ABSTRACT

SONG, KUNCHENG. Bioinformatics and Machine Learning in Human Microbiome Analysis. (Under the direction of Dr. Yi-Hui Zhou).

Microbiome, particularly the microbes that reside within human environments, has been living, sharing, and evolving quietly along with the development of Homo sapiens in obscurity. The first recorded observation of microorganisms was in the latter half of the seventeenth century. Not until 1876, when Dr. Robert Koch isolated and cultivated these "invisible" organisms, scientists slowly yet steadily unraveled the importance of these microorganisms and their fascinating stories that tides closely to human health.

In the past two decades, with the advancement of sequencing technology, we have begun to reveal the reciprocal and imperative host-microbe relationship. Within the sequencing realm of microbiome data sources, there are predominately two methods, 16S ribosomal ribonucleic acid (rRNA) and shotgun sequencing, and while there are multiple platforms for sequencing, short-read sequencing is still the most cost-effective choice. Both 16S rRNA and shotgun sequencing reads can generate taxonomical profiling of the microbe communities for the samples. Generally, the taxonomic profiling results undergo different algorithms or methods to extract the critical microbes positively or negatively associated with a phenotype by comparing the profiles between disease and control samples. A wide range of machine learning models was examined and developed for more accurate phenotype predictions.

In **Chapter 1,** we investigated the background and rationale behind the chapters in this thesis. We systematically evaluate how different assignment methods, filtering, and normalization affect the downstream machine learning performance in **Chapter 2**. We present a novel method and bioinformatics tool to characterize the microbial structures, including processing, evaluating, and visualizing the taxon-taxon networks between the healthy and control

samples in **Chapter 3**. Lastly, we describe an ensemble approach of machine learning methods with differential abundance analyses to gain a better predictive power for internal and external validation in **Chapter 4**.

Bioinformatics and Machine Learning in Human Microbiome Analysis

by
Kuncheng Song

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Bioinformatics

Raleigh, North Carolina
2022

APPROVED BY:

Dr.Yi-Hui Zhou Committee Chair	Dr. Fred Wright
Dr. Benjamin Callahan	Dr. Xinxia Peng

DEDICATION

To my beloved family!

BIOGRAPHY

Kuncheng was born in Beijing, China, and grew up in this metropolitan city until he finished middle school. In 2007, Kuncheng attended the King's School in Australia, where he finished his high school education. In 2010, Kuncheng came to the United States to pursue a bachelor's degree at Syracuse University, and he graduated with a double major in Biology and Nutrition Science. Kuncheng continued his education journal at Boston University, where he completed a Master of Public Health degree with dual concentrations in Epidemiology and Biostatistics. After graduation, Kuncheng worked as a biostatistician for a year. During this time, Kuncheng was intrigued by the computational aspects of biological research and decided to pursue a Master of Science degree in Bioinformatics from Johns Hopkins University, followed by a Ph.D. in Bioinformatics at North Carolina State University.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to Dr. Yi-Hui Zhou for her guidance and wisdom throughout my time at North Carolina State University. She has been a true mentor for me throughout my doctoral training. Dr. Zhou provided me with many opportunities for career building and allowed me to explore my research interests. I would also like to thank my committee members, Drs. Benjamin Callahan, Fred Wright, and Xinxia Peng for their guidance. I also want to extend my gratitude to Dr. Spencer Muse for leading an excellent bioinformatics program. I would like to acknowledge the support from Chris Smith, Dana Ripperton, Jenni Wilson, and Kevin Dudley for technical and advisory support at the Bioinformatics Research Center.

Finally, I am thankful for the support from my family and friends for their unconditional love and prescient guidance throughout my educational journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 2:

CHAPTER 3:

CHAPTER 4:

Chapter 1

Introduction

Background and Challenges

With the advancement and reduced cost of the Next Generation Sequencing (NGS) technologies, microbial sequencing-based studies became more prevalent in detecting patterns among human samples for various diseases. As the research efforts grow, we can discover not only the relationship between gastrointestinal diseases (Bonder et al., 2016) but also metabolic diseases (obesity (Goodrich et al., 2014) and diabetes (Zhou et al., 2019)), allergic disease (Vatanen et al., 2019), and more. There are multiple objectives in understanding the microbiome, including learning the microbe-host relationship, correctly characterizing the microbial composition, and for our research goal, the use of the compositional microbiome data to discover the hidden connections between the microbiome composition and host phenotypes.

There are multiple challenges presented in the goals mentioned above. We can categorize them into three different areas: raw microbiome data processing and optimization, microbe-microbe association within and between phenotypes, and machine learning algorithms optimization for consistent predictive performance. We will address all these issues in the following chapters.

Our Approaches

While different microbial sequencing types are available, the most cost-effective and prevalent method is 16S rRNA gene sequencing, which targets the variable regions of the microbial species. The essential ideology behind this technology for taxonomic profiling is the use of ‘universal’ PCR primers, which bind and amplify ‘all’ microbial 16S rRNA genes. The more prevalent species will amplify more, and these read abundances will later be translated into a compositional representation of the source gut microbiota. Next, there are many taxonomic assignment methods for processing the 16S rRNA data, all of which use different algorithms to

assign the sequencing read to the closest or best matching taxonomic profile. However, there are reservations about using filtering and normalization techniques for the microbial data prior to machine learning models as these factors can affect the performance of different types of machine learning models, and we explore these effects in **Chapter 2.**

In **Chapter 3**, we present a novel method named correlation and consensus-based cross-taxonomy network analysis (C3NA) for extracting and refining the taxa-taxa relationship for a phenotype. This is the first method that methodically explores the microbial species' modularization, which groups the taxa into different modular groups based on their correlation and shared connections. C3NA utilizes modular preservation analysis to compare two phenotypes and extract the differential taxa network and correlations that are only present in one of the phenotypes.

Finally, in **Chapter 4**, we proposed an ensemble method of machine learning methods accompanied by feature selections from differential abundance analyses for better external validations. Our results suggested the importance of filtering the number of taxa prior to a machine learning model to enhance the predictive performance. The filtering criteria were evaluated using differential abundance analyses, and the results supported enhancement in performance. The performance can be maximized through an ensemble method of Random Forest and LASSO regression.

References

Bonder, M. J., Kurilshikov, A., Tigchelaar, E. F., Mujagic, Z., Imhann, F., Vila, A. V., Deelen, P., Vatanen, T., Schirmer, M., Smeekens, S. P., Zhernakova, D. V., Jankipersadsing, S. A., Jaeger, M., Oosting, M., Cenit, M. C., Masclee, A. A. M., Swertz, M. A., Li, Y., Kumar, V., … Zhernakova, A. (2016). The effect of host genetics on the gut microbiome. *Nature Genetics*, *48*(11), 1407–1412. https://doi.org/10.1038/ng.3663

Goodrich, J. K., Waters, J. L., Poole, A. C., Sutter, J. L., Koren, O., Blekhman, R., Beaumont, M., Treuren, W. Van, Knight, R., Bell, J. T., Spector, T. D., Clark, A. G., & Ley, R. E. (2014). Human genetics shape the gut microbiome. *Cell*, *159*(4), 789. https://doi.org/10.1016/J.CELL.2014.09.053

Vatanen, T., Plichta, D. R., Somani, J., Münch, P. C., Arthur, T. D., Hall, A. B., Rudolf, S., Oakeley, E. J., Ke, X., Young, R. A., Haiser, H. J., Kolde, R., Yassour, M., Luopajärvi, K., Siljander, H., Virtanen, S. M., Ilonen, J., Uibo, R., Tillmann, V., … Xavier, R. J. (2019). Genomic variation and strain-specific functional adaptation in the human gut microbiome during early life. *Nature Microbiology*, *4*(3), 470–479. https://doi.org/10.1038/s41564-018-0321-5

Zhou, W., Sailani, M. R., Contrepois, K., Zhou, Y., Ahadi, S., Leopold, S. R., Zhang, M. J., Rao, V., Avina, M., Mishra, T., Johnson, J., Lee-McMullen, B., Chen, S., Metwally, A. A., Tran, T. D. B., Nguyen, H., Zhou, X., Albright, B., Hong, B. Y., … Snyder, M. (2019). Longitudinal multi-omics of host–microbe dynamics in prediabetes. *Nature*, *569*(7758), 663–671. https://doi.org/10.1038/s41586-019-1236-x

Chapter 2

Systematic Comparisons for Composition Profiles, Taxonomic Levels, and Machine Learning Methods for Microbiome-Based Disease Prediction

Reproduced with minor reformatting from: Song, K., Wright, F. A., & Zhou, Y. H. (2020). Systematic comparisons for composition profiles, taxonomic levels, and machine learning methods for microbiome-based disease prediction. Frontiers in molecular biosciences, 423. doi.org/10.3389/fmolb.2020.610845

1 Introduction

With the advancement of sequencing technology and the downward trends in the cost of sequencing, more studies have used microbiome data as a primary source for investigating the relationship between microbiota and host health. In general, human microbiota samples consist of easily collected specimens such as feces, saliva, and skin. Upon collection, the sample can undergo a variety of extraction protocols, including protein, RNA, and DNA. Each of these data types has led to a specific field of emerging research (Weinstock, 2012). In this review, we focus on the targeted extraction of microbial DNA from the16S rRNA region, which is present in most microorganisms but displays high variability across species. The sequenced reads are then typically clustered into Operational Taxonomic Units (OTUs) by matching the reads to a reference database.

Multiple studies have investigated the use of OTUs for phenotype/disease prediction, including inflammatory bowel diseases (Gevers et al., 2014), Type 2 diabetes (Gurung et al., 2020), and lung cancer (Zheng et al., 2020). As a variety of data treatment and prediction methods have been used, there is a pressing need to connect and verify how the upstream processing of the 16S rRNA data affects the downstream prediction performance and compare the different OTU/ASV methods.

There are two primary representations to produce data count matrices: OTUs and Amplicon Sequence Variants (ASVs) (Rosen et al., 2012). Within the realm of OTUs, there are three methods to “cluster” sequences into OTUS: *de novo*, closed-reference, and open-reference, each with its unique advantages and disadvantages depending on the sequence region, reference database, and sample environment (Rideout et al., 2014). ASVs are commonly generated using the Divisive Amplicon Denoising Algorithm 2 (DADA2), and the resultant ASVs represent true

biological sequences obtained from reads (Callahan et al., 2016). In addition, there have been recent efforts to use the occurrence of short-chain k-mer (15~30-mer) (Molik et al., 2020), and very short-chain k-mers (<10-mer) (Asgari et al., 2018, 2019), within reads that offer a unique reference-free and alignment-free approach to provide a data representation upon which a phenotype prediction model is built. We have included both of these k-mer approaches in our review to compare them directly with the OTU/ASV assignment methods.

Additional procedures for handling the OTUs or ASVs include filtering (Duvallet et al., 2017; Goodrich et al., 2014; Zhou & Gallins, 2019) and normalization (McMurdie & Holmes, 2014; Weiss et al., 2017). We included both practices to show the result from different combinations.

Overall, we conducted a systematic review of how different combinations of (i) OTU/ASV assignment methods and k-mer lengths, (ii) the use of normalization and filtering, and (iii) choices of machine learning methods, among eleven commonly used approaches, all affect the prediction accuracy for complex host traits.

2 Methods

2.1 Raw Sequence Data

2.1.1 Inflammatory Bowel Diseases Dataset

This microbiome dataset includes host phenotypes of Crohn's disease, with microbiome data from 16S rRNA gene (V4) sequencing on the Illumina MiSeq platform (version 2) with 175 bp paired-end reads (Gevers et al., 2014). In brief, the samples were collected from 28 participating pediatric gastroenterology centers in North America between 2008 and 2012. Within the metadata, there are three disease diagnoses described: Crohn's Disease (CD), Ulcerative Colitis (UC), Ischemic Colitis (IC), and control. Each of the disease diagnoses was compared separately

to the control group. The data were downloaded from the European Nucleotide Archive (ENA), accession PRJEB13679. The available FASTQ file format is a single-end layout; the QIIME2 pipeline for the microbiota analyses was processed as single-end reads. The full processing workflow is described in the Supplemental Material under Data Processing and **Figure S1**. A summary of the basic patient characteristics for the datasets is provided in **Table 1**.

2.1.2 TwinsUK Dataset

This microbiome dataset contains 1,081 fecal samples collected from 997 individuals, all of whom underwent 16S rRNA-based sequencing. The raw sequences were retrieved from the European Nucleotide Achieve (ENA) accession IDs PRJEB6702 and PRJEB6705. The collection and processing of the data were described previously (Goodrich et al., 2014). The fecal samples were obtained by the participants from their households and stored in a refrigerator up to 2 days prior to the twins' annual visit to King's College London, where the samples were stored at $-80°C$ until the following process. The DNA was extracted from the provided samples, and the 16S rRNA genes (V4) were amplified from bulk DNA through PCR. The sequencing steps were performed on the Illumina MiSeq 2x250 bp platform. The available FASTQ file format is a single-end layout; the QIIME2 pipeline for the microbiota analyses was processed as single-end reads. The full processing is described in section 1 of the Supplementary Material under Data Processing. A summary of the basic patient characteristics is included in **Table 1**.

