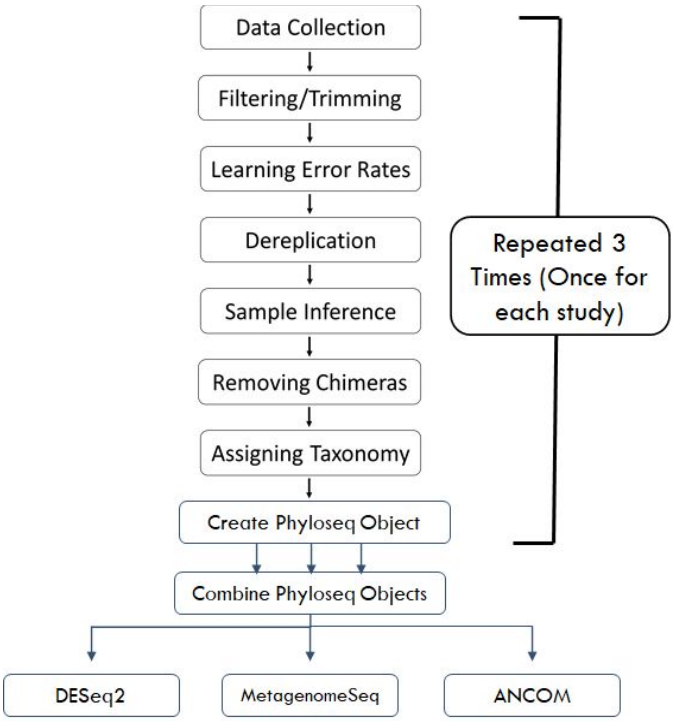
number of reads of each unique sequence and their associated quality scores were retained. The sequence data and their corresponding quality scores were used by the DADA2 core algorithm in order to determine the ASVs present in each study and their respective counts. Chimeras within the ASVs were removed through a function within DADA2 that searches for any ASVs that can be exactly linked to a combination of other ASVs.. The ASVs were assigned a taxonomy by using the Silva 16S database (version 132). A link to the Silva database as well as the basic steps of processing with DADA2 can be found at https://benjjneb.github.io/dada2/tutorial_1_8.html.

Taxonomic information, ASV counts, and metadata from the samples were all combined into a phyloseq object for each study. After proper formatting of metadata, the phyloseq objects created for each study were combined into one. Samples with less than 5000 reads were omitted from the study due to low depth. Samples were then divided into two groups; individuals with anxiety-related conditions (e.g anxiety, depression, ADHD, or ASD) and the neurotypical.

To create a balanced dataset for differential analysis, each individual in the anxious phenotype was paired with one control of similar age (within 3-4 years). Extra controls were omitted except from the Kang study which had a limited amount of samples and only two extra controls. This lowered the sample size from 1582 to 1226.

The balanced dataset consisting of 1266 samples was analyzed with DESeq2, Metagenomeseq, and ANCOM to find ASVs that were significantly different in abundance counts between the anxious and neurotypical phenotypes. The following diagram on page 10 summarizes the DADA2 pipeline and differential analysis process. In addition, a power

analysis using Dirichlet-multinomial distributions were performed. Resulting alpha values for various sample sizes were plotted for each study.



Flow Chart 1 Meta-analysis Workflow for DADA2 and Differential Analysis

3. Identifying Covariates Impacting Microbial Community

DESeq2 normalization and Cumulative sum scaling (CSS) normalization was each done individually on the phyloseq object containing the entire dataset resulting in two different normalized versions of the data. DESeq2 normalization uses variance normalization to stabilize counts (Love, Huber, and Anders 2014). CSS normalization uses a zero-inflated gaussian model to account for under-sampling or abundance differences brought from sample-specific bias (Joseph et al. 2013). Hence, Permanova and the creation of PCOA plots were done twice: once with DESeq2 normalization and another time with CSS normalization.

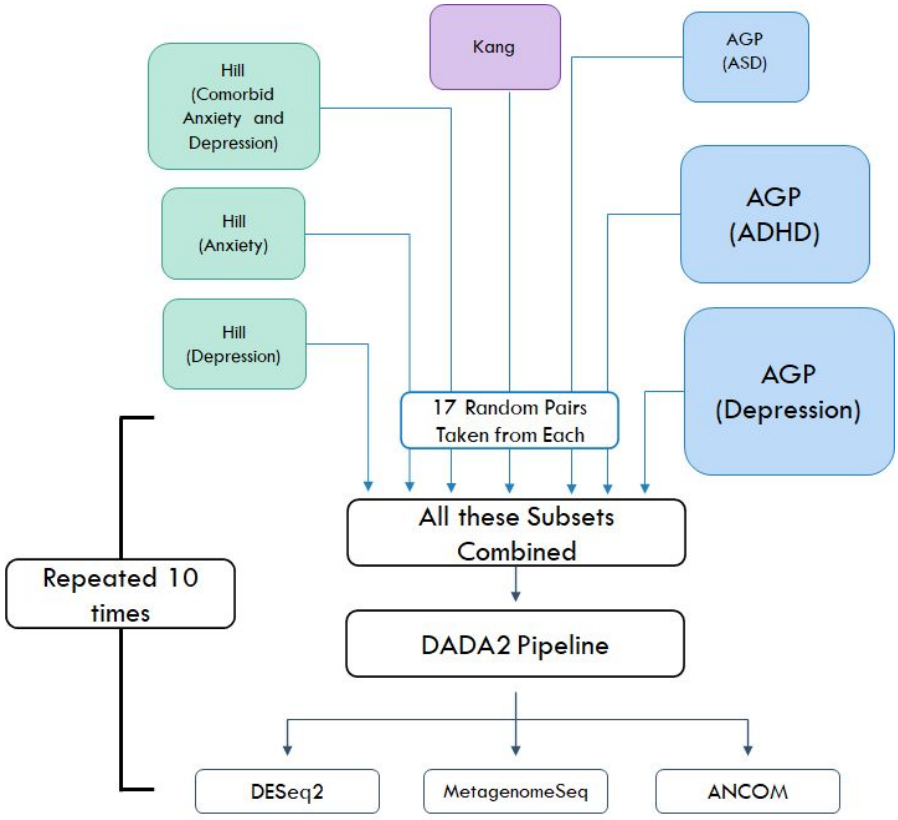
The adonis and betadisper functions were used for all metadata variables in the Permanova analysis. 9999 permutations were used in the adonis function for the dataset. Variables identified as potentially confounding were investigated further using wilcoxon-ranked-sum tests to assess their impact on the results. Constricted PCOAs were constructed using Bray-Curtis distances and DESeq2 normalization since this normalization mitigated the differences between studies better than the CSS normalization as seen by the Permanova results. The R package, ggplot2, was used for PCOA plotting (Wickham 2016).

4. Random Subsetting Analysis

To account for the higher number of samples present in the American Gut Project (AGP) study and to allow for each different type of diagnosis (anxiety, depression, ASD, and ADHD) to match sample size within the dataset, 17 pairs of samples were randomly taken from each study for each different diagnosis present. This number was chosen because the Kang study only consisted of a total of 38 individuals. Hence, three sets of 38 samples were taken from the AGP study: one set containing ASD samples, one set containing ADHD samples, and another set containing samples with depression. Three sets were also taken from the Hill study: one set of samples with anxiety, one set of samples with depression, and another set exhibiting both conditions comorbidly. The Kang study only contained ASD as a diagnosis; thus, one set of 38 samples was taken and used in the analysis. Each pair of samples consisted of one exhibiting the anxious phenotype and one being neurotypical, and these pairs were matched to have similar ages. A total of 238 samples were randomly selected using this process, and the same analytical analysis pipeline was used for the whole

dataset (DESeq2, ANCOM, and Metagenomeseq) were used to find ASVs that were significantly different in abundance between the anxious phenotype and the neurotypical.

In addition, in order to get better representation of the studies that were subsetted, the randomization process of selecting samples described above was completed randomly 10 times. In each instance, we performed the same analytical pipeline (DESeq2, ANCOM, and Metagenomeseq), and the ASVs that were significantly different between the anxiety-related conditions group and the neurotypical were identified. The amount of times the same ASV was detected was also tabulated. The following figure, *Flow Chart 2*, shows the overall sampling method for this meta-analysis.



Flow Chart 2 Random Subset Analysis Workflow

The Kang study only had ASD as a category of diagnosis. The different colors represent different studies

5. 10-cross Validation with Random Forest Models

The ASV sequences found from analyzing the entire dataset and found throughout the 10 different random sample sets were used to train random forest models for classifying the samples as anxious or neurotypical. A 10-cross validation was done on the entire dataset and also with each study individually using the ASV sequences as predictors in the model. The 10-cross validation process consists of using 90% of the data to train the random forest model and using the remaining 10% as testing samples for classification. Using the caret package, the process of taking out 10% of the samples was done 10 times and in a way that allowed all samples to be in the testing subset at least once in order to ensure proper representation of all samples. The resulting correct and incorrect classifications were plotted on a Receiver Operating Characteristic (ROC) curve and the resulting Area Under Curve (AUC) values were calculated.

ROC curves were generated by using multiple different groupings of the ASVs found in the study (see Section 4 in the results). The first set of predictors consisted of the sequences found by analyzing across all dataset. The three subsequent sets of predictors were created based on the number of times they were detected in the 10 different random subsets: one set of predictors of all of the sequences detected, another with sequences detected twice or more, and another with sequences detected three times or more. In addition, a null predictor set was created by generating a mock table of fake ASV counts using a random uniform distribution. Lastly, the metadata of age, sex, and antibiotics use were used as

predictors, as well as the metadata combined with some of the ASV predictors in order to compare how the model behaves without microbial sequence biomarkers.

6. Phylogenetic tree

In order to analyze potential phylogenetic associations between the ASVs that were significantly different between the anxious and nonanxious, a phylogenetic tree was created with the tips as individual ASVs present in all samples. The Phyloseq tree was generated using a outlined workflow from the DADA2 authors at the following link: http://web.stanford.edu/class/bios221/MicrobiomeWorkflowII.html#construct_phylogenetic_tree. (Ben J. Callahan et al. 2016). Packages used for this process included Phangorn, APE, and DECIPHER (Schliep 2012; Paradis, Claude, and Strimmer 2004; Wright, Erik, and Wright 2016). ASVs were colored according to their respective groupings within the predictor sets used in the random forest models. The tree was made from a subset of the taxa within the phyloseq object by removing any bacterial families that were not represented among the significant ASVs.The tree was rooted using the following archaea sequence from *Halorhabdus rudnickae* :

"GATCGATTAGCATGCTAGTCGCACGGGTTTAGGCCCGTGGCGGAAGCTCAGTAACACGTGGCCAAACTACCCTGTGGACGA
GAATACCCTCGGGAAACTGAGGTCAATTCTCGATACGGCTCTCATGCTGGAGTGCAGCGAGCCGGAAATGTTCTGGCGCCAC
AGGATGTGGCTGCGGCCGATTAGGTAGACGGTGAGGTAACGGCTCACCGTGCCAATAATCGGTACGGGTCATGAGAG"

RESULTS

1. Identification of ASVs Significantly enriched in each cohort

After using DESeq2, Metagenomeseq, and ANCOM on the 1226 samples within the phyloseq object, a total of eight ASVs were significantly different in abundance between the anxiety-related conditions and the neurotypical. *Table 2* shows the taxonomic information for

each ASV and their associated information. Full ASV sequences can be found in the supplementary information in *Table 8*.

Table 2 Significant Taxa between the Anxious and Neurotypical Phenotypes

*Taxa highlighted in red were significantly enriched in the gut-microbiome of individuals with anxiety-related conditions. Green were enriched in the neurotypical. *ANCOM does not estimate log-fold change, but it's greater prevalence in the neurotypical phenotype was determined using abundance counts. Method refers to which R package determined the ASV to be significant.*

Family	Genus	Log-2-Fold	q-values	Method
Akkermansiaceae	*Akkermansia*	*0.605896224*	*0.012013994*	*MetagenomeSeq*
Lachnospiraceae	*Roseburia*	*0.31406044*	*0.024243432*	*MetagenomeSeq*
Ruminococcaceae	*Butyricicoccus*	*0.28536855*	*0.026396466*	*MetagenomeSeq*
Ruminococcaceae	*Faecalibacterium*	*-0.26531737*	*0.032092058*	*MetagenomeSeq*
Ruminococcaceae	*NA*	*-0.28370918*	*0.009146485*	*MetagenomeSeq*
Bacteroidaceae	*Bacteroides*	*-0.49029464*	*0.000130443*	*MetagenomeSeq*
Bacteroidaceae	*Bacteroides*	*-0.5370154*	*0.00414062*	*MetagenomeSeq*
Marinifilaceae	*Odoribacter*	*<0.00**	*<0.05*	*ANCOM*

2. Impact of covariates and study origin on microbial community structure

In order to analyze for the effects of metadata variables on the gut-microbial communities, PCoA plots were generated and Permanova was run on the dataset. *Table 3* and *4* show the resulting p-values when the adonis and betadisper functions were performed on each normalization.