2.2 Sample Processing

The detailed sample processing is also listed in the Data Processing section 1 in the Supplementary Material, and the workflow is shown in **Figure 1**. Our analyses can be summarized in four stages. In the first stage, we extracted the OTUs/ASVs using QIIME2 (Bolyen et al., 2019). We then collapsed count matrices at OTUs/ASVs levels to higher

taxonomic order, including phylum, class, order, family, genus, and species. At the same time, we also extracted the very short-chain k-mers and short-chain k-mers directly from raw FASTQ files. In the second stage, we used the DESeq2 package in R to apply normalization to the OTUs/ASVs count matrices and the very short-chain k-mers (Weiss et al., 2017), or did not apply normalization. Short-chain k-mer (15-mer, 21-mer, and 30-mer) were omitted from this analysis because of the large matrix dimensions when including all observed short k-mers. In the third stage, we applied (or did not) a common filtering criterion as follows. The first filter excludes samples with fewer than 100 reads, and the second filter subtracts OTUs with fewer than 10 reads (Duvallet et al., 2017; Zhou & Gallins, 2019). The third filter removes OTUs that are present in fewer than 5% of samples. In the last stage, we applied eleven commonly used machine learning algorithms to the different combinations. Overall, we conducted 1,353 combinations per phenotype and 5,412 total combinations for four diseases against their respective controls.

In our extensive analyses, 102 combinations failed to return any useful results. Thirteen of these involved elastic nets, five used neural networks, and the remaining 84 combinations used logistic regression. These failed runs were likely due to the algorithms being unable to converge.

2.3 Evaluation of Prediction Accuracy

We compared the four disease types prediction at each taxonomic and k-mer level through the Area Under the Curve (AUC) for the Receiver Operating Characteristics (ROC) curve, which is commonly used to evaluate the prediction accuracy for binary traits. The ROC is a plot with the True Positive Rate (Sensitivity) compared to the False Positive Rate (1 – Specificity). Also, we can calculate the Area Under the Precision-Recall Curve (AUPR), which is another way to evaluate the prediction with a plot of recall against precision. We utilized two evaluating

parameters to quantify the prediction ability of a balanced and imbalanced dataset. Ideally, we would want these two values to be both high to indicate good discrimination between the disease and the controls. The full summary of the combinations is in Supplementary **Table S1**. In the following discussion, we focus primarily on the AUC (full abbreviation AUROC for Area Under the ROC) as a performance measure, as it offers a more distinct contrast among combinations compared to the AUPR for our comparisons.

2.4 Evaluation of Machine Learning Results

To investigate the consistency between the feature selected from machine learning algorithms and the discovery studies, we extracted the useful information from the machine learning algorithm outputs and compared them to taxa previously identified as significantly associated with IBD. We based our comparisons on three separate publications. First, we selected eleven critical taxa identified from the original study for our IBD dataset (Gevers et al., 2014): two from the Order level and nine from the Family level. All of these were identified in our results, with the exception of a Family-level assignment *Gemellaceae,* and we used the Genus-level assignment *Gemella* as a substitute. Also, we chose another study that had examined the microbiota associated with IBD; the authors had identified multiple taxa associated with either increased or decreased changes in IBD (Glassner et al., 2020), and we selected nine taxa from the list. The Genus *Bacteroides* and Eubacterium have multiple subgroups and the *Pectinophilus* group was selected for *Bacteroides*, and the *nodatum* group was selected as a stand-in for Eubacterium, and both passed our filtering procedures. In the last example the authors had used a linear discriminant analysis effect size approach to determine three important taxa, two from the Family-level and one from the Order level, all of which are present in our features (Kim et al., 2019).

We focused on two of the most consistent machine learning methods, random forest and xgBoost, and two methods with less consistent performance, a support vector machine and logistic regression. The definition of "important" features is different depending on the method. Each of the features was selected within the 100 iterations of 5-fold cross-validation. As a result, the numeric representation of "important" features here represents an average of over 500 training and testing loops.

The results on the presence of taxa are shown in **Table 2**. Overall, only DADA2 was able to pick all these taxa, while other OTU assignment methods missed a few. After filtering, most of the taxa listed were removed; out of the 22 taxa, only nine taxa remained, and these nine taxa were present in all OTU/ASV assignment methods, except *Eubacterium* with the *nodatum* group that was missing when using the *de novo* method.

2.4.1 Defining Important Features

There are many ways to define the features of machine learning models that are important to the model. For illustrative purposes, we will focus on only one of the available ways to select important features. With xgBoost, we extracted the "Gain" result from the xgBoost output to evaluate the importance of the features (Chen & Guestrin, 2016). Gain represents the relative contribution of the corresponding feature to the model based on each tree in the training data. In other words, the higher the Gain, the more critical that feature is compared to other features. For random forests, we selected the "Mean Decrease in Gini" output to evaluate the importance of features (Breiman, 2001). Mean Decrease in Gini represents how each feature contributes to the homogeneity of the nodes and leaves in the given random forest model. Hence, the higher the Mean Decrease Gini, the more critical the corresponding feature. We utilize the weights associated with each of the features to evaluate their importance in the support vector machine

(Chih-Chung Chang, 2019). These weights represent the feature's discriminative ability to distinguish between two classes: the higher the weights, the more crucial the support vector machine model's feature. Lastly, in logistic regression, we obtained the coefficients from each of the iterations and then checked the consistency of the coefficients across multiple iterations.

2.5 Evaluation of phylogeny-aware distances

Phylogeny-aware distances are used to determine if we can separate species between different communities in an aggregate fashion. In our analyses, we examined the distances using multiple types of distances, including Euclidean (Schloss, 2008), Jaccard (Hancock et al., 2004), Bray-Curtis (J. R. Bray, 1957), UniFrac (C. Lozupone & Knight, 2005), and weighted UniFrac (C. A. Lozupone et al., 2007). Euclidean distance is a traditional distance measure between two species. The Jaccard index is a similarity coefficient using the presence and absence of the features within the OTU/ASV matrices. The Bray-Curtis distance is a widely used technique to highlight the differences in abundance by transforming the count matrix to a distance matrix. UniFrac, in contrast, utilizes the phylogenetic tree structure and its distances to calculate the overall distance matrix. Weight UniFrac takes account of the relative abundance of information and weights the branches of the phylogenetic tree.

3 Results

3.1 Prediction Accuracy

3.1.1 OTUs/ASVs Assignment Methods

For the traditionally-used OTUs and ASVs count matrices (**Figure 2A**), the prediction accuracy was lower at higher taxonomic levels, such as Phylum and Class, and gradually increased for most machine learning methods until reaching the OTU/ASV level of refinement. The highest average prediction accuracies are at the Genus and OTU/ASV levels. This observation provides

support for the common use of this level of taxonomy in phenotype prediction. All machine learning algorithms with an average around or below 0.5 were dropped in the figure because those algorithms do not assist in distinguishing cases and controls. This step excludes support vector machines, K-means, and hierarchical clustering.

The noticeable drop in prediction accuracy with Species-level information is due to incomplete information in the taxonomic assignment of the reference database. As a result, these missing assignments were dropped before running the machine learning algorithms, resulting in decreased performance. The number of unique feature counts for each of the taxonomic level are listed in Supplementary Material **Table S2**. Overall, the number of unique features for the Species-level was ~half that of the Genus level in the unfiltered category. After filtering, the number of unique features is close to the Order-level or Family-level information, explaining the drop we observed in **Figure 2A**.

We also extracted the top-performing combination and its associated ROC curve in **Figure 2B**; the tree-based methods, random forests, and xgBoost, performed the best, followed by neural networks, elastic net, ridge regression, LASSO regression, logistic regression, and KNN. The AUROCs for all of these methods are above 0.8.

To further investigate different machine learning algorithms' performance, we investigated a single machine learning algorithm's performance for each disease type at a single taxonomic level. In **Figures 3A, 3B, 3C,** and **3D**, we observed in density plots for ROC curves the consistent xgBoost performance at the Genus and OTU/ASV levels for both diseases. Each of the plots reflects sixteen different combinations from four OTU/ASV assignment methods, two filtering, and two normalization methods. XgBoost is consistent in its performance under different combinations. In contrast, we also included two inconsistent results. In **Figure 3E**, we

showed the logistic regression for the inflammatory bowel disease (IBD) dataset at the OTU/ASV level; we observed some excellent performing combinations and a cluster of mediocre ROC curves. Another example came from the Phylum-level support vector machine from the TwinsUK dataset shown in **Figure 3F**. This density plot contains two of the best performing combinations in our entire set of 5,412 combinations. The combinations used DADA2 both with filtered features; the non-normalized and normalized AUCs are 0.8977 and 0.8965, respectively. However, we also observed the other combinations from different OTU/ASV assignment methods perform less well.

3.1.2 K-mer-Based Methods

We examined the prediction accuracy of different k-mers, and no clear trend was observed; the prediction accuracy is relatively consistent across all lengths (not shown); thus, we display only the top-performing method (**Figure 2C**), which is the xgBoost combination using 7-mers for predicting Crohn's Disease. This combination has an AUROC of 0.924. The breakdown of AUROC per disease type at different k-mer lengths can be observed in Supplement **Figures S4, S5, S6, and S7** for Crohn's Disease, Interstitial Cystitis, Obesity, and Ulcerative Colitis, respectively.

3.2 Filtered Vs. Unfiltered

To investigate how the filtering affects the final features selected from different OTU/ASV assignment methods at different taxonomic levels and to compare these methods, we utilized upset plots to show the unique taxa shared among filtered and non-filtered methods. Regardless of filtering, the filtered and unfiltered four OTU/ASV assignment methods provide very similar unique features at the Phylum, Class, Order, and Family levels. As expected, there are more unique features identified for the Genus and Species level. In **Figures 4A** and **4B**, we provided

examples from the Genus and Species-level for the Inflammatory Bowel Disease dataset (with expanded plots for other diseases in Supplementary **Figures S2 and S3**). While most of them are shared, each of the OTU/ASV identified different sets of unique features, which might hold keys to better prediction and are important for future investigations. The number of features found per taxonomic level for each of the sub-disease categories is included in **Table S1**.

Filtering, in general, did not cause a severe difference in terms of AUROC for most of the machine learning methods. The exception is Logistic regression and K-means. Filtering improves the prediction accuracy in Logistic Regression in Family, Genus, and Species levels for both normalization categories (**Figure S8**). However, the results are not consistent, and thus, filtering needs to be judged case-by-case. **Table S2** provides the AUROC and AUPR for all prediction combinations and diseases used in this study. The AUROC is more inconsistent at more precise taxonomic levels due to the removal of features as we refine the taxonomic assignments.

3.3 Consistency

Consistency is a key feature when investigating different prediction methods, as we have shown that some machine learning methods might be sensitive to a particular OTU/ASV assignment method. The detailed breakdown for each disease type is included in Supplementary **Figures S4**, **S5**, **S6**, and **S7** for Crohn's Disease, Interstitial Cystitis, Obesity, and Ulcerative Colitis, respectively.

We also investigated the change in prediction accuracy in terms of AUROC by filtering and normalization individually. When we compared the difference between the filtered Vs. unfiltered (**Figure S8**) and normalized vs. un-normalized (**Figure S9**) across the disease categories. Overall, the difference in terms of AUROC are fairly small for most of the machine learning

methods with the exception of K-means and Logistic Regression. Filtering seems to cause more instability in the AUROC as the difference are more obvious. Overall, the decision to use normalization and filtering should be evaluated by the data at hand and the purpose of the study.

3.4 Shiny Application

Considering the vast number of combinations we have tested, we developed help a shiny application to visualize and understand the different types of combinations we have generated. Please visit the following link:

https://github.com/zhouLabNCSU/MicrobiomePredictionExplorer.

3.5 Compare Machine Learning Methods to Discovery Studies

3.5.1 Difference Among OTUs/ASVs Assignment Methods

The feature selection outputs from the four different machine learning methods are consistent within the filtered and unfiltered combination for all OTU/ASV assignment methods. The top-ranked features from both random forest and xgBoost were mostly features that had passed our filtering protocol. The support vector machine approach had a less consistent output; the rankings were similar only within the filtered and unfiltered categories. The ranks from the support vector machine were also quite different compared to xgBoost and random forest. The consistency of the coefficients is also a crucial tool for understanding the properties of a good predictor for logistic regression.

To better understand the feature output, we ranked the output, and the findings are shown in Supplementary Material **Table S3**. Based on the preliminary findings, there is no noticeable difference between the normalized and un-normalized combinations under the same filtering and OTU/ASV assignment methods. Thus, the ranks we shared in Supplementary Material **Table S3** contain only the unnormalized dataset.

3.5.2 Features Selection

The three order-level taxa all displayed average or below-average rankings. XgBoost excluded all of these taxa as they did not help with prediction. For the eleven family-level features, the five taxa that passed the filtering procedure, *Fusobacteriaceae*, *Micrococcaceae*, *Verrucomicrobiaceae, Pseudomonadaceae, and Streptococcaceae* were all ranked around the average, with none of them performing very well. Random forests, xgBoost, and support vector machines shared similar results. Lastly, among the eight genus-level taxa, *Bacteroides* (*Pectinophilus* group) and *Roseburia* ranked among the top 10 for the random forests, xgBoost, and support vector machines with consistent results in logistic regression. The exception is *Roseburia* in support vector machines, which ranked much higher.