Table 3 Resulting p-values from Permanova on DESeq-normalized phyloseq

Adonis is a function that determines if a group centroid is significantly different from the rest of the data. Betadisper is a function that determines if a group has a heterogeneous dispersion or not.Red rows highlight variables that have a homogenous group dispersion and significant different centroid as seen by significance in adonis, but not betadisper.

Variable	Adonis p-value	Betadisper p-value
Sex	0.00015	0.21
Age	0.00015	0.002
Depth	0.00015	0.937
Antibiotics within 6 months	0.00024	0.002
Anxiety-Related Condition	0.055	0.937
Study	0.00015	0.002

Table 4 Resulting p-values from Permanova on CSS-normalized phyloseq

Variable	Adonis p-value	Betadisper p-value
Sex	0.00012	0.0984
Age	0.00012	0.002
Depth	0.00012	0.909
Antibiotics within 6 months	0.00012	0.002
Anxiety-Related Condition	0.0726	0.084
Study	0.00012	0.002

In the Permanova analysis, the adonis function found sex, age, antibiotics usage, study, and sequencing depth to be signficant using both CSS and DESeq2 normalizations.

The presence of an anxiety-related condition variable had a p-value of ~0.07 and ~0.06 for each of these nomalizations. However, only sex, depth, and the anxious phenotype had homogeneous dispersions after using the betadisper function on both normalizations. *Figure 1* displays two unconstrained PCoA plots; one with samples colored by study and another with samples colored by phenotype (anxious or neurotypical). Both of these were normalized using DESeq2. These PCoAs were also constructed using CSS normalizations. These plots can be found in the supplementary information in *Figure 8* and *Figure 9*.

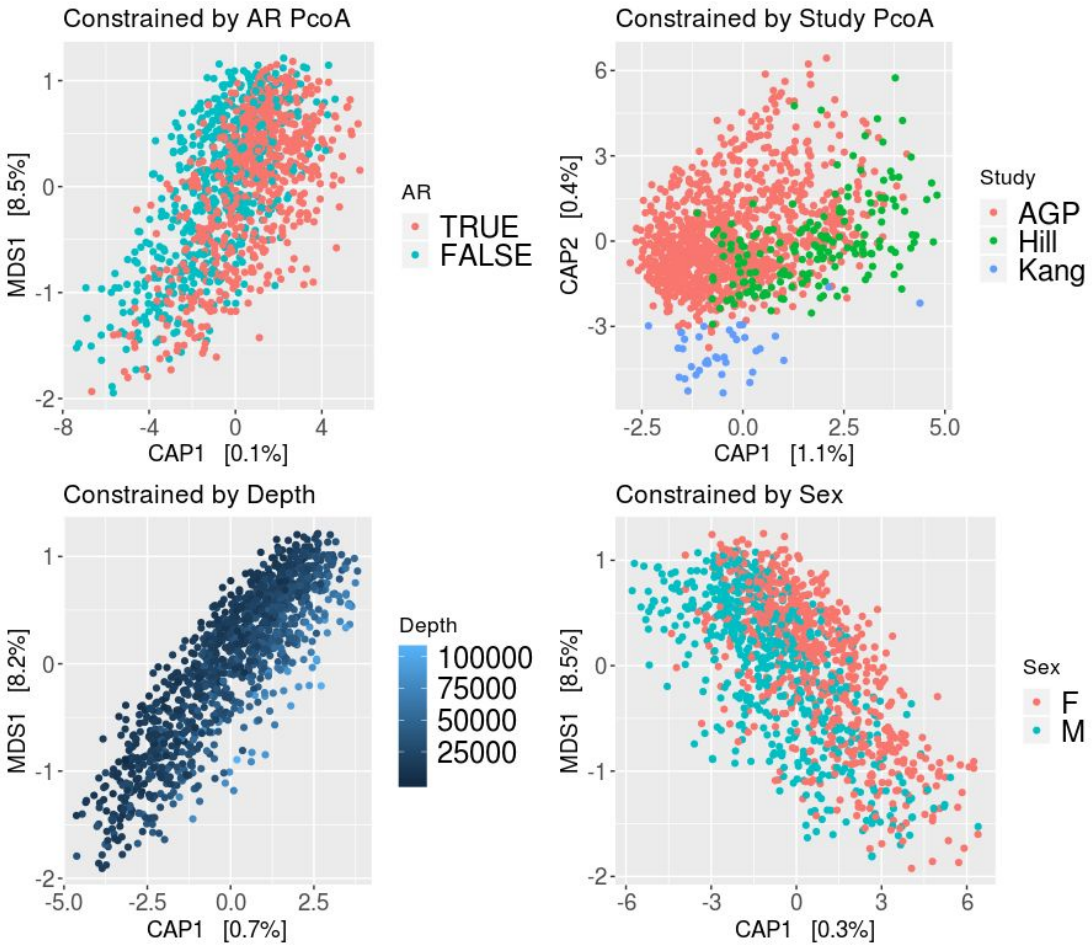


Figure 1 Constrained PCoAs of all 1226 Samples using DESeq2 Normalization

Samples were colored according to the study they came from. AGP stands for "American Gut Project"The label, "AR", found in the legend stands for "Anxiety-Related" and refers to individuals with anxiety-related conditions.

3. Analysis of Random Subsets

As mentioned in the Methods, an analysis for differentially abundant taxa was also performed on 10 random subsets containing 238 samples per subset. Each subset consisted of 17 pairs for each different kind of anxiety-related condition present in each study in order to avoid the study with the highest number of samples from being over-represented, and each pair consisted of one with the anxiety-related condition and an age-matched control. *Table 5* and *6* on the next pages display the taxonomic classification of significant ASVs among the 10 subsets and the amount of subsets in which were significant. A total of 33 different ASVs were significant in at least one of the subsets.

Table 5 Significantly Enriched Taxa in Anxious Individuals in at least Two Random Subsets

Some ASVs were unable to be classified at the genus level, and thus were labeled as NA for the genus.Full sequences of ASVs can be found in the supplementary information.

Family	Genus	Enrichment	Times Detected in 10-fold Sampling
Tannerellaceae	*Parabacteroides*	A	6
Rikenellaceae	*Alistipes*	A	6
Veillonellaceae	*Dialister*	A	6
Marinifilaceae	*Odoribacter*	A	6
Lachnospiraceae	*NA*	A	4
Lachnospiraceae	*Coprococcus_3*	A	4
Ruminococcaceae	*Ruminococcaceae_UCG-003*	A	4
Rikenellaceae	*Alistipes*	A	3
Erysipelotrichaceae	*Turicibacter*	A	3
Akkermansiaceae	*Akkermansia*	A	3
Ruminococcaceae	*Faecalibacterium*	A	2
Ruminococcaceae	*Faecalibacterium*	A	2
Desulfovibrionaceae	*Bilophila*	A	2
Lachnospiraceae	*Tyzzerella*	A	2
Lachnospiraceae	*Lachnospiraceae_UCG-010*	A	2
Lachnospiraceae	*Lachnoclostridium*	A	2
Erysipelotrichaceae	*Erysipelatoclostridium*	A	2

Table 6 Significantly Enriched Taxa in the Neurotypical in at least Two Random Subsets

Some ASVs were unable to be classified at the genus level, and thus were labeled as NA for the genus. Each row represents a different ASV despite similar "NA" genus labels in the Lachnospiraceae family. Full sequences of ASVs can be found in the supplementary information.

Family	*Genus*	*Enrichment*	Times Detected in 10-fold Sampling
Enterobacteriaceae	*Escherichia/Shigella*	*N*	*6*
Lachnospiraceae	*NA*	*N*	*4*
Ruminococcaceae	*Ruminococcaceae_UCG-005*	*N*	*3*
Bacteroidaceae	*Bacteroides*	*N*	*3*
Barnesiellaceae	*Coprobacter*	*N*	*2*
Rikenellaceae	*Alistipes*	*N*	*2*
Bacteroidaceae	*Bacteroides*	*N*	*2*
Lachnospiraceae	*Lachnospiraceae_NK4A136_group*	*N*	*2*
Lachnospiraceae	*NA*	*N*	*2*
Family_XIII	*Family_XIII_UCG-001*	*N*	*2*
Christensenellaceae	*Christensenellaceae_R-7_group*	*N*	*2*
Ruminococcaceae	*Ruminococcaceae_UCG-002*	*N*	*2*
Ruminococcaceae	*NA*	*N*	*2*
Lachnospiraceae	*Agathobacter*	*N*	*2*

ASVs from the genera, *Escherichia/Shigella, Alistripes, Parabacteroides, Odoribacter,* and *Dialister* were significant in six out of the ten random subsets and were the most commonly significant.

4 Random Forest Model

Figure 2, *Figure 3*, *Figure 4*, and *Figure 5* show the resulting ROC (Receiver Operating Characteristic) curves from using the ASVs found above as predictors in a variety of combinations. ROC curves are created by plotting the true positive rate over the false positive rate. The greater the area under the curve (AUC) is, the more often the classifier correctly identifies the phenotype. Metadata was also used for comparison. *Figure 2* shows the random forest model on all 1266 samples while *Figure 3*, *Figure 4*, and *Figure 5* show random forest model performance using samples exclusively from one study (AGP, Hill, and Kang respectively), using taxa identified as significant in one of the cohorts.

As seen in *Figure 2*, all predictors performed above 0.60 AUC (Area Under Curve) except the 8 ASVs identified across all datasets and the null model. The random subset sequences seen two or more times (ASVs from *Table 5 and 6*) paired with metadata performed the highest out of them all with about 0.70 AUC. The 8 ASVs identified across all datasets (ASVs from *Table 2*) performed at ~0.56 AUC and the Null model performed at ~0.50 .

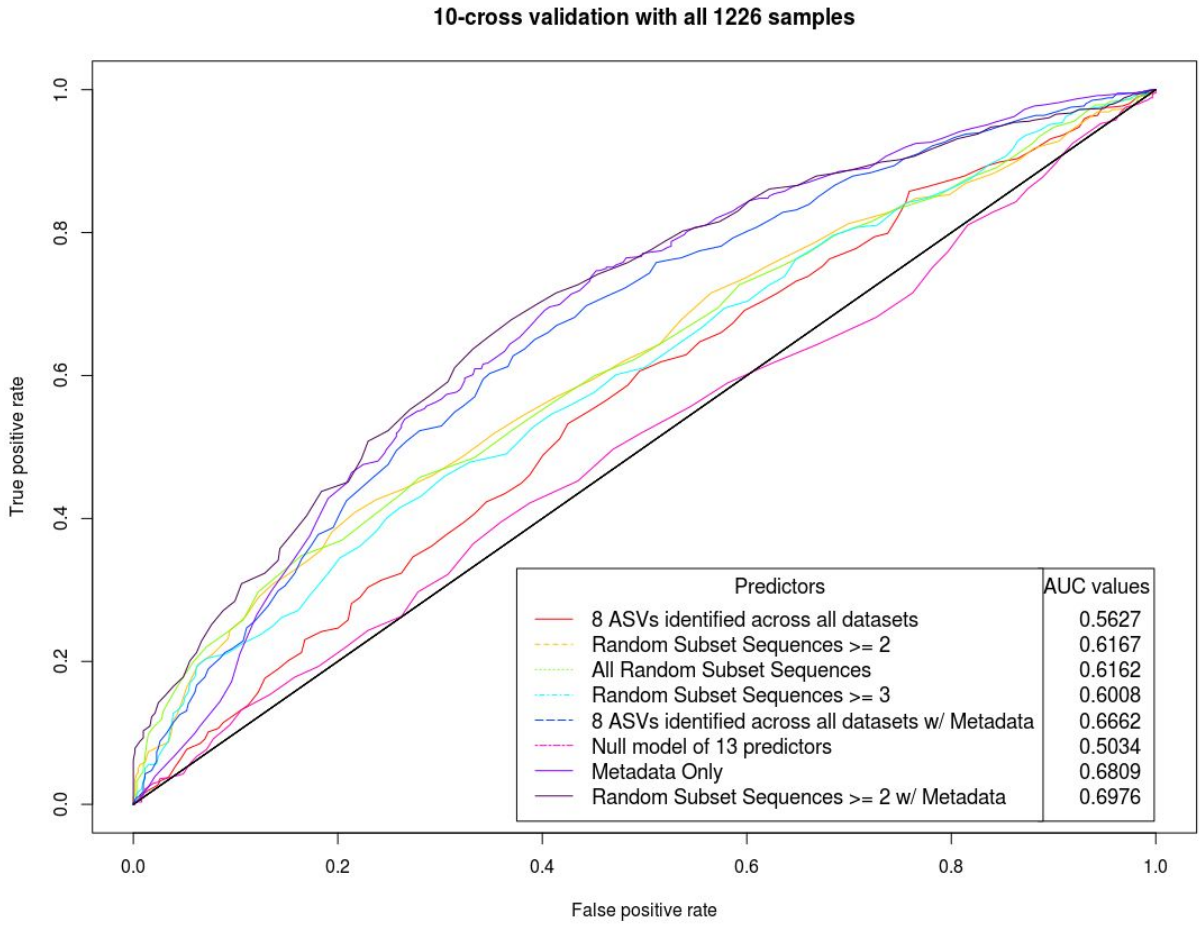


Figure 2 Random Forest Model Performance with all 1226 samples

AUC stands for "Area under the Curve". Predictors are used by the Random Forest Models as variables to train on for their machine learning algorithm. The True Positive Rate on the y-axis and plots the proportion of positives that were correctly identified as positive (e.g. the proportion of people with anxiety classified as anxious, also known as the sensitivity). The False Positive Rate on the x-axis is the opposite and represents how often the classifier labels a sample is positive when it is not(e.g labeling a sample as anxious when they are neurotypical).

The black line represents an AUC value of 0.50 and displays what a line would look like if the model was randomly classifying."8 ASVs identified across all datasets" refers to the 8 ASVs found when Metagenomeseq, ANCOM, DESeq2 was used on the entire data set as seen in Table 2. "Random Subset Sequences >= 2" and "Random Subset Sequences >= 3" refer to the ASVs found among the random subsets greater than two times and greater than 3 times respectively as seen in Table 5 and 6. The Null model refers to a mock dataset created by generating a random uniform distribution of counts for fake ASV sequences.

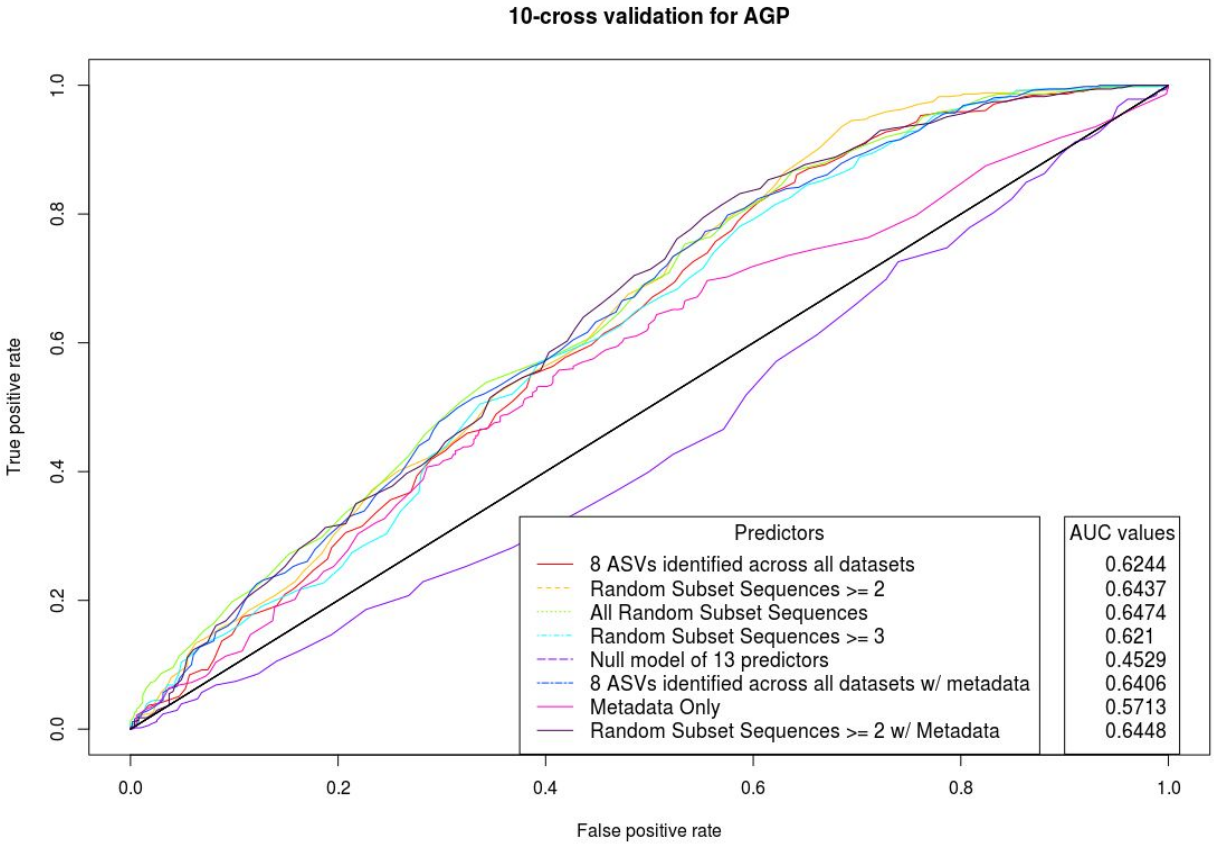


Figure 3 Random Forest Model Performance with American Gut Project samples (AGP)

Predictors are used by the Random Forest Models as variables to train on for their machine learning algorithm."8 ASVs identified across all datasets" refers to the 8 ASVs found when Metagenomeseq, ANCOM, DESeq2 was used on the entire data set as seen in Table 2. "Random Subset Sequences >= 2" and "Random Subset Sequences >= 3" refer to the ASVs found among the random subsets greater than two times and greater than 3 times respectively as seen in Table 5 and 6. The Null model refers to a mock dataset created by generating a random uniform distribution of counts for fake ASV sequences.

For the American Gut Project samples, all predictors except the metadata alone and the null model performed had greater than 0.60 AUC. The metadata alone had an AUC of ~0.57 and the null model had an AUC of ~0.45.

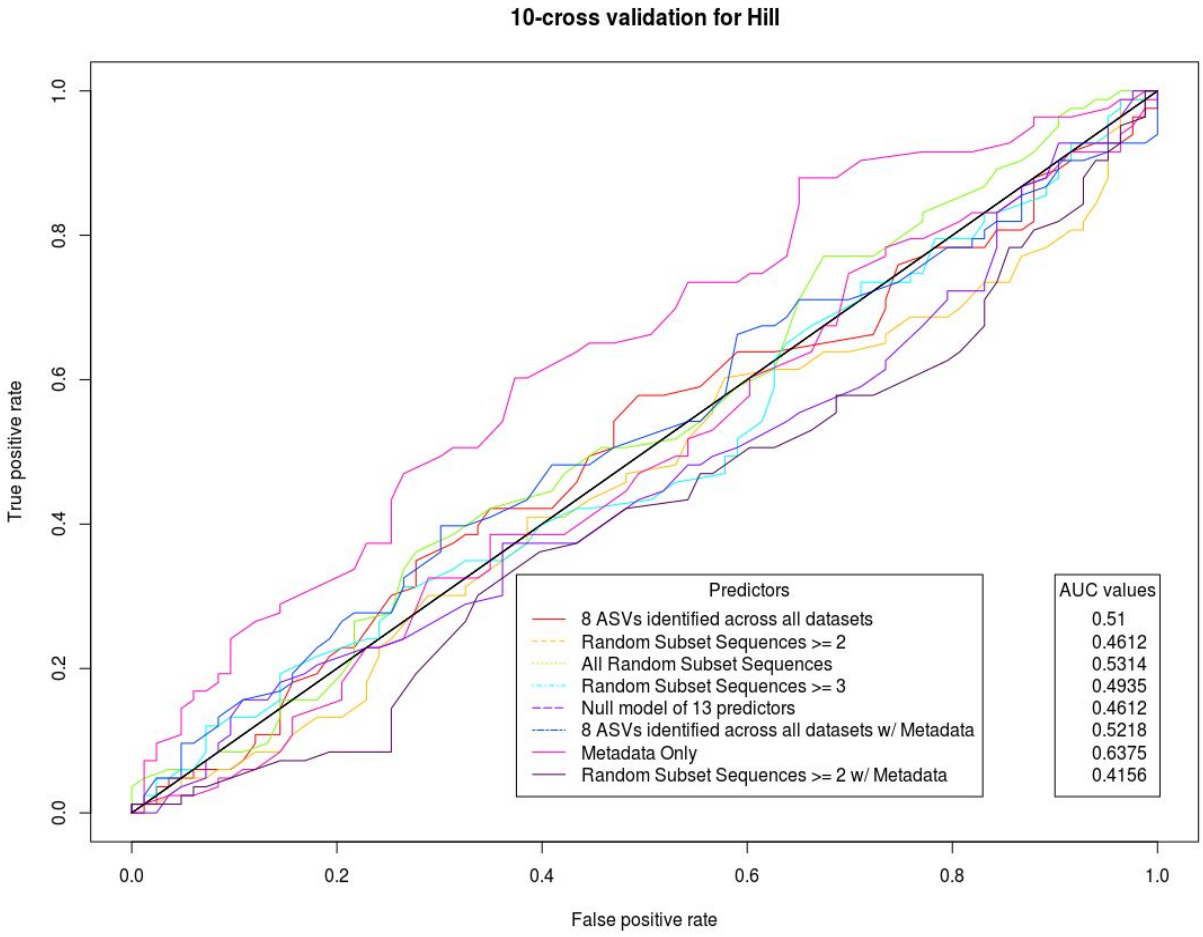


Figure 4 Random Forest Model Performance with Study by Hill et al.

Predictors are used by the Random Forest Models as variables to train on for their machine learning algorithm. "8 ASVs identified across all datasets" refers to the 8 ASVs found when Metagenomeseq, ANCOM, DESeq2 was used on the entire data set as seen in Table 2. "Random Subset Sequences >= 2" and "Random Subset Sequences >= 3" refer to the ASVs found among the random subsets greater than two times and greater than 3 times respectively as seen in Table 5 and 6. The Null model refers to a mock dataset created by generating a random uniform distribution of counts for fake ASV sequences.

As seen in *Figure 4* containing the Hill Study random forest ROC curves, all predictors except the Metadata alone had AUC values less than 0.60. Four of the predictors performed

with less than 0.50 AUC. Out of the sequence-based predictors, the "All Random Subset Sequences" performed the greatest with 0.53 AUC.

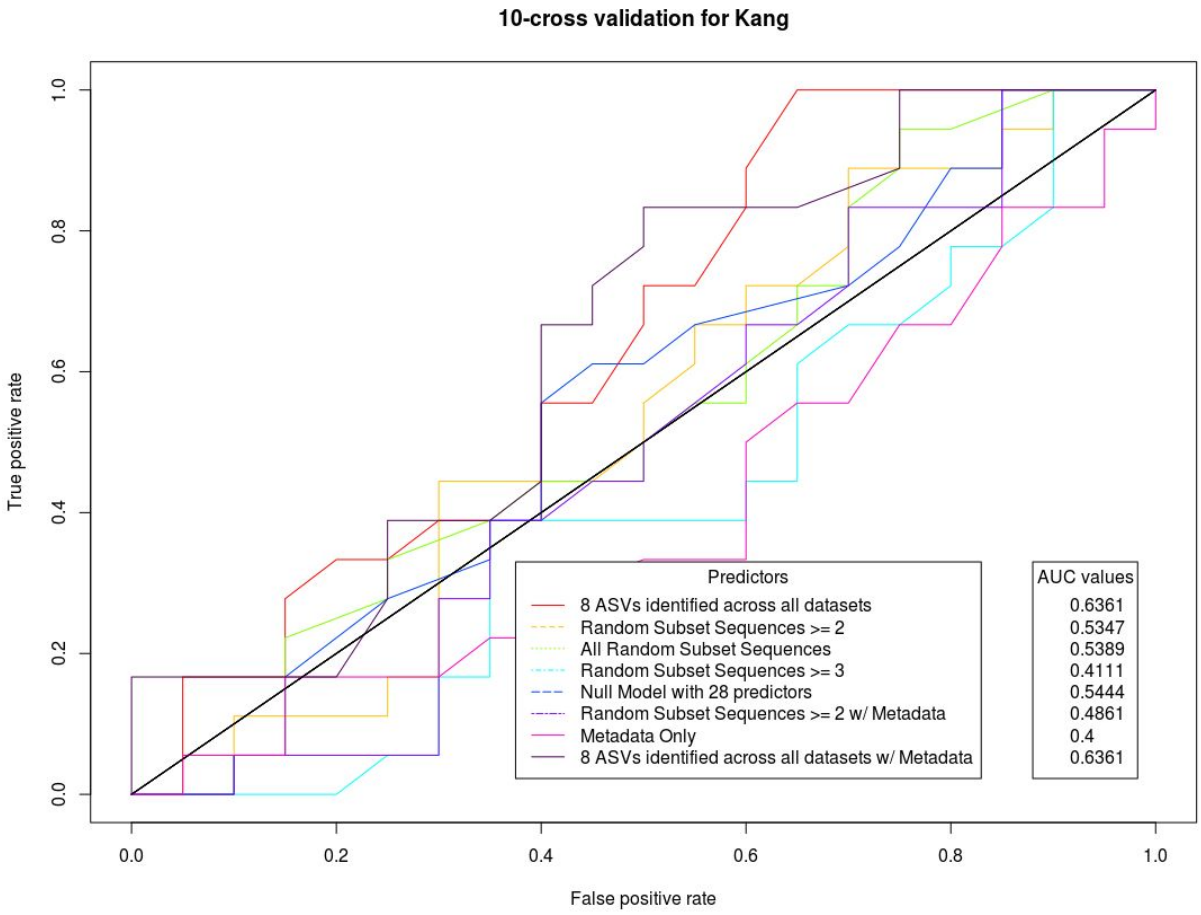


Figure 5 Random Forest Model Performance with Study by Kang et al.