Overall, the rankings for random forests and xgBoost were similar between the filtered and unfiltered combinations across all four OTU/ASV assignment methods. In other words, the excess taxa unique to the unfiltered dataset did not improve the prediction accuracy, as the ranks did not change much even after adding a large number of features to the model. However, in the support vector machine, the taxa ranks were inconsistent between the filtered and unfiltered OTU/ASV assignment methods. The ranks remain roughly around the same percentile.

3.6 Phylogenetic Analyses

3.6.1 Differential Taxa in Crohn's Disease

Overall, the weighted UniFrac was the best performing way to separate the Crohn's Disease and Control subjects.

We investigated how different OTU/ASV assignment methods react to the combination of a variety of ordination and the distance measure. The best performing OTU/ASV assignment method was DADA2, with the first and second axis separating 70.982% and 8.786%, which

means the combination of the first two axes explained roughly 80% of the total variance between Crohn's Disease and Control subjects (**Figure 5B**). While the other methods perform relatively well, DADA2 worked much better on distinguishing Crohn's Disease subjects with control with weighted UniFrac (**Figures 5A, 5C, and 5D**).

Moreover, we followed through with the statistical tests to determine if the first two axes were significantly affected by the disease category between Crohn's Disease and Control. A previous study determined the usefulness of using the two axes from the Multidimensional scaling techniques to discriminate between the case and control cases and with consistencies across different OTU assignment percentage matches, i.e., 99%, 95%, 90%, and 85% (Frank et al., 2007). We followed a similar protocol and examined using different combinations of distance methods and OTU/ASV assignment methods; our results did not replicate the significant separation between Crohn's disease and control. However, we observe an adjusted R^2 of 0.864 with a p-value of 0.081 using the Jaccard distance and de novo OTU assignment methods. The second-best test, PERMANOVA, uses weighted UniFrac on Closed-Reference OTU assignment methods with an adjusted R^2 of 0.6525 and a p-value of 0.073. The full table is in Supplementary Material **Table S4**.

3.6.2 Relationship between the phylogenetic trees among clustering methods

As we discussed earlier, the number of unique features reported by different OTU/ASV methods are different, so the resultant phylogenetic trees also differ. Here, we focus our investigation on the Family-level taxa, and we extracted the taxa from the eleven important taxa that previous studies had identified. We calculated the log-transformed average of Crohn's disease and Control OTU/ASV counts per the taxon assignment. Examining the taxonomic tree closely, we detected some unique taxon assignments from Crohn's disease group or the control group. There are

different observable patterns between the case and control (**Figure 6**), including the log-scale differences and the present/absent differences.

3.7 K-mers

Finally, we examined two separate types of k-mers, the very short-chain k-mers (4, 5, 6, and 7-mers) and the short-chain k-mers (15, 21, 30-mers). Both very short-chain k-mers and short-chain k-mers, when combined with effective machine learning methods, perform as well as the top-ranked OTU/ASV clustering methods for host trait prediction. From the computation side, very short-chain k-mers can be calculated quickly by parsing the raw FASTQ files, but short-chain k-mers take longer to extract, and due to the enormous number of possible combinations, we filtered the count matrices to make the final table computationally feasible. Here, any unique-k-mers with fewer than 5 reads were removed. The advantage of short-chain k-mers is the potential of mapping back to genomic data to better understand the underlying biology (Koslicki & Falush, 2016). With the short-chain k-mers, we could study them by mapping them back to a 16S rRNA database and extracting their taxonomic information. Using these mappings will be an interesting area to explore for future projects. Very-short k-mers cannot be mapped uniquely back to a reference database, as they are ubiquitous in all samples.

3.8 Benchmarking

While computational cost is not the primary goal of this journal, we nevertheless conducted benchmarking investigation by using our best phenotype, Crohn's Disease. We evaluated the difference in terms of time consumption by running the 100 iterations of 5-fold validation for the eleven machine learning methods we tested on a single core. Overall, the results suggested Elastic Net and xgBoost are the most time-consuming (**Figure S10**). Also, normalization did not

cause significant computation changes, and filtered combinations generally cause slightly shorter computation time (**Figure S11**).

3.9 Variable Importance

For each of the machine learning methods, we calculated the mean and standard deviation across all 500 rounds to evaluate the importance of features, as defined in 2.4.1. These outputs are included in the Supplementary Material **Tables S5-S16** for Order, Family, and Genus level feature outputs. For each of the combinations, the mean and standard deviation for the features from the machine learning models are shown.

4 Discussion

This article aims to explore and compare the different upstreaming processes and how they can affect downstream machine learning predictions. Despite the introduction of a large number of data pre-processing steps and machine learning methods, there has been little systematic exploration of the massive number of possible combinations of these approaches. While many of our findings accord with earlier smaller explorations, the definitive nature of our combination "search-space" provides important assurance that the community is generally applying best-practice methods for host-trait prediction. All of the completed combinations can be explored in the Shiny application in terms of their corresponding AUROC curve.

Firstly, we reviewed the impact of filtering and normalization on four OTU/ASV assignment methods. Normalization has only a modest impact on the downstream machine learning algorithm performance, while filtering has a more impact on the overall performance of the algorithms. We also observed that the filtering criteria might throw out important taxa that had been identified as important from discovery studies. Depending on the goal of the machine learning methods, filtering criteria might need to be adjusted.

We also explored the usefulness of short-chain and very short-chain k-mers and their ability to differentiate between diseases and controls. Both types of k-mers can provide high-quality predictions that are equally as good as Genus and OTU/ASV assignment methods. This area needs further research to uncover the additional potential of using k-mers as predictors.

While we tried many combinations of different processing steps, it is impossible to consider all scenarios, and there are limitations to our conclusions and in the available data. Both of the datasets we used are based on 16S rRNA from the V4 hypervariable region. Previous studies have shown that other hypervariable regions, or a combination of variable regions, affect biodiversity and community state types, which could eventually cause differences in prediction accuracy (Bukin et al., 2019; Graspeuntner et al., 2018). Moreover, the choice of the reference database may also affect the quality of the OTU/ASV assignment results, and it is recommended to use a curated database. Lastly, we employed only a single combination of filtering criteria, and different studies might require more exclusive or inclusive filtering standards, depending on the disease of interest. The current filtering criteria focus on removing rare taxonomic features.

Overall, we provided a comprehensive comparison of commonly used machine learning algorithms and how upstream methods affect overall outcomes.

References

Asgari, E., Garakani, K., McHardy, A. C., & Mofrad, M. R. K. (2018). MicroPheno: predicting environments and host phenotypes from 16S rRNA gene sequencing using a k-mer based representation of shallow sub-samples. *Bioinformatics (Oxford, England)*, *34*(13), i32–i42. https://doi.org/10.1093/bioinformatics/bty296

Asgari, E., Münch, P. C., Lesker, T. R., McHardy, A. C., & Mofrad, M. R. K. (2019). DiTaxa: nucleotide-pair encoding of 16S rRNA for host phenotype and biomarker detection. *Bioinformatics*, *35*(14), 2498–2500. https://doi.org/10.1093/bioinformatics/bty954

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., … Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, *37*(8), 852–857. https://doi.org/10.1038/s41587-019-0209-9

Breiman, L. (2001). *Random Forests* (Vol. 45).

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

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

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 785–794. https://doi.org/10.1145/2939672.2939785

Chih-Chung Chang, C.-J. L. (2019). *LIBSVM -- A Library for Support Vector Machines*. https://www.csie.ntu.edu.tw/~cjlin/libsvm/

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

Frank, D. N., St. Amand, A. L., Feldman, R. A., Boedeker, E. C., Harpaz, N., & Pace, N. R. (2007). Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(34), 13780–13785. https://doi.org/10.1073/pnas.0706625104

Gevers, D., Kugathasan, S., Denson, L. A., Vázquez-Baeza, Y., Van Treuren, W., Ren, B., Schwager, E., Knights, D., Song, S. J., Yassour, M., Morgan, X. C., Kostic, A. D., Luo, C., González, A., McDonald, D., Haberman, Y., Walters, T., Baker, S., Rosh, J., … Xavier, R. J.

(2014). The Treatment-Naive Microbiome in New-Onset Crohn's Disease. *Cell Host & Microbe*, *15*(3), 382–392. https://doi.org/10.1016/j.chom.2014.02.005

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

Goodrich, J. K., Waters, J. L., Poole, A. C., Sutter, J. L., Koren, O., Blekhman, R., Beaumont, M., Treuren, W. Van, Knight, R., Bell, J. T., Spector, T. D., Clark, A. G., & Ley, R. E. (2014). Human genetics shape the gut microbiome. *Cell*, *159*(4), 789. https://doi.org/10.1016/J.CELL.2014.09.053

Graspeuntner, S., Loeper, N., Künzel, S., Baines, J. F., & Rupp, J. (2018). Selection of validated hypervariable regions is crucial in 16S-based microbiota studies of the female genital tract. *Scientific Reports*, *8*(1), 9678. https://doi.org/10.1038/s41598-018-27757-8

Gurung, M., Li, Z., You, H., Rodrigues, R., Jump, D. B., Morgun, A., & Shulzhenko, N. (2020). Role of gut microbiota in type 2 diabetes pathophysiology. In *EBioMedicine* (Vol. 51, p. 102590). Elsevier B.V. https://doi.org/10.1016/j.ebiom.2019.11.051

Hancock, J. M., Zvelebil, M. J., & Hancock, J. M. (2004). Jaccard Distance (Jaccard Index, Jaccard Similarity Coefficient). In *Dictionary of Bioinformatics and Computational Biology*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780471650126.dob0956

J. R. Bray, J. T. C. (1957). An ordination of the upland forest communities of Southern Wisconsin. *Ecol. Monogr.*, *27*, 325–349.

Kim, S., Thapa, I., Zhang, L., & Ali, H. (2019). A novel graph theoretical approach for modeling microbiomes and inferring microbial ecological relationships. *BMC Genomics*, *20*(Suppl 11). https://doi.org/10.1186/s12864-019-6288-7

Lozupone, C. A., Hamady, M., Kelley, S. T., & Knight, R. (2007). Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. In *Applied and Environmental Microbiology* (Vol. 73, Issue 5, pp. 1576–1585). https://doi.org/10.1128/AEM.01996-06

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

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

Molik, D. C., Pfrender, M. E., & Emrich, S. J. (2020). Uncovering Effects from the Structure of Metabarcode Sequences for Metagenetic and Microbiome Analysis. *Methods and Protocols*, *3*(1), 22. https://doi.org/10.3390/mps3010022

Rideout, J. R., He, Y., Navas-Molina, J. A., Walters, W. A., Ursell, L. K., Gibbons, S. M., Chase, J., McDonald, D., Gonzalez, A., Robbins-Pianka, A., Clemente, J. C., Gilbert, J. A., Huse, S. M., Zhou, H. W., Knight, R., & Gregory Caporaso, J. (2014). Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. *PeerJ*, *2014*(1), e545. https://doi.org/10.7717/peerj.545

Rosen, M. J., Callahan, B. J., Fisher, D. S., & Holmes, S. P. (2012). Denoising PCR-amplified metagenome data. *BMC Bioinformatics*, *13*(1), 283. https://doi.org/10.1186/1471-2105-13-283

Schloss, P. D. (2008). Evaluating different approaches that test whether microbial communities have the same structure. *ISME Journal*, *2*(3), 265–275. https://doi.org/10.1038/ismej.2008.5

Weinstock, G. M. (2012). Genomic approaches to studying the human microbiota. In *Nature* (Vol. 489, Issue 7415, pp. 250–256). NIH Public Access. https://doi.org/10.1038/nature11553

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E. R., & Knight, R. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, *5*(1), 27. https://doi.org/10.1186/s40168-017-0237-y

Zheng, Y., Fang, Z., Xue, Y., Zhang, J., Zhu, J., Gao, R., Yao, S., Ye, Y., Wang, S., Lin, C., Chen, S., Huang, H., Hu, L., Jiang, G. N., Qin, H., Zhang, P., Chen, J., & Ji, H. (2020). Specific gut microbiome signature predicts the early-stage lung cancer. *Gut Microbes*, *11*(4), 1030–1042. https://doi.org/10.1080/19490976.2020.1737487

Zhou, Y.-H., & Gallins, P. (2019). A Review and Tutorial of Machine Learning Methods for Microbiome Host Trait Prediction. *Frontiers in Genetics*, *10*(JUN), 579. https://doi.org/10.3389/fgene.2019.00579