Predictors are used by the Random Forest Models as variables to train on for their machine learning algorithm. "8 ASVs identified across all datasets" refers to the 8 ASVs found when Metagenomeseq, ANCOM, DESeq2 was used on the entire data set as seen in Table 2. "Random Subset Sequences >= 2" and "Random Subset Sequences >= 3" refer to the ASVs found among the random subsets greater than two times and greater than 3 times respectively as seen in Table 5 and 6. The Null model refers to a mock dataset created by generating a random uniform distribution of counts for fake ASV sequences.

In the Kang study, the 8 ASVs identified across all datasets had the highest AUC at ~0.64 and while the random subset sequences and metadata alone were all under 0.60 AUC. Excluding metadata predictors, the random subset sequence predictors had the highest AUC value within AGP, the Hill study, and when all 1226 samples were combined. However, random subset sequences predictors performed poorly for the Kang study with AUC values lower around 0.55 and lower. All ASV predictors in the AGP study had AUC values greater than 0.62 and were almost comparable to the metadata alone. *Table 7* summarizes all of the AUC values within the 10-cross validation process.

Table 7 Summary of AUC values of Random Forest 10-cross validation

"Green" indicates AUC values above 0.60. "Yellow indicates AUC values between 0.50 and 0.60. Red indicates AUC values below 0.50.

Predictor Category	All 1226 Samples	AGP	Hill	Kang
Null Model	0.5034	0.4529	0.4612	0.5444
8 ASVs Identified across all datasets	0.5627	0.6244	0.51	0.6361
All Random Subset Sequences	0.6162	0.6474	0.5314	0.5389
Random Subset Sequences >= 2	0.6167	0.6437	0.4612	0.5347
Random Subset Sequences >= 3	0.6008	0.621	0.4935	0.4111
Metadata Alone	0.6809	0.5713	0.6375	0.4
8 ASVs Identified across all datasets w/ Metadata	0.6662	0.6406	0.5218	0.6361
Random Subset Sequences >= 2 w/ Metadata	0.6976	0.6448	0.4156	0.4861

Overall, the "All Random Subset Sequences" predictor performed the best out of the categories without metadata. When the 8 ASVs identified across all datasets were combined with the metadata, it yielded more consistent results and higher AUCs than the other

categories. Among the different sets of random subset sequences, the higher thresholds of amount of times seen significant (greater than two and three) cause equal or even poorer classification performance.

5. Phylogenetic analysis of biomarkers of interest

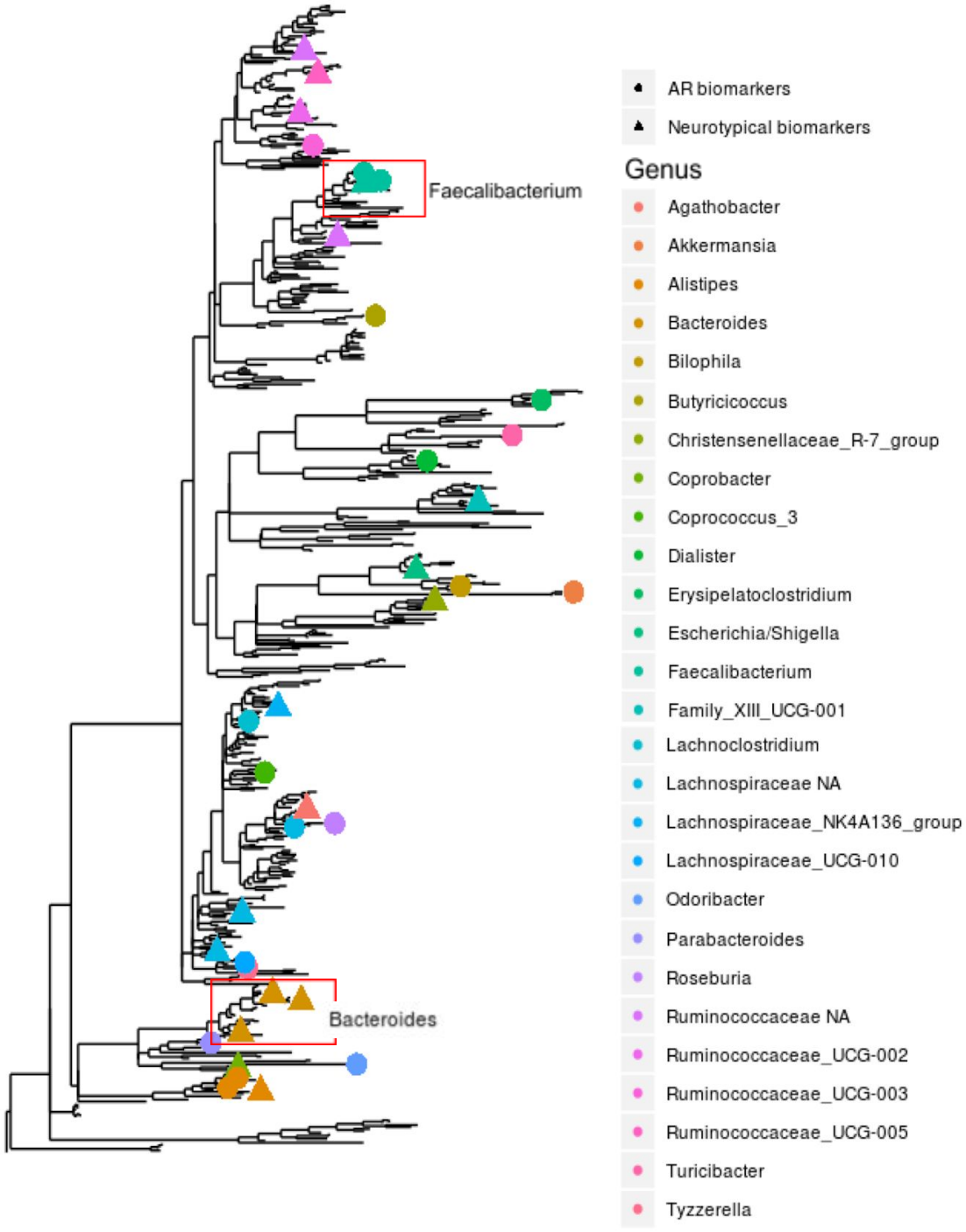
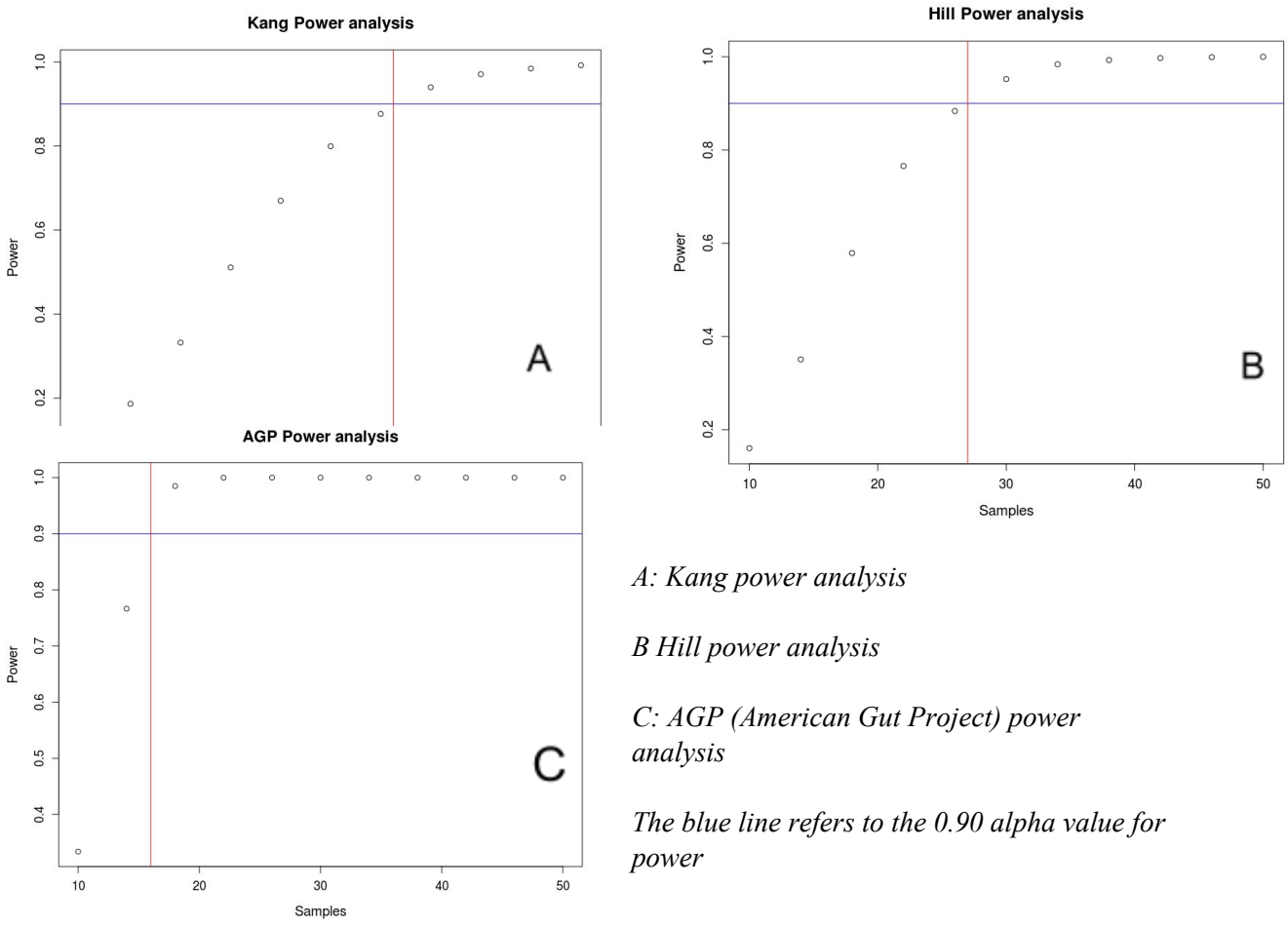


Figure 6 Phylogenetic Tree Labeled by Significant ASVs and their Enrichment

Figure 6 displays the phylogenetic tree generated from a subset of taxa families that were present among any of the ASVs that were significant and were used as a predictor. AR biomarkers refer to the anxious phenotype (Anxiety-related conditions). On the tree, there appears to be a small cluster of ASVs from the *Faecalibacterium* genus and the *Bacteroides* genus. The *Bacteroides* group boxed in red is also all enriched in the neurotypical and is on a branched section of its own. The enriched taxa in both the neurotypical and the anxious are relatively dispersed distributed throughout the tree.The following figure, *Figure 7*, shows the power analysis performed on the samples from each study.

Figure 7 Power Analysis



A: Kang power analysis

B Hill power analysis

C: AGP (American Gut Project) power analysis

The blue line refers to the 0.90 alpha value for power

The power for the Kang study was above 0.90 at around 35 samples while the Hill crossed the 0.90 threshold at about 27 samples. AGP was above 0.90 at about 17 samples. All of these studies had more samples than these amounts and thus had greater power than 0.90

DISCUSSION

1. ASVs taxonomy comparison

Several of the ASVs found relevant in this study have already been associated with neurological conditions in other studies. For instance, altered levels of *Akkersmania*, *Bacteroides, Parabacteroides, Coprococcus, Odoribacter,* and *Faecalibacterium* have been seen within the gut-microbiome of individuals with ASD (Xu et al. 2019; Kang et al. 2013; Zhang et al. 2018). *Faecalibacterium* and *Odoribacter* have been associated with mood disorders like bipolar disorder and major depressive disorder as well as other genera such as *Alistripes* and *Roseburia*. (Huang et al. 2019; Winter et al. 2018; Cheung et al. 2019). As seen in *Table 2, 5, and 6,* ASVs from all these genera mentioned were significantly different between individuals with anxiety-related conditions and the neurotypical within this meta-analysis which indicate our findings are generally consistent with the current literature. Our meta-analysis also found multiple members of the *Ruminococcaceae* family which have been associated with social avoidance behaviors in mice and major depressive disorders in humans (Schnorr and Bachner 2016; Cheung et al. 2019).

This analysis of differentially abundant taxa within anxiety-related conditions however also adds specific ASV sequences of these associated genera to the scientific literature. Having this sequence information can help future researchers identify more specific bacteria and is another step closer to identifying specific species of gut-microbiota

that are associated with certain neurological conditions. However, whether or not these associations are causing symptoms of neurological conditions or if they are a product of the conditions themselves is still undetermined. Our increased sample size of 1266 that yielded eight specific ASVs also provide increased clarity and support regarding the significance of these taxa within our grouping of anxiety-related conditions. The ASVs found to be significant within the random subsets also each possess a sample size of 238 which is more than many studies within the current literature. While we hypothesized that we would find more novel bacteria significantly different between the two cohorts through this particular means of meta-analysis than what was seen, finding these taxa to be significant despite grouping across a variety of four different anxiety-related conditions is unique to this meta-analysis.

Mapping the relevant markers to a phylogenetic tree highlighted that while only a few families were relevant to this study, members from the same families can be enriched in one cohort or another. While this observation is difficult to interpret, it supports the relevance of using ASVs, rather than 97% clustered sequences in order to identify biomarkers of interest.

2.Covariates impacting the microbial structure

The only two variables that were significant in adonis and also had a homogenous distribution were sequencing **depth** and **sex** as seen in *Table 3*. This result was consistent across the two normalizations used. These two factors have already been reported in the literature as potentially affecting beta-diversity(Zaheer et al. 2018; Dominianni et al. 2015). However, these two variables are not significantly different between the two cohorts. This is demonstrated for depth by the boxplot on Supplemental *Figure 10* which did not show a

significant difference between the two cohorts when using a wilcoxon-ranked sum test , and therefore should not constitute as a confounding factor in our study. The same observation was made for the sex of the indivuals involved in the study: a chi-squared test between sex and phenotype revealed the two to have no association with one another which also invalidates it as a confounding variable despite the anxious cohort presenting 10 more women and lacking 12 males over a total of 1266 samples. The fact that the sex ratio was almost 1:1 demonstrates that our sampling was biased towards the male cohort since studies have shown that women experience anxiety and depression at twice the rate as men (Hankin 2009). This bias is seen in mental health counseling and research as well (Danzinger and Welfel 2000). In addition, the ASD cohort was predominantly male at approximately a 3:1 ratio. This may have been why we did not see sex bias in our anxious cohort. Taking these findings and factors into account, it is unlikely that sex or depth had a significant impact on our results.

Note that the adonis test performed when considering the phenotype (*i.e.* the presence of an anxiety-related condition) was showed close to having resulted in a significant p-value from the adonis function at ~0.06, and also had a homogenous group dispersion (as displayed in *Table 3)*. While this may not classify as significant for this study, it may imply that future meta-analyses could see significant changes within the gut-microbial communities between anxious and neurotypical phenotypes.

3. Comparing ASVs significantly enriched across all samples with ASVs enriched during sub-sampling

Out of all the ASVs identified as significantly different between the two cohorts, three ASVs were significant across the entire dataset and within the random subsets. These ASVs were from the genera *Odoribacter, Akkersmania,* and *Bateroides*. The *Odoribacter* ASV was significant in six of the ten random subsets. *Akkersmania* and *Bacteroides* were detected as significant in three of the 10 random subsets which was more than most of the random subset sequences. The fact that these three exact sequences were identified as significant in the entire subset and in many of the random subsets demonstrates that these particular ASVs may be especially consistent across datasets. Taxa at the order levels represented by these ASVs have already been characterized in studies mentioned previously in Discussion Section 1 .

4. Random Forest Model Performances

The second part of our hypothesis was that these ASVs could be used as predictors to increase the performance of a random forest model in classifying the two groupings in a 10-cross validation. As shown in *Figures 2, 3, 4, and 5*, using the ASV sequences without metadata increased the AUC to above 0.60 in many cases. While AUC values of this magnitude are not sufficient for a proper classification model, it is much greater than random as well as the AUC values from the null predictors, which supports the notion that these taxa are associated with anxiety and comorbid disorders.

However, as seen in *Figure 2*, the random subset sequence predictors (ASVs from *Table 5* and *6* with an AUC of ~0.61) all outperformed the 8 sequences found across all

datasets (AUC of ~0.56) when the 10-cross validation was performed on all 1226 samples. We hypothesized that the 8 significant sequences found across all datasets would outperform ones found in the 10 random subsets when building a model with all 1226 samples, since the random subsets were more representative of the smaller studies and not the dataset as a whole. This may be due to the fact that the random subset predictors had more sequences to use as predictors (33 ASVs instead of 8 ASVs found across all datasets). Having more sequences as predictors could have helped train the model more proficiently.

In *Figure 2*, the metadata alone (consisting of age, sex, and antibiotics use within the past six months) had an AUC value of ~0.68. The fact that the best of the random subset sequence predictors had an AUC value of ~0.62, only 0.06 lower than the metadata, shows that these taxa are almost as meaningful to the classifier as variables such as age and sex which play a signficant role in gut-microbial communities (Dominianni et al. 2015; Jašarević, Morrison, and Bale 2016). Additional exploration into random forest models with both metadata and the ASV predictors for this study in particular could be done to better determine if the addition of 16S amplicon improves the classification, or, rather, adds more noise.

Figures 3, *4* , and *5* show the performance of the random forest models on samples from individual studies. *Figure 3* displaying AGP's classifier, shows that the classifier performed best on this dataset, with AUC values of 0.62 to 0.64. These values can be compared to the AUC value of the 8 sequences found across all datasets (AUC = ~0.57 as seen in *Figure 3*). Such a result was expected given that AGP has the most samples out of the three studies. This may be due to the greater number of ASV sequences present within

the random subset sequence predictors available for the model to train with. This explanation is also consistent with the fact that the random subset sequences detected three or more times had less sequences and also performed slightly worse.

All classifiers performed very poorly with the Hill study (*Figure 4*), and the best being the random subset sequences at ~0.53. While it was expected that the random subset sequences would perform the best on this study due to how the proportion of Hill samples were greater in the random subsets than in the entire dataset, a value of 0.53 is minimally above random and much less than the 0.63 AUC value of the metadata alone. One potential reasoning as to why performance of ASV predictors were so poor could be due to the prevalence of older individuals within this study. As mentioned in the Methods, the average age of individuals in this study was ~68 years old. While across all studies no significance in age was detected between the two cohorts, age affects the gut-microbiome (as discussed earlier in this section), and could render identified predictors for anxiety-related conditions irrelevant for this specific dataset. Given the small sample size of this dataset and the number of samples required (around 27 samples as seen Figure 7 in Results Section 5), we did not aim to identify markers specific to this study. In addition, while individuals with Parkinson's diseases included in this study were always matched in the control cohort with individuals possessing the opposite phenotype (i.e. not without anxiety) and Parkinson's disease as well, this may have played a role in affecting model performance since Parkinson's disease is more comorbid with anxiety than the general population and may have a unique effect on microbial structure on it's own (Chen and Marsh 2014; Hill-Burns et al. 2017).

Finally, the 8 predictors significant across all studies performed better in the Kang datasets (*Figure 5*, AUC of 0.63*)* while all the random sequence predictor categories had AUC values below ~0.55. This was unexpected considering that the Kang samples were only a small portion of the total 1226 samples across all datasets, One reason for this may be due to the fact that the Kang study only consisted of individuals with ASD and their respective controls, which also constituted samples from the largest study here, AGP (the American Gut Project). Note that such an explanation would imply that these ASVs could be of relevance for relevance to ASD specifically. Furthermore, within the Kang study 10-cross validation, the metadata performed the lowest even when compared with the other studies. This is likely due to the fact that all the individuals in this study were below the age of 17 and were mostly male as seen in *Table 1*, and hence training a machine learning classifier using these variables would be ineffective since differences between the phenotypes would be minimal.

Figure 6 shows the degree to which the positively and negatively enriched ASVs are grouped phylogenetically. On this phylogenetic tree, there appears to be a small cluster of ASVs from the *Faecalibacterium* genus and the *Bacteroides* genus; however, there are many sequences that are dispersed across the entire tree. While these individual clusters may be worthy of analysis and further identification (especially since the *Bacteroides* ASVs were all enriched in the neurotypical), overall positioning of positively or negatively enriched ASVs in the anxious phenotype appear to be dispersed fairly evenly across the tree.

5. Limitations and Future Research

Within this meta-analysis, there are multiple issues within the dataset that must be addressed. One of which is the fact that most of the different diagnoses (ADHD, Depression,

Anxiety, etc.) were self-reported. This could cause false positives to be present in the dataset. In addition, by the nature of reporting true or false to these queries of neurological conditions, we are unable to assess the degree or severity of a given condition. If levels of potentially associated taxa are affected by condition severity, then certain taxa may not be found to be significant if a high number of mild or light cases were present. Another factor in this study lies in the differences between the different anxiety-related conditions. Different conditions may result in individual differences in the gut-microbiome despite their links to anxiety. The fact that significant taxa were found among these conditions when grouped despite their inherent differences highlight their potential underlying commonalities and the power of larger sample sizes brought about by this meta-analysis.

Difference in samples between studies also may have been an obstacle to our analysis. The difference in average age of samples between these studies were stark, and inherent differences brought about by sequencing done by different labs may have played a role in the sequence data. While the Permanova analysis did not reveal significant differences in gut-microbial communities due to the study by which they came from, the PCoAs did a degree of separation between studies. Even if the magnitude of these differences were minimal as seen in the small percentages on the axis for the PCoA for study in Figure 1, it should be acknowledged that the three studies in this meta-analysis used different extraction methods that could have introduced unwanted variation in the sequence data. While study was not significant in Permanova within this meta-analysis, the methods of extraction and preparation should always be examined in any meta-analysis of this nature.

Data acquisition of sequence data and metadata for this study was also a difficulty in this study. Originally more studies were to be included in this meta-analysis; however, many studies had their sequence data uploaded, but not their meta-data. Emails to those responsible for these study data resulted in no response, or in one case, an admission that there were issues with the metadata that they would need to fix and re-release to the public. Hence, the process of this meta-analysis also reveals the need for greater clarity in uploaded sequence data so that meta-analyses can be performed properly and with greater ease in order to provide more contributions to the scientific community at large.

One important contributing factor that was lacking in this meta-analysis is diet. Diet plays a key role in the gut microbial community, and future meta-analyses or studies should include this information if possible while still maintaining high sample sizes. For meta-analyses of neurological conditions in particular, using standardized scales of anxiety, stress, or depression would be important to include in order to determine if differences in severity affect the magnitude of gut-microbial community changes. In addition, multiple timepoints for each sample would also provide valuable information on the association of anxiety or anxiety-related conditions and the gut-microbiome. In conclusion, if a study could incorporate these changes while still maintaining a high sample size, it could provide a clearer picture on the relationship between anxious conditions, the gut-microbiome, and the gut-brain axis.

Supplementary Information:

Table 8 Significant ASVs within the 1266-Sample Analysis

Some ASVs were unable to be classified at the genus level, and thus were labeled as NA for the genus.