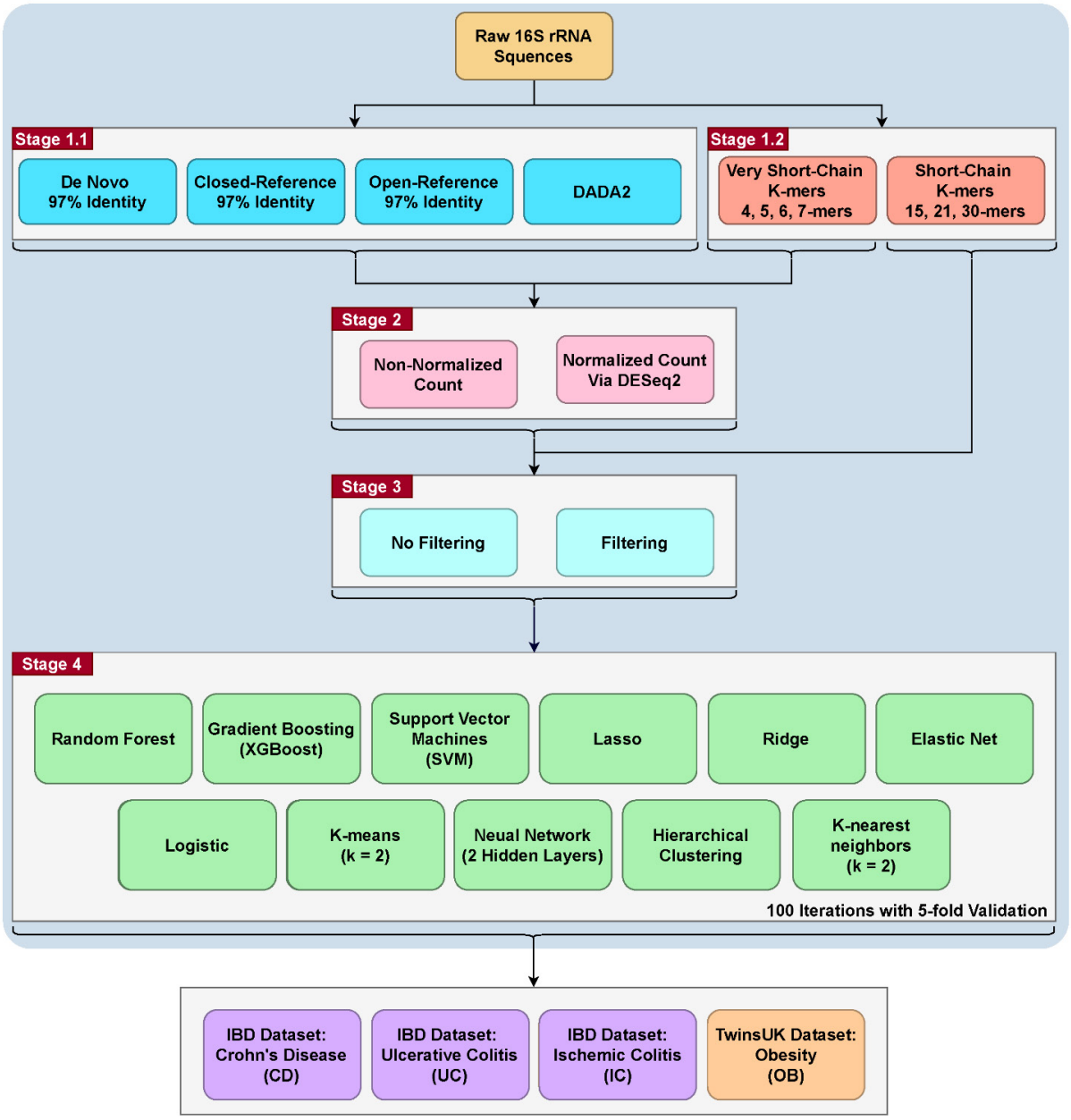


Figure 1. The workflow of the project. The project is roughly split to four stages. The first stage is the generation of count matrices via QIIME2 for the OTU/ASV assignment methods, while the k-mer matrices were generated using R (resulting in 35 count matrices). In the second stage, DESeq2 normalization are performed for all stage 1 count matrices except for the short-chain k-mers (resulting in 67 count matrices). In the third stage, filtering was performed for all the above count matrices (resulting in 123 count matrices). In the fourth stage, we ran eleven commonly used machine learning methods on the 123 count matrices with 100 iterations of 5-fold validation (resulting in 1,353 combinations). Lastly, we tested these combinations with 4 binary comparisons: Crohn's Disease, Ulcerative Colitis, Ischemic Colitis and Obesity with the corresponding control in their respective dataset (resulting in 5,412 combinations). A more detailed workflow is in Figure **S1**.

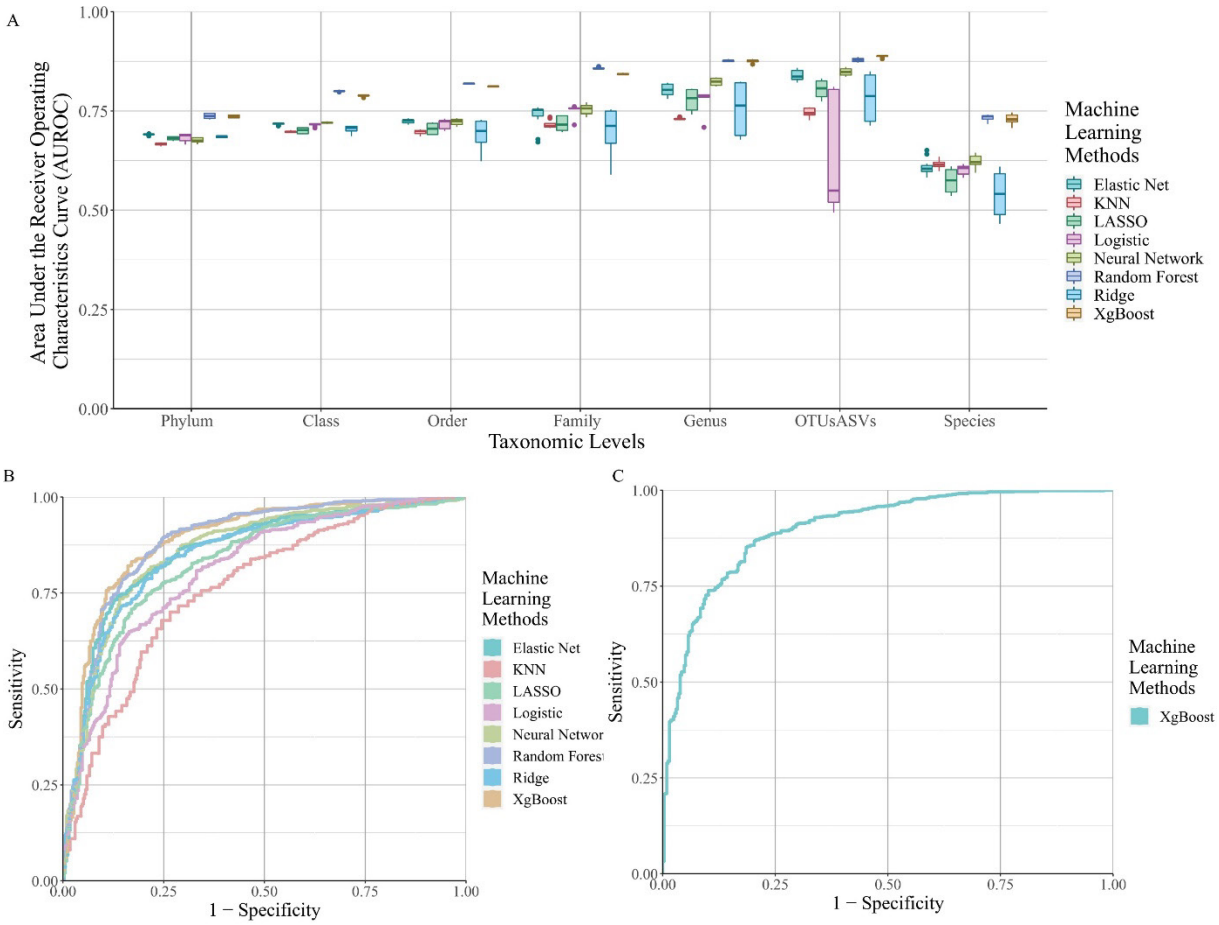


Figure 2. The area under the ROC curve (AUROC) for selected machine learning methods across different taxonomic levels and k-mer lengths. A. Boxplots of the AUROC for eight machine learning methods from OTU/ASV assignment methods across all seven taxonomic levels for Crohn's Disease. **B**. ROC curves for the eight machine learning methods from OTU/ASV assignment methods across all seven taxonomic levels. Hierarchical clustering, K-means and Support Vector Machine were removed from the figure due to their poor performance. **C**. ROC curve for the best k-mer methods to predict Crohn's Disease, which is from the xgBoost algorithms on the 7-mer level.

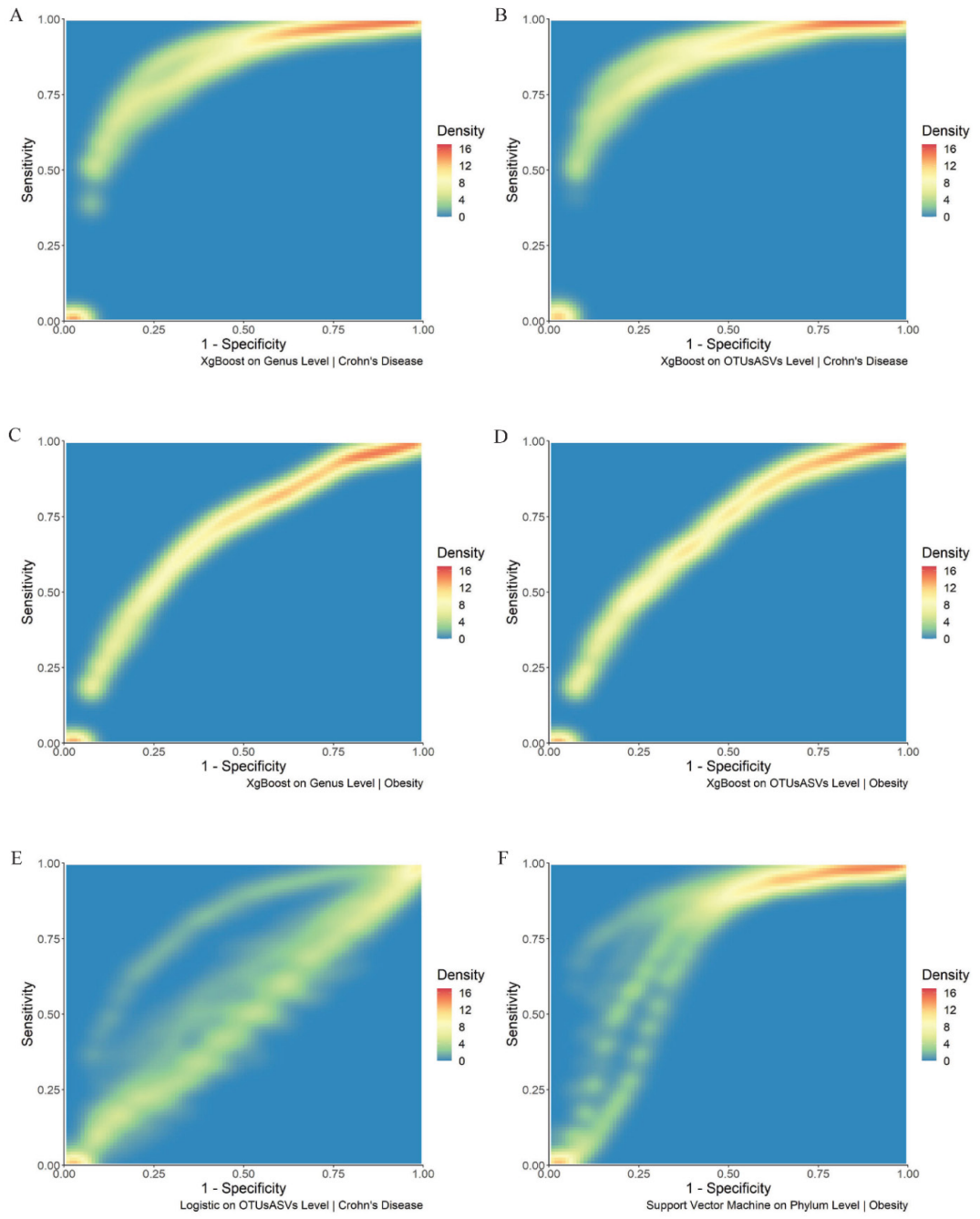


Figure 3. Density plots of the selected combination of machine learning methods, taxonomic levels, and dataset. **A.** Density plot of the ROC curve for xgBoost at the Genus level for the IBD dataset. **B.** Density plot of the ROC curve for xgBoost at the OTU/ASV level for the IBD dataset. **C.** Density plot of the ROC curve for xgBoost at Genus level for the TwinsUK dataset. **D** Density plot of the ROC curve for xgBoost at the OTU/ASV level for the TwinsUK dataset. **E.** Density plot of the ROC curve for logistic regression at the OTU/ASV level for the IBD dataset. **F.** Density plot of the ROC curve for support vector machines at the Phylum level for the TwinsUK dataset.