Family	Genus	ASV Sequence
Akkermansiaceae	*Akkermansia*	*TACAGAGGTCTCAAGCGTTGTTCGGAATCACTGGGCGTAAAGCGTGCGTAGGCTGTTTCGTAAGTCGTGTGTGAAAGGCGCGGGCTCAACCCGCGGACGGCACATGATACTGCGAGACTAGAGTAATGGAGGGGGAACCGGAATTCTCGG*
Lachnospiraceae	*Roseburia*	*TACGTATGGTGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGCAGGCGGAAGGCTAAGTCTGATGTGAAAGCCCGGGGCTCAACCCCGGTACTGCATTGGAAACTGGTCATCTAGAGTGTCGGAGGGGTAAGTGGAATTCCTAG*
Ruminococcaceae	*Butyricicoccus*	*TACGTAGGGAGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGCGCGCAGGCGGGCCGGTAAGTTGGAAGTGAAATCTATGGGCTTAACCCATAAACTGCTTTCAAAACTGCTGGTCTTGAGTGATGGAGAGGCAGGCGGAATTCCGTG*
Ruminococcaceae	*Faecalibacterium*	*AACGTAGGTCACAAGCGTTGTCCGGAATTACTGGGTGTAAAGGGAGCGCAGGCGGGAGAACAAGTTGGAAGTGAAATCCATGGGCTCAACCCATGAACTGCTTTCAAAACTGTTTTTCTTGAGTAGTGCAGAGGTAGGCGGAATTCCCGG*
Ruminococcaceae	*NA*	*TACGTAGGGAGCGAGCGTTGTCCGGAATTACTGGGTGTAAAGGGAGCGTAGGCGGGAAAGCAAGTTGGAAGTGAAATGCATGGGCTTAACCCATGAGCTGCTTTCAAAACTGTTTTTCTTGAGTGAAGTAGAGGCAGGCGGAATTCCTAG*
Bacteroidaceae	*Bacteroides*	*TACGGAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGATGGATGTTTAAGTCAGTTGTGAAAGTTTGCGGCTCAACCGTAAAATTGCAGTTGATACTGGATGTCTTGAGTGCAGTTGAGGCAGGCGGAATTCGTGG*
Bacteroidaceae	*Bacteroides*	*TACGGAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGGCGGACTATTAAGTCAGCTGTGAAAGTTTGCGGCTCAACCGTAAAATTGCAGTTGATACTGGTCGTCTTGAGTGCAGTAGAGGTAGGCGGAATTCGTGG*
Marinifilaceae	*Odoribacter*	*TACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGTGCGTAGGCGGTTTATTAAGTTAGTGGTTAAATATTTGAGCTAAACTCAATTGTGCCATTAATACTGGTAAACTGGAGTACAGACGAGGTAGGCGGAATAAGTTAA*

Table 9 Significantly Enriched ASV in Anxious Individuals in at least Two Random Subsets

Some ASVs were unable to be classified at the genus level, and were labeled as NA for the genus.

ASV Sequence	Family	Genus
TACGGAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGTGCGTAGGCGGCCTTTTAAGTCAGCGGTGAAAGTCTG TGGCTCAACCATAGAATTGCCGTTGAAACTGGGGGGCTTGAGTATGTTTGAGGCAGGCGGAATGCGTGG	Tannerellaceae	Parabacteroides
TACGGAGGATTCAAGCGTTATCCGGATTTATTGGGTTTAAAGGGTGCGTAGGCGGTTTGATAAGTTAGAGGTGAAATTTCG GGGCTCAACCCTGAACGTGCCTCTAATACTGTTGAGCTAGAGAGTAGTTGCGGTAGGCGGAATGTATGG	Rikenellaceae	Alistipes
TACGTAGGTGGCAAGCGTTGTCCGGAATTATTGGGCGTAAAGCGCGCGCAGGCGGCTTCCCAAGTCCCTCTTAAAAGTGCG GGGCTTAACCCCGTGATGGGAAGGAAACTGGGAAGCTGGAGTATCGGAGAGGAAAGTGGAATTCCTAGT	Veillonellaceae	Dialister
TACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGTGCGTAGGCGGTTTATTAAGTTAGTGGTTAAATATTT GAGCTAAACTCAATTGTGCCATTAATACTGGTAAACTGGAGTACAGACGAGGTAGGCGGAATAAGTTAA	Marinifilaceae	Odoribacter
TACGTAGGGGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGAATGGCAAGTCTGATGTGAAAGGC CGGGGCTCAACCCCGGGACTGCATTGGAAACTGTCAATCTAGAGTACCGGAGGGGTAAGTGGAATTCCTAG	Lachnospiraceae	NA
TACGTATGGTGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGCTGTGTAAGTCTGAAGTGAAAGCCC GGGGCTCAACCCCGGGACTGCTTTGGAAACTATGCAGCTAGAGTGTCGGAGAGGTAAGTGGAATTCCCAG	Lachnospiraceae	Coprococcus_3
TACGGAGGATCCAAGCGTTATCCGGATTTATTGGGTTTAAAGGGTGCGTAGGCGGTTTGATAAGTTAGAGGTGAAATACC GGGGCTCAACTCCGGAACTGCCTCTAATACTGTTGAACTAGAGAGTAGTTGCGGTAGGCGGAATGTATGG	Rikenellaceae	Alistipes
TACGTAGGTGGCAAGCGTTGTCCGGATTTACTGGGTGTAAAGGGCGTGCAGCCGGGCATGCAAGTCAGATGTGAAATCTC AGGGCTTAACCCTGAAACTGCATTTGAAACTGTATGTCTTGAGTGCCGGAGAGGTAATCGGAATTCCTTG	Ruminococcaceae	Ruminococcaceae (UCG-003)
TACGTAGGTGGCGAGCGTTATCCGGAATTATTGGGCGTAAAGAGCGCGCAGGTGGTTGATTAAGTCTGATGTGAAAGCCC ACGGCTTAACCGTGGAGGGTCATTGGAAACTGGTCGACTTGAGTGCAGAAGAGGGAAGTGGAATTCCATG	Erysipelotrichaceae	Turicibacter
AACGTAGGTCACAAGCGTTGTCCGGAATTACTGGGTGTAAAGGGAGCGCAGGCGGGAAGACAAGTTGGAAGTGAAATCT ATGGGCTCAACCCATAAACTGCTTTCAAAACTGTTTTTCTTGAGTAGTGCAGAGGTAGGCGGAATTCCCGG	Ruminococcaceae	Faecalibacterium
AACGTAGGTCACAAGCGTTGTCCGGAATTACTGGGTGTAAAGGGAGCGCAGGCGGGCGATCAAGTTGGAAGTGAAATCC ATGGGCTCAACCCATGAACTGCTTTCAAAACTGGTCGTCTTGAGTAGTGCAGAGGTAGGCGGAATTCCCGG	Ruminococcaceae	Faecalibacterium
TACAGAGGTCTCAAGCGTTGTTCGGAATCACTGGGCGTAAAGCGTGCGTAGGCTGTTTCGTAAGTCGTGTGTGAAAGGCG CGGGCTCAACCCGCGGACGGCACATGATACTGCGAGACTAGAGTAATGGAGGGGGAACCGGAATTCTCGG	Akkermansiaceae	Akkermansia
TACGGAGGGTGCAAGCGTTAATCGGAATCACTGGGCGTAAAGCGCACGTAGGCGGCTTGGTAAGTCAGGGGTGAAATCCC ACAGCCCAACTGTGGAACTGCCTTTGATACTGCCAGGCTTGAGTACCGGAGAGGGTGGCGGAATTCCAGG	Desulfovibrionaceae	Bilophila
TACGTAGGGGGCAAGCGTTATCCGGAATTACTGGGTGTAAAGGGAGCGTAGGCGGCATGGTAAGCCAGATGTGAAAGCC TTGGGCTTAACCCGAGGATTGCATTTGGAACTATCAAGCTAGAGTACAGGAGAGGAAAGCGGAATTCCTAG	Lachnospiraceae	Tyzzerella
TACGTAGGGGGCAAGCGTTATCCGGAATTACTGGGTGTAAAGGGTGAGTAGGCGGCATGGCAAGTAAGATGTGAAAGCC CGAGGCTTAACCTCGGGATTGCATTTTAAACTGCTAAGCTAGAGTACAGGAGAGGAAAGCGGAATTCCTAG	Lachnospiraceae	Lachnospiraceae (UCG-010)
TACGTAGGGGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGTAAAGCAAGTCTGAAGTGAAAGCCC GGGGCTCAACCCCGGGACTGCTTTGGAAACTGTTTAACTAGAGTGCTGGAGAGGTAAGCGGAATTCCTAG	Lachnospiraceae	Lachnoclostridium
TACGTAGGTGGCAAGCGTTATCCGGAATTATTGGGCGTAAAGAGGGAGCAGGCGGCAGCAAGGGTCTGTGGTGAAAGCC TGAAGCTTAACTTCAGTAAGCCATAGAAACCAGGCAGCTAGAGTGCAGGAGAGGATCGTGGAATTCCATGT	Erysipelotrichaceae	Erysipelatoclostridium

Table 10 Significantly Enriched Taxa in the Neurotypical in at least Two Random Subsets

Some ASVs were unable to be classified at the genus level, and thus were labeled as NA for the genus.

ASV Sequence	Family	Genus
TACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCAGGCGGTTTGTTAAGTCAGATGTGAAATCCCCGGGCTCAACCTGGGAACTGCATCTGATACTGGCAAGCTTGAGTCTCGTAGAGGGGGGTAGAATTCCAGG	Enterobacteriaceae	Escherichia/Shigella
TACGTATGGGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGGCGGCATGGCAAGTCAGAAGTGAAAGCCTGGGGCTCAACCCCGGAATTGCTTTTGAAACTGTCAGGCTAGAGTGTCGGAGGGGTAAGCGGAATTCCTAG	Lachnospiraceae	NA
TACGTAGGTGGCAAGCGTTGTCCGGATTTACTGGGTGTAAAGGGCGTGTAGGCGGAGAAGCAAGTCAGAAGTGAAATCCATGGGCTTAACCCATGAACTGCTTTTGAAACTGTTTCCCTTGAGTATCGGAGAGGCAGGCGGAATTCCTAG	Ruminococcaceae	Ruminococcaceae (UCG-005)
TACGGAAGATGCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGTGCGTAGGCGGAAGAATAAGTCAGCGGTGAAATGCTTCAGCTCAACTGGAGAATTGCCGATGAAACTGTTTTTCTAGAGTATAAAAGAGGTATGCGGAATGCGTGG	Barnesiellaceae	Coprobacter
TACGGAGGATCCAAGCGTTATCCGGATTTATTGGGTTTAAAGGGTGCGTAGGCGGTTTAGTAAGTCAGCGGTGAAATTTTGGTGCTTAACACCAAACGTGCCGTTGATACTGCTGGGCTAGAGAGTAGTTGCGGTAGGCGGAATGTATGG	Rikenellaceae	Alistipes
TACGGAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGATGGATGTTTAAGTCAGTTGTGAAAGTTTGCGGCTCAACCGTAAAATTGCAGTTGATACTGGATGTCTTGAGTGCAGTTGAGGCAGGCGGAATTCGTGG	Bacteroidaceae	Bacteroides
TACGGAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGGTGGACTGGTAAGTCAGTTGTGAAAGTTTGCGGCTCAACCGTAAAATTGCAGTTGATACTGTCAGTCTTGAGTACAGTAGAGGTGGGCGGAATTCGTGG	Bacteroidaceae	Bacteroides
TACGTAGGGGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGTTTTGCAAGTCTGAAGTGAAAGCCCGGGGCTTAACCCCGGGACTGCTTTGGAAACTGTAGAACTAGAGTGCAGGAGAGGTAAGTGGAATTCCTAG	Lachnospiraceae	Lachnospiraceae (NK4A136_group)
TACGTAGGGGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGTGCGTAGGTGGCAAGGCAAGTCAGATGTGAAAGCCCGGGGCTCAACCCCGGTACTGCATTTGAAACTGTCTAGCTAGAGTGCAGGAGAGGTAAGCGGAATTCCTAG	Lachnospiraceae	NA
TACGTAGGGGGCAAGCGTTGTCCGGAATTATTGGGCGTAAAGAGTACGTAGGCGGTTTGCTAAGCGCAAGGTGAAAGGCAGTGGCTTAACCATTGTAAGCCTTGCGAACTGACAGACTTGAGTGCAGGAGAGGAAAGCGGAATTCCTAGT	Family_XIII	Family_XIII_UCG-001
TACGTAGGGGGCGAGCGTTGTCCGGAATGATTGGGCGTAAAGGGCGCGTAGGCGGCCTGCTAAGTCTGGAGTGAAAGTCCTGCTTTCAAGGTGGGAATTGCTTTGGATACTGGTGGGCTGGAGTGCAGGAGAGGAAAGCGGAATTACCGG	Christensenellaceae	Christensenellaceae (R-7_group)
TACGTAGGTGGCAAGCGTTGTCCGGATTTACTGGGTGTAAAGGGCGTGCAGCCGGGTCTGCAAGTCAGATGTGAAATCCATGGGCTCAACCCATGAACTGCATTTGAAACTGTAGATCTTGAGTGTCGGAGGGGCAATCGGAATTCCTAG	Ruminococcaceae	Ruminococcaceae (UCG-002)
TACGTAGGTGGCAAGCGTTGTCCGGATTTACTGGGTGTAAAGGGCGTGTAGGCGGGATTGCAAGTCAGGCGTGAAAACCAGGGGCTCAACCTCTGGCCTGCGTTTGAAACTGTAGTTCTTGAGTACTGGAGAGGTTGACGGAATTCCTAG	Ruminococcaceae	NA
TACGTATGGTGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGCAGGCGGTGCGGCAAGTCTGATGTGAAAGCCCGGGGCTCAACCCCGGTACTGCATTGGAAACTGTCGTACTAGAGTGTCGGAGGGGTAAGTGGAATTCCTAG	Lachnospiraceae	Agathobacter