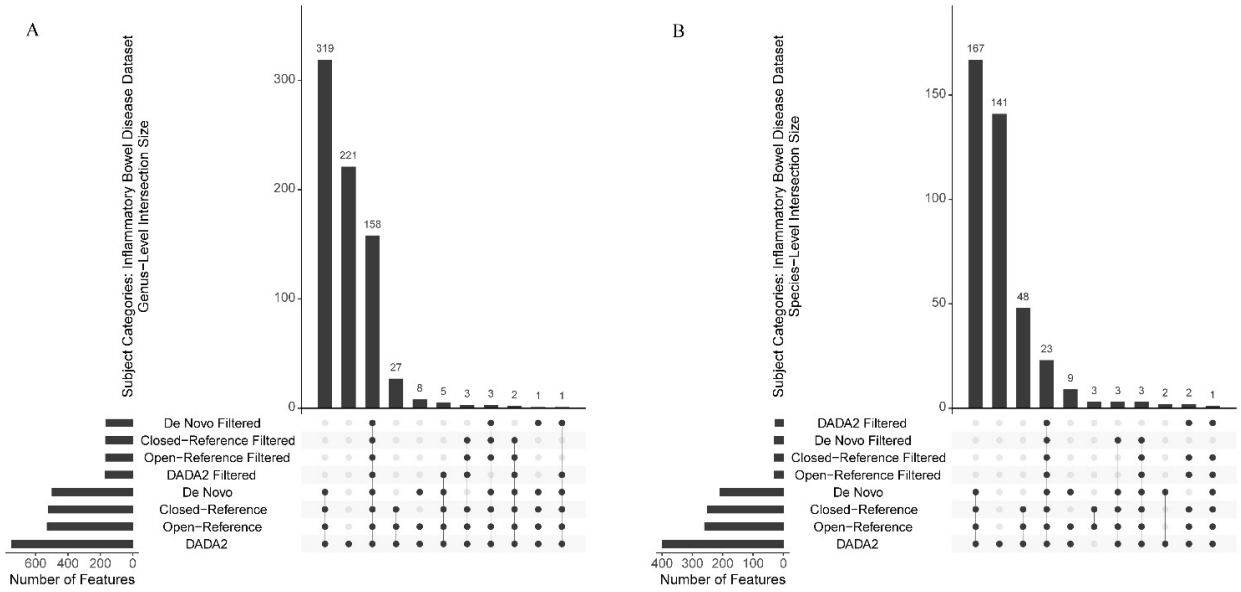


Figure 4. Upset plots of the Genus and Species-level interaction of features among the filtered and non-filtered OTU/ASV picking methods from the Inflammatory Bowel Disease Dataset. **A**. Genus-level. **B**. Species-level.

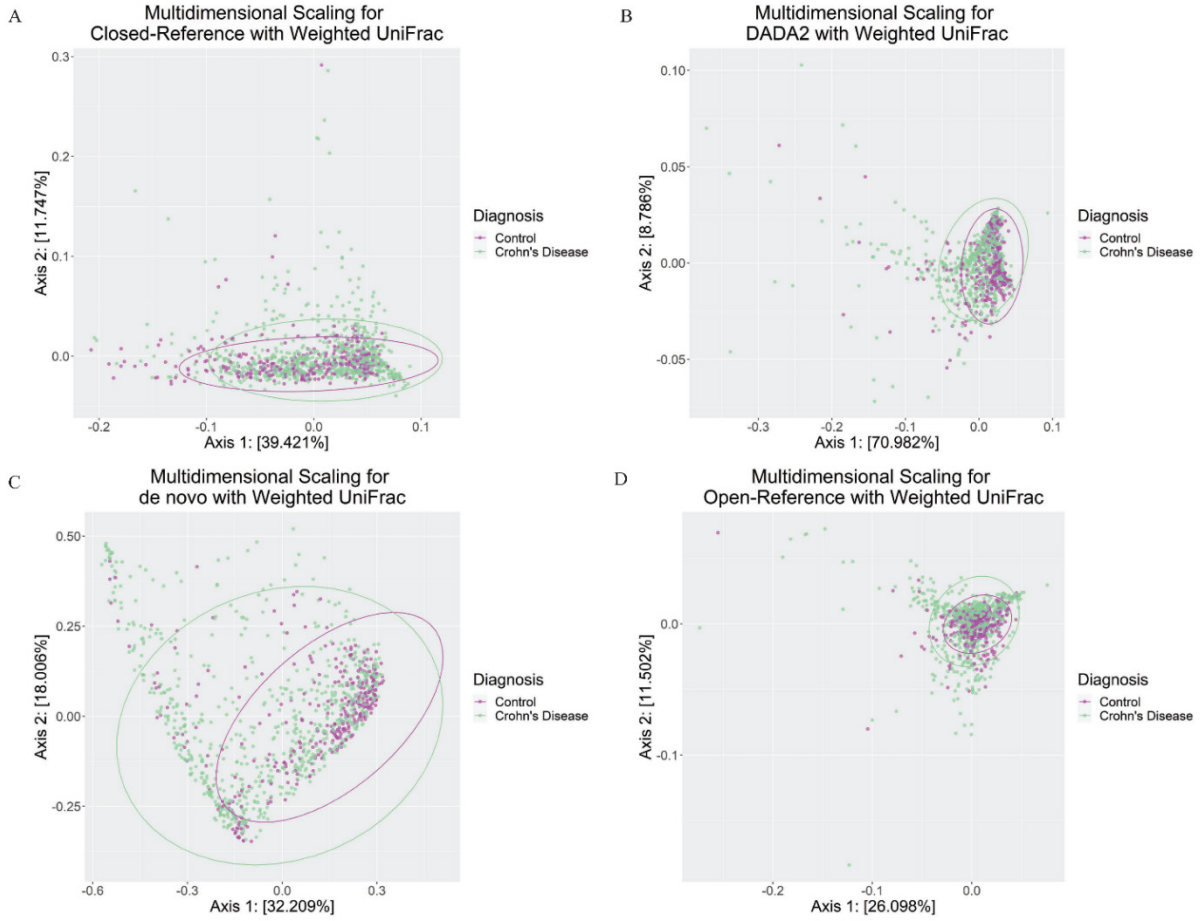


Figure 5. Weighted UniFrac ordination plot from four OTU/ASV assignment methods. **A.** Closed-Reference, an OTU assignment method. **B.** DADA2, an ASV assignment method. **C.** *de novo*, an OTU assignment method. **D.** Open-Reference, an OTU assignment method. The ellipses were drawn based on the multivariate t-distribution respectively for cases/controls.

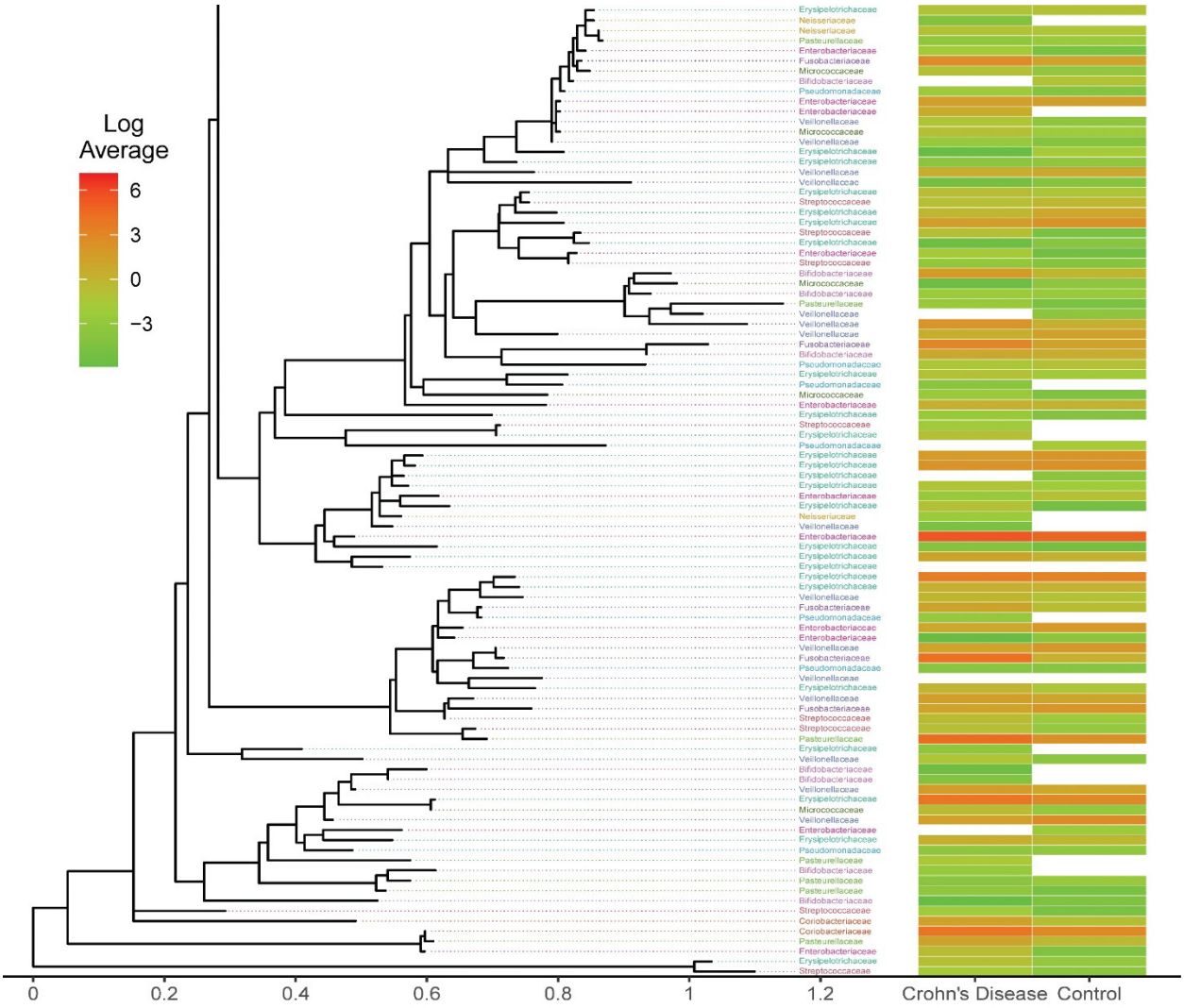


Figure 6. Phylogenetic tree from the open-reference clustering methods showing the mean log-transformed average count between the Crohn's Disease and Control. This is a subset of the full plot due to the large dimension of the original files. The full figure for all four OTU/ASV assignment methods are included in Supplement Files S9.

Table 1. Brief Summary of Datasets

	Study	**Inflammatory Bowel Diseases**				**Twins UK**	
	Disease Type	Crohn's Disease	Ischemic Colitis	Ulcerative Colitis	Control	Obesity	Healthy
	n	731	73	217	335	193	451
Sex	**Female**	337	38	96	161	192	447
	Male	394	35	123	174	1	4
Bases per FASTQ File	**Mean**	6,381,116	6,414,195	6,884,041	7,247,420	19,776,834	20,355,603
	SD	7,184,600	5,026,312	5,857,603	8,069,537	5,421,961	6,004,283
Sequence Length	**Mean**	172.44	172.52	172.52	172.99	250.84	250.84
	SD	1.37	1.18	1.36	1.12	0.40	0.41
Age	**Mean**	19.92	18.15	26.93	13.78	60.49	59.84
	SD	14.47	10.77	18.31	9.78	9.56	9.57

Table 2. Presence of important taxa in our clustering methods

Gevers et al.	Glassner et al.	Kim et al.	Taxonomic Level	Open-Reference	Closed-Reference	*de novo*	DADA2
Bacteroidales			Order				Present
Clostridiales			Order	Present	Present	Present	Present
		Lactobacillales	Order				Present
Coriobacteriaceae			Family	Present	Present	Present	Present
Enterobacteriaceae			Family	Present	Present	Present	Present
Erysipelotrichaceae			Family	Present	Present	Present	Present
*Fusobacteriaceae**			Family	Present	Present	Present	Present
*Micrococcaceae**			Family	Present	Present	Present	Present
Neisseriaceae			Family	Present	Present	Present	Present
Pasteurellaceae	*Pasteurellaceae*		Family	Present	Present	Present	Present
Veillonellaceae	*Veillonellaceae*		Family	Present	Present	Present	Present
*Verrucomicrobiaceae**			Family	Present	Present	Present	Present
		*Pseudomonadaceae**	Family	Present	Present	Present	Present
		*Streptococcaceae**	Family	Present	Present	Present	Present
Gemella			Genus	Present	Present		Present
	*Bacteroides**‡		Genus	Present	Present	Present	Present
	Bifidobacterium		Genus				Present
	*Eubacterium**†		Genus	Present	Present		Present
	*Faecalibacterium**		Genus	Present	Present	Present	Present
	Fusobacterium		Genus	Present	Present	Present	Present
	*Roseburia**		Genus	Present	Present	Present	Present
	Solobacterium		Genus	Present	Present	Present	Present

***** The Taxa that were kept after filtering was performed

‡ There are multiple groups under *Bacteroides*, *Pectinophilus* group was selected as it is the only Eubacterium group that remains after our filtering procedure

† There are multiple groups under *Eubacterium* - the *nodatum* group was selected as it is the only Eubacterium group that remains after our filtering procedure

Present: the taxa occurred in the features from the corresponding non-filtered OTU/ASV assignment method

An empty cell means the taxon is absent in the corresponding OTU/ASV assignment method

Chapter 3

C3NA – correlation and consensus-based cross-taxonomy network analysis for compositional microbial data

1 Introduction

In recent years, researchers have discovered specific yet complex links between the human gut microbial ecology and diseases such as colorectal cancer (Saus et al. 2019; Zhang et al. 2020; Mo et al. 2020) and inflammatory disorders (Sultan et al. 2021; Glassner, Abraham, and Quigley 2020; Mancabelli et al. 2017). According to a widely recognized model for the microbe-human interaction, a dysbiosis of gut microbiota is associated with the development of illnesses (Degruttola et al. 2016). Different microorganisms will thrive or decline depending on the host illness progression, medicine, food, and other variables. Studies on the essential bacteria linked with a disease have shown various integral microbes that play vital roles in illness progression. These findings aided in the development of predictive models and targeted therapies. 16S ribosomal RNA amplicon sequencing is the most cost-effective and accessible way to obtain microbial composition data. It is supported by established downstream taxonomic classification pipelines/software, curated 16S ribosomal reference databases, and statistical methods designed to analyze microbial compositional data (Bolyen et al. 2019).

Two methods for detecting important microbes are differential abundance (DA) analysis and co-occurrence network analysis. DA method is a popular tool for finding microbial taxa associated with ill or healthy people, and co-occurrence networks aim to decipher the taxa-taxa co-occurrence patterns for unique disease enriched or depleted pattern discovery. The concordances of differentially abundant taxa are one of the main challenges in DA approaches. While numerous methods can be used to evaluate consensus taxa, the results are influenced by sample/study variances and filtering criteria (Nearing et al. 2022). Previous co-occurrence networks provided helpful information about taxonomic phenotypic associations (Chen et al. 2020), and finding significant taxa-taxa correlations among many highly correlated pairs requires

different correlation approaches, module identification, and post-module practice methods which lacks a concordant approach.

To address these challenges, we present C3NA, a user-friendly and open-source R-package that includes data processing and interactive visualization functionalities. The goal of the C3NA is to maximize the available biologically inferable information in terms of taxonomic assignments across Phylum, Class, Order, Family, Genus, and Species levels via co-occurrence network analysis to extract optimal numbers of co-occurring taxa modules with similar taxonomic abundance patterns.

2 Materials and Methods

2.1 Microbial Data Processing

There are many established pipelines available for processing the 16S rRNA amplicon sequencing to summarize the raw sequencing data into taxonomic profiles, such as the operational taxonomic units (OTUs) and amplicon sequencing variants (ASVs). In addition, many taxonomic assignment methods were published as well as established pipelines to streamline the process. For C3NA, we utilized the QIIME2 pipeline with two of the frequently used taxonomic assignment methods, de novo clustering and DADA2 algorithm (Bokulich et al. 2018; Callahan et al. 2016). Regardless of the methods and reference database, the resulting taxonomic profile includes an OTUs/ASVs table, a taxonomic table, and a metadata table. To ensure the accuracy of these data prior to loading into the C3NA pipeline, we utilize the Phyloseq R package to ensure the correct formatting of these tables, and the Phyloseq object is the initial input for the C3NA pipeline. Prior to running C3NA, the user need to perform a subset function to extract a single phenotype Phyloseq object. The C3NA pipeline is divided into six sections as shown in **Fig.1**.

2.1.1 Cross-Taxonomy Table

Taxonomic-specific tables from Phylum, Class, Order, Family, Genus, and Species-level are summarized by their corresponding taxonomic profile. We recommended removing samples with a library size of fewer than 1,000 reads as these samples are known to suffer from low-quality issues in terms of microbial diversity as well as sequencing-related issues (Navas-Molina et al. 2013). For each of the taxonomic levels mentioned above, we will create a taxonomic-specific names, and then the final count matrix $\boldsymbol{M}$ is the stacked-taxa result of all these six matrices. Lastly, as we are focusing on higher taxonomic levels than OTUs/ASVs by summing the OTUs/ASVs counts, we filtered out the taxa that did not present in at least 10% of the samples (Friedman and Alm 2012). This filtering criteria coincide with one of the assumptions for SparCC in which the kept taxa are assumed to be present among samples, and this approach also drastically reduces the computational complexity with the reduced number of taxa. For each of the phenotypes within the study, a phenotype-specific $\boldsymbol{M}'$ matrix is created by applying the aforementioned filtering procedures.

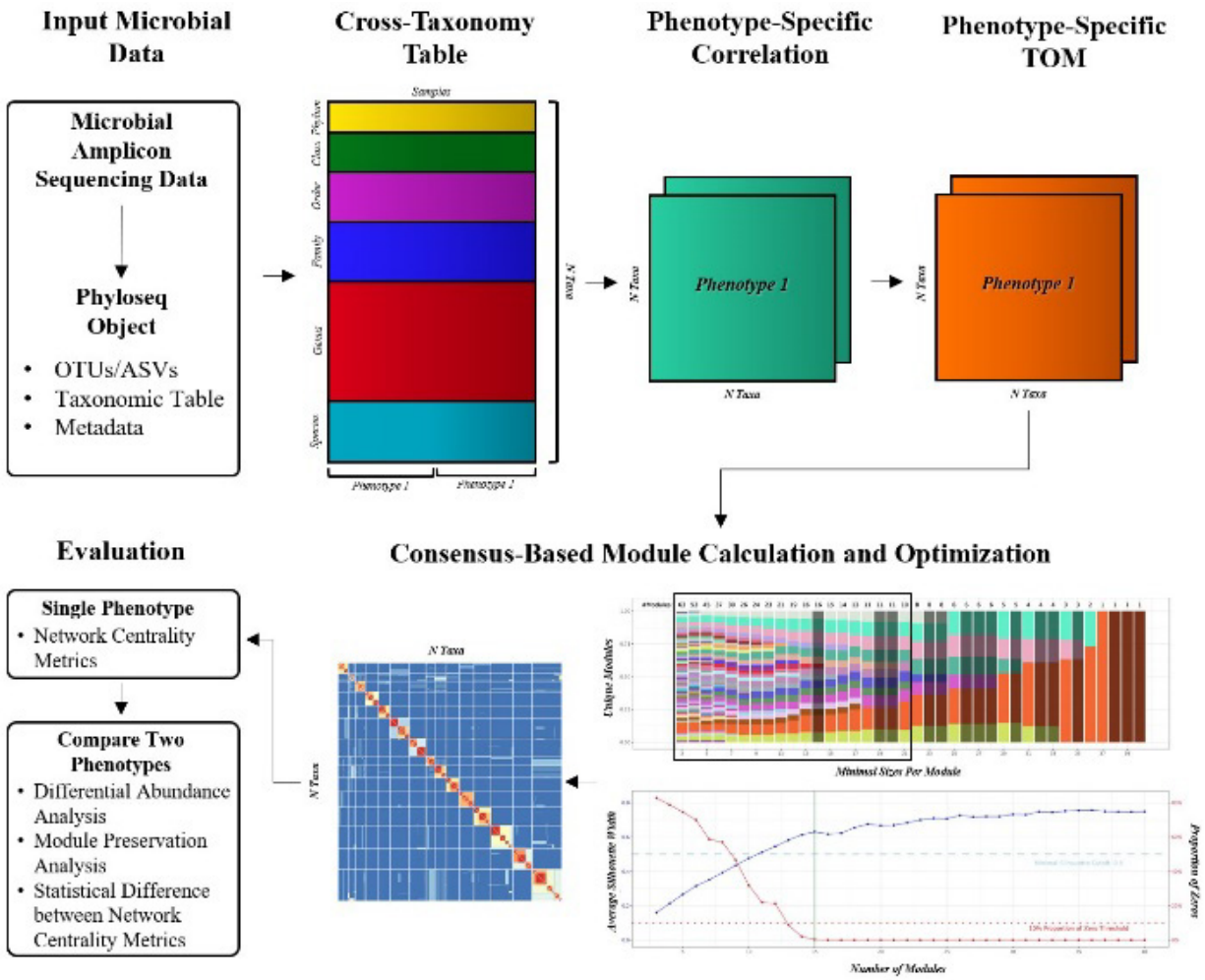


Figure 1. C3NA framework for two phenotypes comparison. For every single phenotype, the data is loaded into Phyloseq; the stack-taxa count matrix is extracted by combining Phylum, Class, Order, Family, Genus, and Species-level raw count matrix together. Then, the matrix undergoes SparCC correlation calculation with 1,000 bootstraps followed by the topological overlap matrix (TOM) calculation under the "signed" network setting. Next, the dissimilarity TOM matrix (1-TOM) is used for hierarchical clustering with a range of minimal numbers of taxa per module (3 to 40) to extract a range of clustering patterns. A selected range of patterns is used to generate a consensus matrix, in which the intra-modular connections are the key taxa-taxa correlations we focus on in the subsequent network analysis. Module preservation analyses are performed when we compare two phenotypes.

2.1.2 Phenotype-specific Correlation

We employed Sparse Correlation for Compositional data (SparCC) correlation with 1,000 bootstrap settings employed using the SPIEC-EASI R Package (Friedman and Alm 2012; Kurtz et al. 2015). The bootstrap resulted in a correlation value between each taxa pair and a p-value,

and the Benjamini-Hochberg (BH) method will subsequently be applied to adjust for multiple testing.

We further validated the stability of the SparCC on how the stacked-taxa affect the taxa-taxa correlations (detailed in Supplement Results). We investigated the impact of using the stacked-taxa correlation compared to the single-taxonomic level correlations and found minimal differences for the taxa-taxa pairs, especially for the correlations above 0.2 (Supplementary Results). There is no drastic difference between the stacked-taxa and single-taxonomic level correlations (Supplementary Results). As a result, we recommended a minimal correlation cutoff at 0.2 to remove uncertain and weak correlations.

SparCC algorithms enable parallel programming, but they are still computationally expensive in both storage requirements and time. The computation time varies depending on the number of taxa extracted as well the number of strong correlations presented at each bootstrap; our runs take between one to five days using 12 cores on Intel(R) Xeon(R) CPU E5-2670 0 @ 2.60GHz with 30 GB of RAM for four studies we have tested. Time consumption for each of the examined phenotypes is recorded in Supplementary Table 1. Also, we investigated the impact of using fewer iterations, and the results indicated smaller number of iterations would have more significant correlations, most of which are below 0.3 (Supplementary Result). For preliminary investigation, it is possible to run as little as ten iterations, and the user is advised to adjust the display on the Shiny application for the correlation to increase the correlation cutoff to 0.3.

2.1.3 Phenotype-specific Topological Overlap Matrix (TOM)

The Topological Overlap Matrix (TOM) is constructed from the correlation matrix under the signed network setting using the WGCNA R package (Langfelder and Horvath 2008). The signed network is chosen because negative correlations should not be interpreted the same as

positive correlations, as negative correlations carry different biological meanings. The result in the TOM matrix represents the network connections strength, especially for spurious connections (Yip and Horvath 2007).

2.1.4 Taxa-based Module Calculation

We obtained the dissimilarity TOM via (1 – TOM) and used the complete linkage hierarchical clustering to classify the taxa into different modules. By default, the minimal module size range is between 3 and 40, as shown in Fig. 1. There is a clear, dynamic change with numerous unique modules, and the module became more stable as the number of unique modules reduced to less than 10, where more repetitive module patterns emerged. By default, C3NA combines all the unique module patterns equal to or greater than ten modules (Fig. 1, Supplement Data). We investigated the difference between choosing different numbers of unique patterns and the optimal number of clusters and discovered that there is minimal effect as long as the user chooses the dynamic region of the patterns and a reasonable number of consensus-based clusters (Supplementary Results).

2.1.5 Consensus-Based Module determination and Optimization

The consensus-based module determination used the Cluster-based Similarity Partitioning Algorithm (CSPA) (Strehl and Joydeep 2002). We focus on the region in which the module membership changes continuously, incrementing minimal module sizes. The module membership becomes stable, defined as no modular membership changes, three sequential incrementations of the minimal size, and the rest remains broadly stable. We extract each of the taxa-taxa modular assignments and create an individual binary similarity matrix with the presence of taxa-taxa pairs in the same module as 1, otherwise 0. The consensus matrix is obtained via averaging all these similarity matrices.

Two parameters are used to determine the optimal number of clustering. Firstly, we utilize the proportion of zeros per module; this proportion should be below 10%. Secondly, we calculated the average silhouette width for each of the clusterings based on the consensus matrix, and we will choose the first local maximum, which represents the first drop in the average silhouette score. We use the corresponding modular membership to construct the intra-modular taxa modules. The module patterns, silhouette results, consensus, and correlation matrices for our examined datasets with different taxonomic assignment methods are illustrated in Supplement Data. Once the module is determined, the user can utilize our shiny application to investigate the single phenotype taxa-taxa relationship or compare two phenotypes to determine the preserved or perturbed modules.