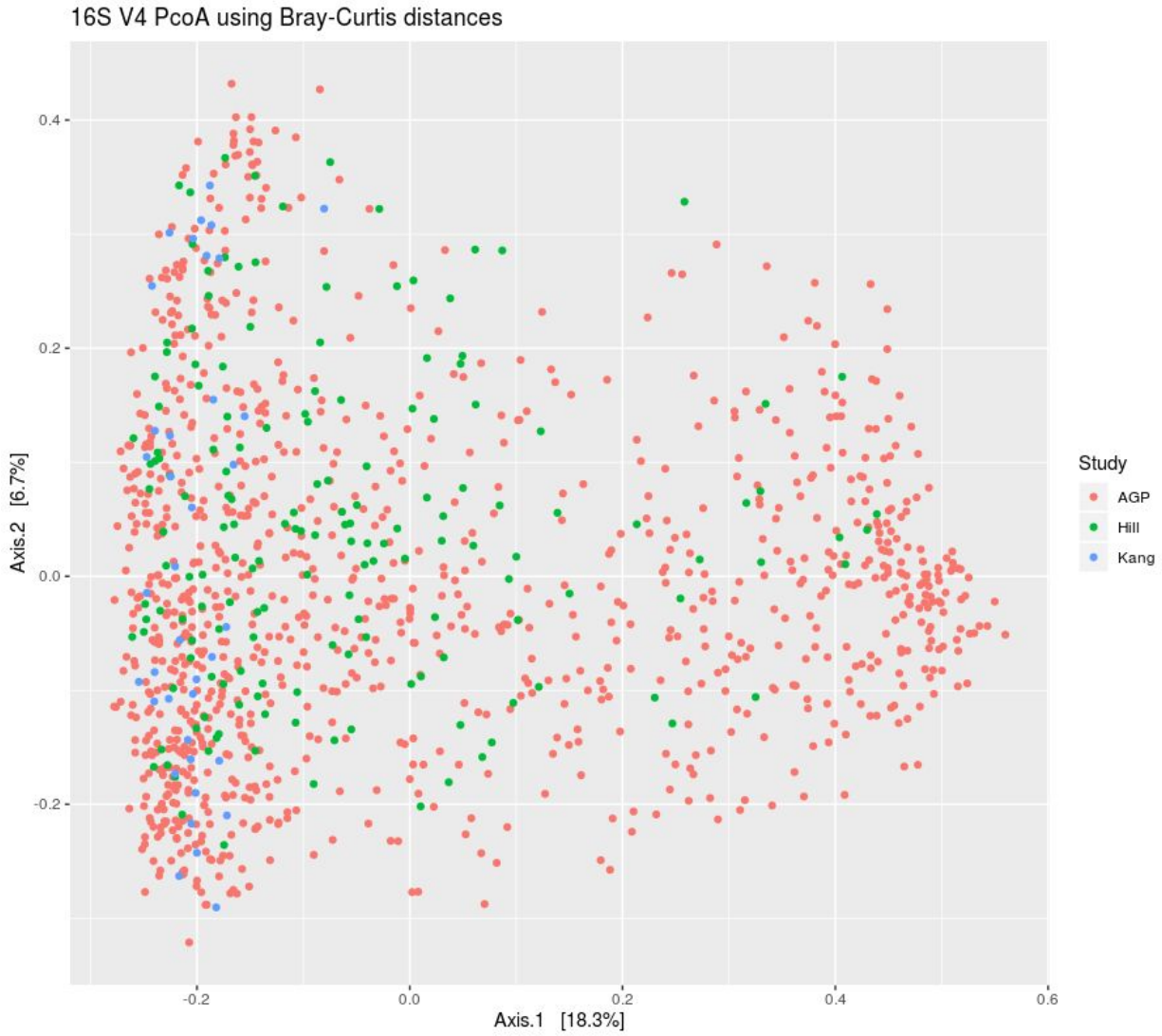


Figure 8 Unconstrained PCoA of all 1226 Samples by Study using CSS Normalization

Samples were colored according to the study they came from. AGP stands for "American Gut Project".

Figure 9 Unconstrained PCoA of all 1226 Samples by Phenotype using CSS Normalization

The label, "AR", found in the legend stands for "Anxiety-Related" and refers to individuals with anxiety-related conditions.

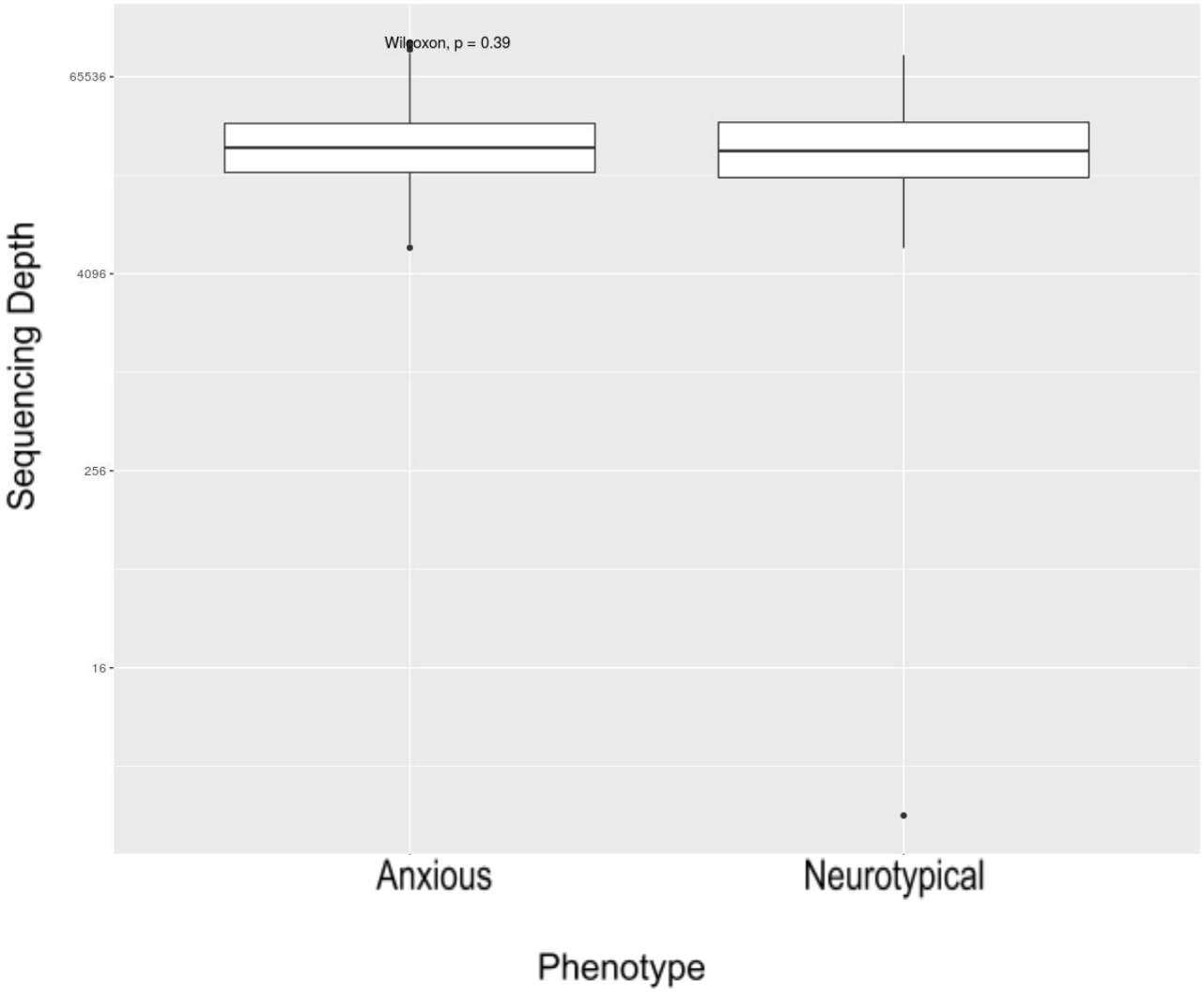


Figure 10 Boxplot of Sequence Depth of both Phenotypes

As seen above, the p-value of the wilcoxon-ranked test was insignificant at 0.39.

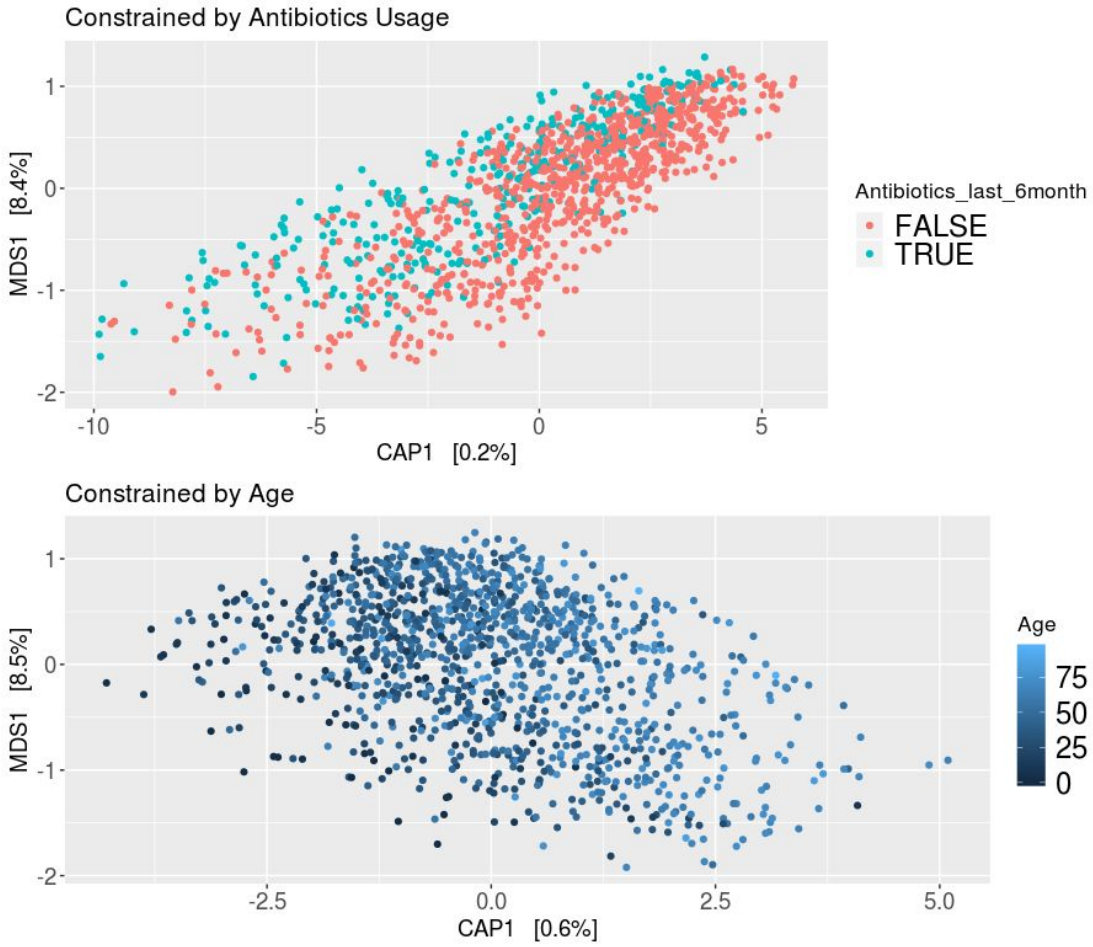


Figure 11 Constrained PCoA of Antibiotics and Age

"Antibiotics_last_6month" refers to whether or not antibiotics was taken within the last 6 months.

Data Accession

Kang Study: Available in the open-source microbiome database “Qiita” with the study ID number 10532 (https://qiita.microbio.me)

Hill Study: Sequence and metadata are in EBI and NCBI under the accession number ERP016332.

American Gut Project: Sequence and metadata are in the NCBI under study accession number PRJEB11419.

References

Ai, Dongmei, Hongfei Pan, Rongbao Han, Xiaoxin Li, Gang Liu, and Li C. Xia. 2019. "Using Decision Tree Aggregation with Random Forest Model to Identify Gut Microbes Associated with Colorectal Cancer." *Genes* 10 (2). https://doi.org/10.3390/genes10020112.

Albert, Paul R., and Chawki Benkelfat. 2013. "The Neurobiology of Depression—revisiting the Serotonin Hypothesis. II. Genetic, Epigenetic and Clinical Studies." *Philosophical Transactions of the Royal Society B: Biological Sciences*. https://doi.org/10.1098/rstb.2012.0535.

Amir, Amnon, Daniel McDonald, Jose A. Navas-Molina, Evguenia Kopylova, James T. Morton, Zhenjiang Zech Xu, Eric P. Kightley, et al. 2017. "Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns." *mSystems* 2 (2). https://doi.org/10.1128/mSystems.00191-16.

Anderson, Marti J. 2017. "Permutational Multivariate Analysis of Variance (PERMANOVA)." *Wiley StatsRef: Statistics Reference Online*. https://doi.org/10.1002/9781118445112.stat07841.