2.1.6 Single Phenotypes Extraction and Two Phenotypes Comparison

We evaluate every phenotype's result and use module preservation analysis to assess the differential taxa in network structure alteration between the two phenotypes. We use the $Z_{Summary}$, a composite preservation statistic proposed by Langfelder et al., to evaluate the module preservation between the disease phenotype and control. $Z_{Summary}$ compared the connectivities of the intramodular nodes and the highly connected nodes between the comparison groups. The *medianRank,* which is less sensitive to module size, is also selected to assist the definition of preserved modules (Horvath 2011; Bakhtiarizadeh et al. 2018; Li et al. 2015). Ideally, the higher the $Z_{Summary}$ and the lower the *medianRank,* the more preserved the module is.

Moreover, as microbial modules can be tiny, users should distinguish a more extensive and more diverse module from a smaller taxon with very similar phylogenetic information, i.e., taxa from the same phylogenetic branch. From a biological point of view, high and low preservation modules are critical. The high preservation module contains the connections between two

phenotypes, and the standard preservation modules have modules and elements that are either no interest or phenotype-specific perturbation. However, it is essential to evaluate all modules, and modules with differentially abundant taxa are often significant.

2.2 Evaluation of the Optimal Number of Clusters

Two significant considerations determine the appropriate number of clusters. The first is the distinct patterns of modules picked from a different minimum number of taxa per module for building the consensus matrix. The silhouette width evaluation of the consensus matrix with hierarchical clustering is the second factor. We investigated how different selections of these two parameters affect the downstream analysis in terms of intra-modular correlations, and our results show that it is essential to select one less dynamic region of the module patterns, around ten modules (Fig. S8). And for the optimal cluster selection, it is vital to choose a minimum number as the curve turns into a plateau region. Once these two are selected carefully, the resulting intra-modular correlations are very similar and should not drastically affect the preservation and network analysis. A complete investigation is in Supplement Result.

2.3 Network Centrality Metrics

There are three network centrality metrics used by the C3NA, degree centrality, closeness centrality, and transitivity using the igraph R package (version 1.2.8) (Gabor Csardi and Nepusz 2006). The purpose of these parameters is to infer the significance of the taxon between the comparing phenotypes. We choose the normalized version for the degree and closeness to account for the total number of vertices in the graph, making the results more comparable. We calculate the local transitivity for the node's importance within the local network. We extract the corresponding intra-modular members from the two phenotypes for each taxon to construct a network. We subsequently calculated the three parameters with and without the selected taxon,

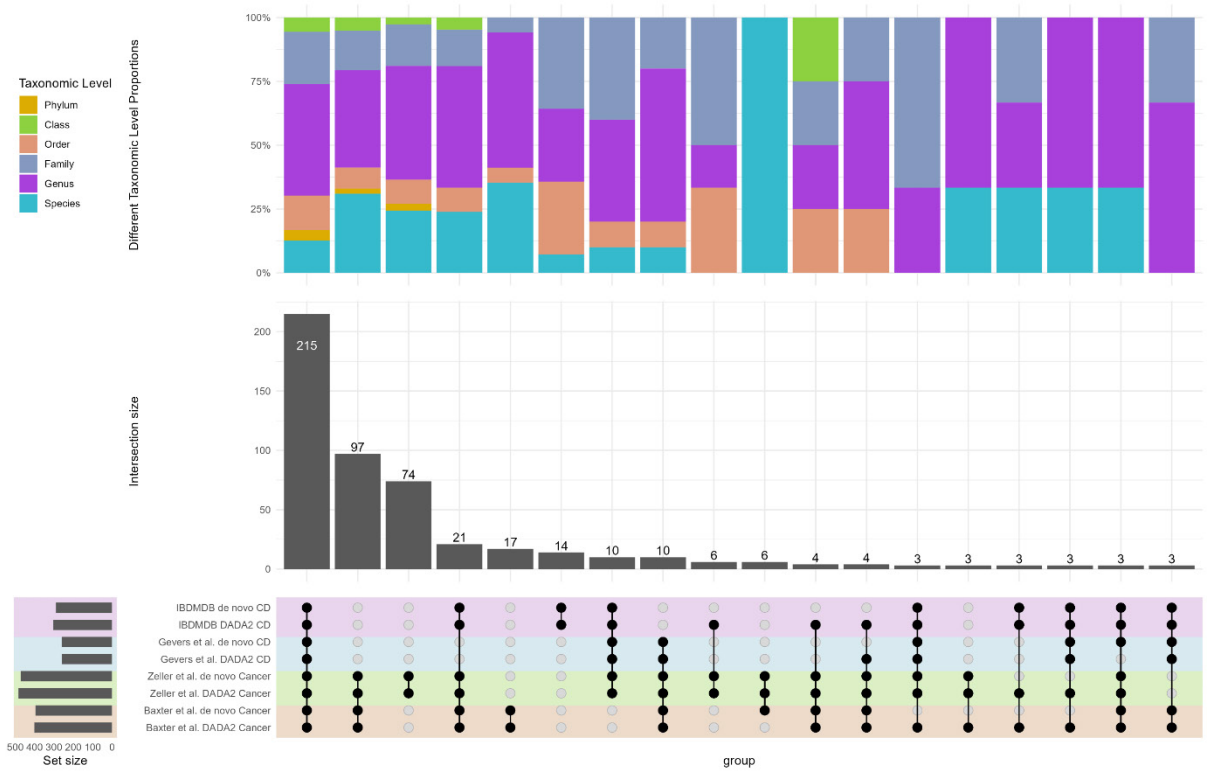


Figure 2. The shared taxa patterns among the studies and taxonomic assignment methods. The top bar plot illustrates the relative number of taxa from each of the six taxonomic levels. The bottom upset plot and the interaction plot illustrate the number of shared taxa and patterns among the examined datasets and taxonomic assignment methods.

then used the paired-sample Wilcoxon test to compare the changes among these three network metrics. We define the influential taxon as one with at least one statistically significant difference after BH-adjusted p-values ≤ 0.05.

2.4 Intra-Modular Evaluation

For each of the modules, we will keep the taxa-taxa correlation greater or equal to 0.2 with a BH-adjusted p-value no higher than 0.05. Next, we obtain the threshold by comparing correlations at the stack-taxonomic and single-taxonomic levels (Supplement Results). The combination of these two parameters will help estimate the influential taxa within each module.

Table 1. Dataset information from four microbiome study and their associated information from C3NA and differential abundance.

					C3NA		Differential Abundance		
Dataset	**Taxonomic Assignment Method**	**Phenotype**	**Number of Samples**	**Number of Modules**	**Intra-Modular Taxa-Taxa Correlation**	**Influential Taxa**	**ANCOM-BC**	**ALDEx2**	**MaSaLin2**
Baxter et al.	DADA2	Cancer	127	15	1,549	92	53	31	29
		Control	134	20	1,014				
	de novo	Cancer	127	19	1,361	76	72	23	28
		Control	134	17	1,061				
Zeller et al.	DADA2	Cancer	41	23	2,405	206	51	21	5
		Control	50	15	2,580				
	de novo	Cancer	41	21	2,306	182	48	16	5
		Control	50	22	1,953				
Gevers et al.	DADA2	Crohn's Disease	731	13	1,045	65	177	137	119
		Control	335	22	839				
	de novo	Crohn's Disease	731	16	793	36	160	132	117
		Control	335	16	786				
IBDMDB	DADA2	Crohn's Disease	86	16	1,376	156	107	35	44
		Control	46	15	1,742				
	de novo	Crohn's Disease	86	18	930	98	101	26	33
		Control	46	17	1,198				

2.5 Differential Abundance Analyses

In our analyses, we chose three validated DA methods and executed them on each of the six taxonomic levels, including ANOVA-Like Differential Gene Expression Analysis (ALDEx2) (Fernandes et al. 2013), Analysis of Compositions of Microbiomes with Bias Correction (ANCOM-BC) (Lin and Peddada 2020), and Multivariable Association Discovery (MaAsLin2) (Mallick et al. 2021). As expected, there are only a small number of taxa consistently identified by the methods from different clustering methods, and the proportion of the consensus taxa among the OTUs/ASVs clustering methods is much larger (Supplement Results). We utilized the recommended 10% prevalence filtering for each of the Phylum to Species-level assignments prior to running the methods on our dataset to obtain more robust results (Nearing et al. 2021). Each of these differential abundance analyses was performed between the disease and control samples with a binary outcome.

For the ANCOM-BC, we used the Benjamini-Hochberg (BH) adjusted p-value instead of the default Benferroni-Holm methods. The parameter will include the detection of structural zero for better suiting the unique structure of the microbiome data. The determination of the differential abundance taxa is by a BH-adjusted p-value less or equal to 0.05.

For the ALDEx2, we went with the safest approach to maximize the amount of taxa identification from the Wilcoxon output with less than 0.05. Ideally, the best taxa should overlap (95% CI of the effect size omit null point of zero) and an effect size cutoff of 1. However, both are rare with microbiome data, and the significantly abundant taxa will be defined with a BH-adjusted p-value less or equal to 0.05 from the ALDEx2 output.

For the MaAsLin2, we ran with Arcsine Square Root (AST) transformation without including any covariates. The differential abundance taxa were determined using the q-value, which is calculated using BH adjusted p-value less or equal to 0.05.

2.6 Raw 16S rRNA Data Source and Processing Procedures

We evaluated C3NA using two colorectal cancer (CRC) datasets and two inflammatory bowel diseases with Crohn's Disease (CD) 16S rRNA datasets. The first CRC dataset labeled as "cancer" was from PRJNA290926 (Baxter et al. 2016), and the second CRC dataset labeled as "cancer2" was extracted from PRJEB6070 (Zeller et al. 2014). For both CRC datasets, we only used the samples labeled as "Cancer" and "Normal." The first Crohn's disease dataset was downloaded from PRJEB13679 (Gevers et al. 2014), and we used the "CD" and "no" as the case and control, respectively. The second Crohn's disease dataset was from the IBDMDB (Lloyd-Price et al. 2019) website, and we used the "CD" and "non-IBD" as the case and control,

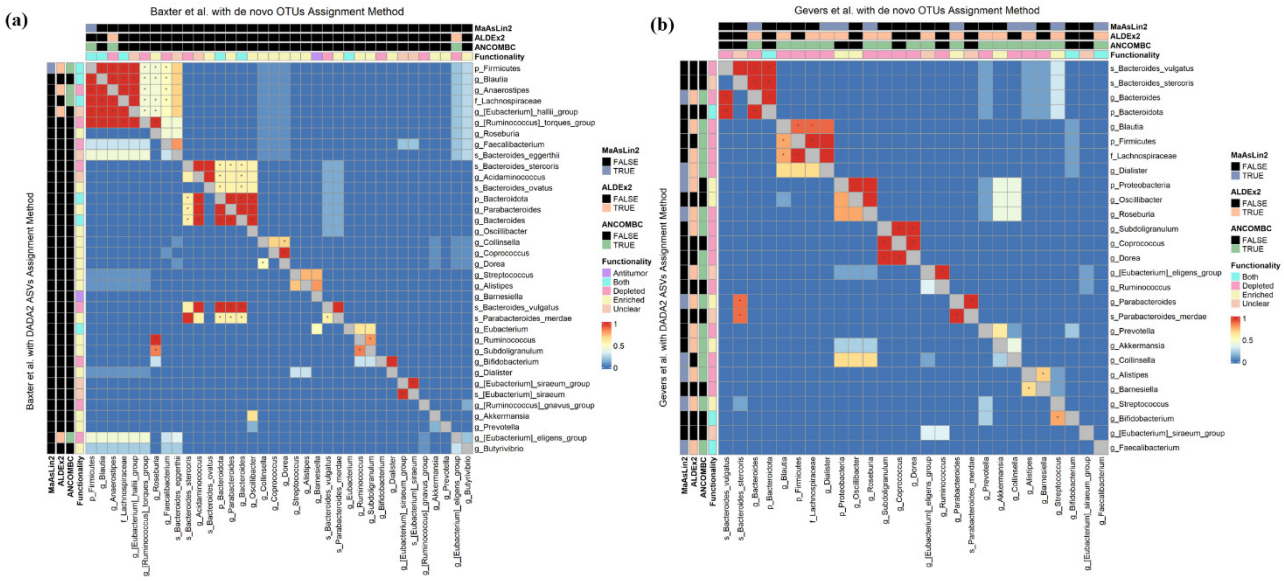


Figure 3. C3NA consensus comparison between the de novo and DADA2 taxonomic assignment methods. (**a**) *de novo* (**b**) DADA2. Asteroids highlight the intra-modular correlation, and the tile color represents the consensus score based on averaging over the selected modular patterns.

respectively. For the evaluation, we only utilize the forward reads for each of the 16S rRNA datasets. This is due to the unforeseeable format restriction on different datasets the user might use. For most cases, the reverse reads tended to have lower QC and needed to be merged with the forward reads before the taxonomic assignment procedure. The merging quality varies between studies and can negatively impact the taxonomic assignment. To ensure the quality of this method for more consistent usability, we run our taxonomic assignment for all the methods using only the forward reads. Next, we evaluated two of the most prevalent different taxonomic assignment methods, de novo clustering and DADA2 (Callahan et al. 2016). All assignments used the same SILVA 138 (Quast et al. 2013) reference under QIIME2 (version 2021.4) environment (Bolyen et al. 2019). The dataset information is shown in Table 1.