Bilgiç, Ayhan, Serhat Türkoğlu, Ozlem Ozcan, Ali Evren Tufan, Savaş Yılmaz, and Tuğba Yüksel. 2013. "Relationship between Anxiety, Anxiety Sensitivity and Conduct Disorder Symptoms in Children and Adolescents with Attention-Deficit/hyperactivity Disorder (ADHD)." *European Child & Adolescent Psychiatry* 22 (9): 523–32.

Callahan, Benjamin J., Paul J. McMurdie, Michael J. Rosen, Andrew W. Han, Amy Jo Johnson, and Susan P. Holmes. 2015. "DADA2: High Resolution Sample Inference from Amplicon Data." *bioRxiv*. https://doi.org/10.1101/024034.

Callahan, Ben J., Kris Sankaran, Julia A. Fukuyama, Paul J. McMurdie, and Susan P. Holmes. 2016. "Bioconductor Workflow for Microbiome Data Analysis: From Raw Reads to Community Analyses." *F1000Research* 5 (June): 1492.

Carabottia, Marilia, Annunziata Sciroccoa, Maria Antonietta Masellib, and Carola Severia. 2015. "The Gut-Brain Axis: Interactions between Enteric Microbiota, Central and Enteric Nervous Systems." *Annales de Gastroenterologie et D'hepatologie* 28: 1–7.

Case, Rebecca J., Yan Boucher, Ingela Dahllöf, Carola Holmström, W. Ford Doolittle, and Staffan Kjelleberg. 2007. "Use of 16S rRNA and rpoB Genes as Molecular Markers for Microbial Ecology Studies." *Applied and Environmental Microbiology* 73 (1): 278–88.

Chen, Jack J., and Laura Marsh. 2014. "Anxiety in Parkinson's Disease: Identification and Management." *Therapeutic Advances in Neurological Disorders* 7 (1): 52–59.

Cheung, Stephanie G., Ariel R. Goldenthal, Anne-Catrin Uhlemann, J. John Mann, Jeffrey M. Miller, and M. Elizabeth Sublette. 2019. "Systematic Review of Gut Microbiota and Major Depression." *Frontiers in Psychiatry / Frontiers Research Foundation* 10 (February): 34.

Cryan, John F., Kenneth J. O'Riordan, Caitlin S. M. Cowan, Kiran V. Sandhu,

Thomaz F. S. Bastiaanssen, Marcus Boehme, Martin G. Codagnone, et al. 2019. "The Microbiota-Gut-Brain Axis." *Physiological Reviews* 99 (4): 1877–2013.
Danzinger, Paula R., and Elizabeth Reynolds Welfel. 2000. "Age, Gender and Health Bias in Counselors: An Empirical Analysis." *Journal of Mental Health Counseling* 22 (2): 135.
Dominianni, Christine, Rashmi Sinha, James J. Goedert, Zhiheng Pei, Liying Yang, Richard B. Hayes, and Jiyoung Ahn. 2015. "Sex, Body Mass Index, and Dietary Fiber Intake Influence the Human Gut Microbiome." *PloS One* 10 (4): e0124599.
Foster, Jane A., and Karen-Anne McVey Neufeld. 2013. "Gut–brain Axis: How the Microbiome Influences Anxiety and Depression." *Trends in Neurosciences* 36 (5): 305–12.
Gao, Kan, Chun-Long Mu, Aitak Farzi, and Wei-Yun Zhu. 2019. "Tryptophan Metabolism: A Link Between the Gut Microbiota and Brain." *Advances in Nutrition* , December. https://doi.org/10.1093/advances/nmz127.
Hankin, Benjamin L. 2009. "Development of Sex Differences in Depressive and Co-Occurring Anxious Symptoms during Adolescence: Descriptive Trajectories and Potential Explanations in a Multiwave Prospective Study." *Journal of Clinical Child and Adolescent Psychology: The Official Journal for the Society of Clinical Child and Adolescent Psychology, American Psychological Association, Division 53* 38 (4): 460–72.
Hill-Burns, Erin M., Justine W. Debelius, James T. Morton, William T. Wissemann, Matthew R. Lewis, Zachary D. Wallen, Shyamal D. Peddada, et al. 2017. "Parkinson's Disease and Parkinson's Disease Medications Have Distinct Signatures of the Gut Microbiome." *Movement Disorders: Official Journal of the Movement Disorder Society* 32 (5): 739–49.
Huang, Ting-Ting, Jian-Bo Lai, Yan-Li Du, Yi Xu, Lie-Min Ruan, and Shao-Hua Hu. 2019. "Current Understanding of Gut Microbiota in Mood Disorders: An Update of Human Studies." *Frontiers in Genetics* 10 (February): 98.
Jašarević, Eldin, Kathleen E. Morrison, and Tracy L. Bale. 2016. "Sex Differences in the Gut Microbiome–brain Axis across the Lifespan." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 371 (1688): 20150122.
Joseph, N., C. Paulson, H. Corrada Bravo, and M. Pop. 2013. "Robust Methods for Differential Abundance Analysis in Marker Gene Surveys." *Nature Methods* 10: 1200–1202.
Kang, Dae-Wook, James B. Adams, Ann C. Gregory, Thomas Borody, Lauren Chittick, Alessio Fasano, Alexander Khoruts, et al. 2017. "Microbiota Transfer Therapy Alters Gut Ecosystem and Improves Gastrointestinal and Autism Symptoms: An Open-Label Study." *Microbiome*. https://doi.org/10.1186/s40168-016-0225-7.
Kang, Dae-Wook, Jin Gyoon Park, Zehra Esra Ilhan, Garrick Wallstrom, Joshua Labaer, James B. Adams, and Rosa Krajmalnik-Brown. 2013. "Reduced Incidence of Prevotella and Other Fermenters in Intestinal Microflora of Autistic Children." *PloS One* 8 (7): e68322.

Kuczynski, J., J. Stombaugh, W. A. Walters, A. González, J. G. Caporaso, and R. Knight. 2005. “Using QIIME to Analyze 16S rRNA Gene Sequences from Microbial Communities, in Current Protocols in Microbiology.” Wiley.

Liu, Lu, and Gang Zhu. 2018. “Gut–Brain Axis and Mood Disorder.” *Frontiers in Psychiatry / Frontiers Research Foundation* 9: 223.

Loomba, Rohit, Victor Seguritan, Weizhong Li, Tao Long, Niels Klitgord, Archana Bhatt, Parambir Singh Dulai, et al. 2017. “Gut Microbiome-Based Metagenomic Signature for Non-Invasive Detection of Advanced Fibrosis in Human Nonalcoholic Fatty Liver Disease.” *Cell Metabolism*. https://doi.org/10.1016/j.cmet.2017.04.001.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15 (12): 550.

Mandal, Siddhartha, Will Van Treuren, Richard A. White, Merete Eggesbø, Rob Knight, and Shyamal D. Peddada. 2015. “Analysis of Composition of Microbiomes: A Novel Method for Studying Microbial Composition.” *Microbial Ecology in Health and Disease* 26 (May): 27663.

Martin, Clair R., Vadim Osadchiy, Amir Kalani, and Emeran A. Mayer. 2018. “The Brain-Gut-Microbiome Axis.” *Cellular and Molecular Gastroenterology and Hepatology* 6 (2): 133–48.

Mayer, Emeran A., Kirsten Tillisch, and Arpana Gupta. 2015. “Gut/brain Axis and the Microbiota.” *The Journal of Clinical Investigation* 125 (3): 926–38.

McDonald, Daniel, Embriette Hyde, Justine W. Debelius, James T. Morton, Antonio Gonzalez, Gail Ackermann, Alexander A. Aksenov, et al. 2018. “American Gut: An Open Platform for Citizen Science Microbiome Research.” *mSystems* 3 (3). https://doi.org/10.1128/mSystems.00031-18.

Neufeld, K. M., N. Kang, J. Bienenstock, and J. A. Foster. 2011. “Reduced Anxiety-like Behavior and Central Neurochemical Change in Germ-Free Mice.” *Neurogastroenterology & Motility*. https://doi.org/10.1111/j.1365-2982.2010.01620.x.

Paradis, E., J. Claude, and K. Strimmer. 2004. “APE: Analyses of Phylogenetics and Evolution in R Language.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btg412.

Paulson, Joseph N., O. Colin Stine, Héctor Corrada Bravo, and Mihai Pop. 2013. “Differential Abundance Analysis for Microbial Marker-Gene Surveys.” *Nature Methods* 10 (12): 1200–1202.

Schliep, Klaus P. 2012. “Estimating Phylogenetic Trees with Phangorn (Version 1.7-0).” https://mran.microsoft.com/snapshot/2014-09-26/web/packages/phangorn/vignettes/Trees.pdf.

Schloss, Patrick D., Sarah L. Westcott, Thomas Ryabin, Justine R. Hall, Martin Hartmann, Emily B. Hollister, Ryan A. Lesniewski, et al. 2009. “Introducing Mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities.” *Applied and Environmental Microbiology* 75 (23): 7537–41.

Schnorr, Stephanie L., and Harriet A. Bachner. 2016. “Integrative Therapies in Anxiety Treatment with Special Emphasis on the Gut Microbiome.” *The Yale Journal of Biology and Medicine* 89 (3): 397–422.

Strang, John F., Lauren Kenworthy, Peter Daniolos, Laura Case, Meagan C. Wills, Alex Martin, and Gregory L. Wallace. 2012. “Depression and Anxiety Symptoms in Children and Adolescents with Autism Spectrum Disorders without Intellectual Disability.” *Research in Autism Spectrum Disorders* 6 (1): 406–12.

Vuong, Helen E., and Elaine Y. Hsiao. 2017. “Emerging Roles for the Gut Microbiome in Autism Spectrum Disorder.” *Biological Psychiatry* 81 (5): 411–23.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer.

Winter, Gal, Robert A. Hart, Richard P. G. Charlesworth, and Christopher F. Sharpley. 2018. “Gut Microbiome and Depression: What We Know and What We Need to Know.” *Reviews in the Neurosciences* 29 (6): 629–43.

Wright, Eriks, Erik, and S. Wright. 2016. “Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R.” *The R Journal*. https://doi.org/10.32614/rj-2016-025.

Xu, Mingyu, Xuefeng Xu, Jijun Li, and Fei Li. 2019. “Association Between Gut Microbiota and Autism Spectrum Disorder: A Systematic Review and Meta-Analysis.” *Frontiers in Psychiatry / Frontiers Research Foundation* 10 (July): 473.

Yang, Bo, Yong Wang, and Pei-Yuan Qian. 2016. “Sensitivity and Correlation of Hypervariable Regions in 16S rRNA Genes in Phylogenetic Analysis.” *BMC Bioinformatics* 17 (March): 135.

Yano, Jessica M., Kristie Yu, Gregory P. Donaldson, Gauri G. Shastri, Phoebe Ann, Liang Ma, Cathryn R. Nagler, Rustem F. Ismagilov, Sarkis K. Mazmanian, and Elaine Y. Hsiao. 2015. “Indigenous Bacteria from the Gut Microbiota Regulate Host Serotonin Biosynthesis.” *Cell* 161 (2): 264–76.

Zaheer, Rahat, Noelle Noyes, Rodrigo Ortega Polo, Shaun R. Cook, Eric Marinier, Gary Van Domselaar, Keith E. Belk, Paul S. Morley, and Tim A. McAllister. 2018. “Impact of Sequencing Depth on the Characterization of the Microbiome and Resistome.” *Scientific Reports* 8 (1): 5890.

Zhang, Mengxiang, Wei Ma, Juan Zhang, Yi He, and Juan Wang. 2018. “Analysis of Gut Microbiota Profiles and Microbe-Disease Associations in Children with Autism Spectrum Disorders in China.” *Scientific Reports* 8 (1): 13981.

Zhou, Yongjie, Zhongqiang Cao, Mei Yang, Xiaoyan Xi, Yiyang Guo, Maosheng Fang, Lijuan Cheng, and Yukai Du. 2017. “Comorbid Generalized Anxiety Disorder and Its Association with Quality of Life in Patients with Major Depressive Disorder.” *Scientific Reports* 7 (January): 40511.

ACKNOWLEDGEMENTS

I would like to thank Dr. Maude David and everyone in the David Lab for their mentorship and guidance. I have learned so much throughout my years in this lab, and I am extremely blessed to have had them as teachers as I learned about the world of bioinformatics.

I would also like to thank the SURE program and the Summer Undergraduate Pharmacy Fellowship, which both allowed me to learn the skills necessary for this thesis. Without these, I would have not been able to tackle this project.

Lastly, I would like to thank all my friends in the Running Club, in Isang Bansang Pilipino, and my family for their continued support throughout my years at OSU. I would have been overloaded with stress and anxiety had it not been for them.