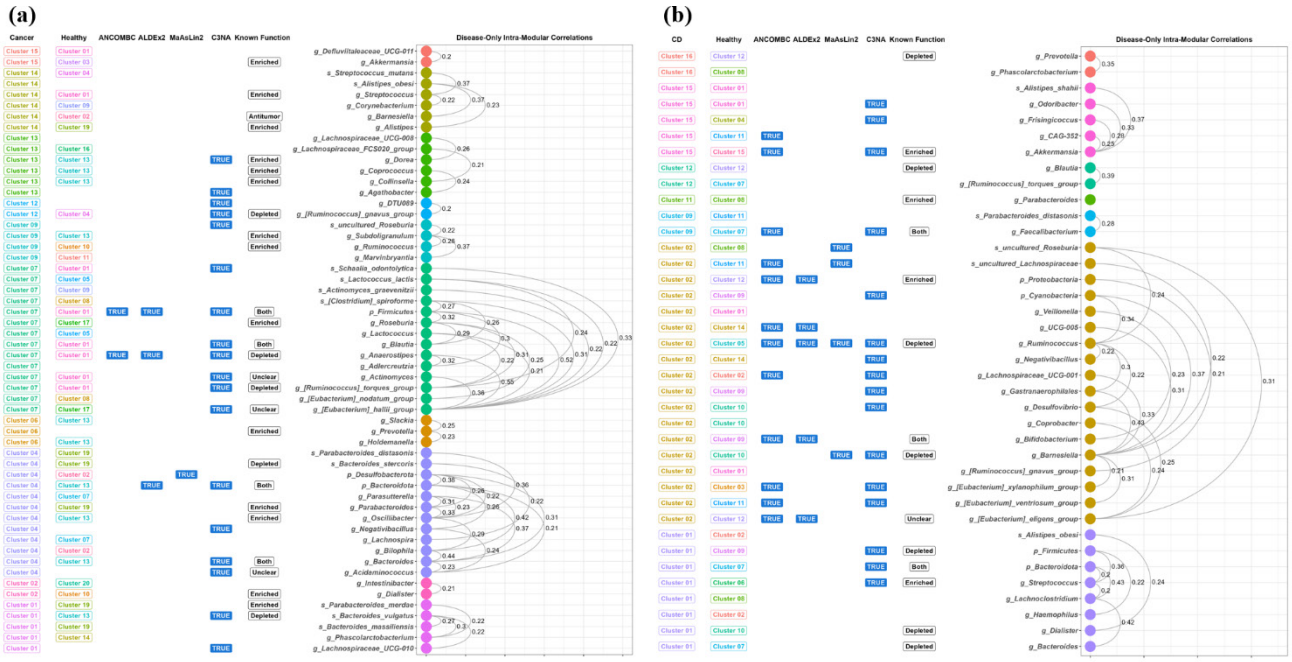


Figure 4. Functional inferences among taxa with disease-only intra-modular correlations. The Cancer/CD and Healthy columns represents the cluster in which the taxa belongs to, if present. The ANCOMBC, ALDex2, and MaAsLin2 with TRUE represent they are differentially abundant. The C3NA column shows if the taxa are influential in the network between disease and control. The known function is obtained from a selection of publications. The disease-only intra-modular correlations identify the important taxa and their corresponding correlation values on the connecting arcs.

3 Results

3.1 Differential Abundance Analyses

There is a clear pattern of inconsistency in which taxa are DA among the three DA methods evaluated, and they are summarized in Fig. S2-S5 for the four datasets we examined. In addition, ANCOM-BC obtained the most number of DA taxa compared to the other methods.

3.2 Post-filtering Taxa Comparison

We apply 10% sample prevalence filtering is used to remove rare taxa from each phenotype. The remaining taxa showed consistency and specificity toward the disease. There are more taxa left with CRC than CD, which conform with the lower bacterial diversity in CRC compared to CD (Fig. 2, Supplement Result) (Ma et al. 2021).

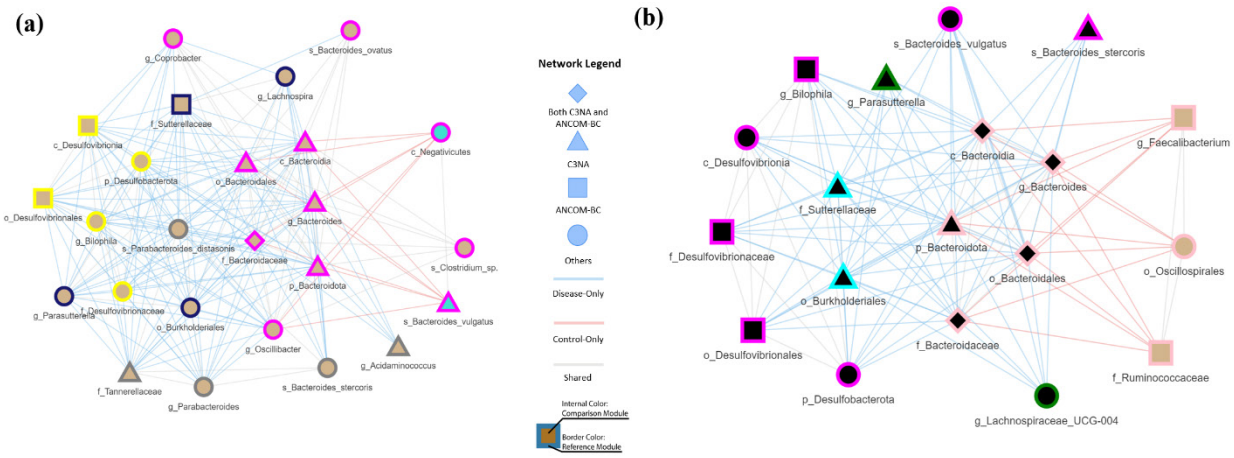


Figure 5. The networks created from taxa related to Genus *Bacteroides* from Baxter et al. and Gevers et al. between the disease (CRC or CD) and control. The taxonomic assignment method used is DADA2. These figures are generated from the build-in Shiny application.

3.3 Consistency Between the Taxonomic Assignment Methods

There are study-dependent and OTUs/ASVs assignment methods dependent on co-occurrence patterns, and C3NA will construct different networks with the important taxa-taxa connections preserved across studies that are specific to the phenotype (Supplementary Results). We selected the most prevalent taxa among our datasets and reported by other studies for each of the disease phenotypes with a combination of multiple taxonomic levels, including 36 and 27 taxa for the CRC and CD, respectively. Previously reported functionalities associated with these taxa and identified a few categories for the CRC (Ma et al. 2021; Baxter et al. 2016; Dai et al. 2018; Nistal et al. 2015) and CD (Ma et al. 2021; Matsuoka and Kanai 2015; Santoru et al. 2017; Mancabelli et al. 2017; Metwaly et al. 2020) phenotypes, and the functionality match well with the consensus clustering (Fig. 3a-b).

There is a high consistency in terms of which taxa are grouped into the modules and which taxa-taxa correlations are intra-modular. This illustrates the practicality of using the C3NA pipeline regardless of the taxonomic assignment methods (Fig. 3a-b).

3.4 Identification of Phenotype-Specific Taxa-Taxa Correlations

One of the critical findings from C3NA among these four datasets is the connection between the Genus *Bacteroides* and *Parabacteroides*, which has an intra-modular correlation found among CRC and absent among CD datasets (Fig. 5a-b). This correlation between Bacteroides and Parabacteroides is 0.641 and 0.416 for CRC and Healthy samples. On the other hand, consensus-based clustering groups them into the same module in CRC with a consensus of 1 and 0 for the control. Moreover, C3NA can help assign functions to less-studied taxa and guide biomarker discovery studies, as shown in Fig. 4a-b, where we extracted the disease-only correlation from the taxa shown in Fig. 3a-b. The results show good consistency in the Category with minor discrepancies, and this demonstrates the potential for C3NA to identify new functions to taxa for biomarker discovery. Lastly, C3NA enables an interactive comparison of taxa across multiple taxonomic levels between phenotypes. For instance, Order Burkholderiales, the higher taxonomic level for the genera *Parasutterella* (Sobhani et al. 2019) and *Sutterella* (Mori et al. 2018), both known to be enriched in CRC samples, had an exceptionally high intra-modular correlation with Bacteroides solely in Cancer samples (Fig. 5a-b).

4 Discussion

In this paper, we presented a correlation and consensus-based investigation of microbial sequencing data to extract and refine the taxa-taxa co-occurrence network for inferring biological relationships between the microbes. C3NA has a wide range of applications, including detecting specific co-occurrence patterns and identifying, confirming, and assigning functionality to microorganisms. By comparing the co-occurrence patterns that differ between different phenotypes, C3NA was able to detect unique microbial patterns that represent phenotype-specific and study-specific key taxa-taxa interactions. These interactions can be examined further regarding

their functional potentials; C3NA can assist in resolving disagreements regarding the contribution of a single microorganism to a given phenotype by examining the relationship of all its biologically inferable taxonomic categories.

The main advantage of the co-occurrence network approach with the ability to integrate a range of differential abundance analyses is to broaden our understanding of the microbial contribution towards a particular phenotype. Given the variability of the results from DA methods, C3NA enables the incorporation of as many DA as possible for concordant analyses to extract the most valuable groups of taxa that are differentially abundant or connected between phenotypes.

In conclusion, we presented a novel microbial data analysis pipeline for enhanced and methodological investigation of microbial communities and their compositional difference between phenotypes.

References

Bakhtiarizadeh, Mohammad Reza, Batool Hosseinpour, Maryam Shahhoseini, Arthur Korte, and Peyman Gifani. 2018. "Weighted Gene Co-Expression Network Analysis of Endometriosis and Identification of Functional Modules Associated with Its Main Hallmarks." *Frontiers in Genetics* 9 (OCT). https://doi.org/10.3389/FGENE.2018.00453/FULL.

Baxter, Nielson T., Mack T. Ruffin, IV, Mary A. M. Rogers, and Patrick D. Schloss. 2016. "Microbiota-Based Model Improves the Sensitivity of Fecal Immunochemical Test for Detecting Colonic Lesions." *Genome Medicine* 8 (1). https://doi.org/10.1186/S13073-016-0290-3.

Bokulich, Nicholas A., Benjamin D. Kaehler, Jai Ram Rideout, Matthew Dillon, Evan Bolyen, Rob Knight, Gavin A. Huttley, and J. Gregory Caporaso. 2018. "Optimizing Taxonomic Classification of Marker-Gene Amplicon Sequences with QIIME 2's Q2-Feature-Classifier Plugin." *Microbiome* 6 (1): 1–17. https://doi.org/10.1186/s40168-018-0470-z.

Bolyen, Evan, Jai Ram Rideout, Matthew R. Dillon, Nicholas A. Bokulich, Christian C. Abnet, Gabriel A. Al-Ghalith, Harriet Alexander, et al. 2019. "Reproducible, Interactive, Scalable and Extensible Microbiome Data Science Using QIIME 2." *Nature Biotechnology* 37 (8): 852–57. https://doi.org/10.1038/s41587-019-0209-9.

Callahan, Benjamin J, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. 2016. "DADA2: High-Resolution Sample Inference from Illumina Amplicon Data." *Nature Methods* 13 (7): 581–83. https://doi.org/10.1038/nmeth.3869.

Chen, Lianmin, Valerie Collij, Martin Jaeger, Inge C.L. L. van den Munckhof, Arnau Vich Vila, Alexander Kurilshikov, Ranko Gacesa, et al. 2020. "Gut Microbial Co-Abundance Networks Show Specificity in Inflammatory Bowel Disease and Obesity." *Nature Communications 2020 11:1* 11 (1): 1–12. https://doi.org/10.1038/s41467-020-17840-y.

Dai, Zhenwei, Olabisi Oluwabukola Coker, Geicho Nakatsu, William K.K. Wu, Liuyang Zhao, Zigui Chen, Francis K.L. Chan, et al. 2018. "Multi-Cohort Analysis of Colorectal Cancer Metagenome Identified Altered Bacteria across Populations and Universal Bacterial Markers." *Microbiome* 6 (1): 70. https://doi.org/10.1186/S40168-018-0451-2.

Degruttola, Arianna K., Daren Low, Atsushi Mizoguchi, and Emiko Mizoguchi. 2016. "Current Understanding of Dysbiosis in Disease in Human and Animal Models." *Inflammatory Bowel Diseases* 22 (5): 1137. https://doi.org/10.1097/MIB.0000000000000750.

Fernandes, Andrew D., Jean M. Macklaim, Thomas G. Linn, Gregor Reid, and Gregory B. Gloor. 2013. "ANOVA-Like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq." *PLoS ONE* 8 (7): 67019. https://doi.org/10.1371/JOURNAL.PONE.0067019.

Friedman, Jonathan, and Eric J. Alm. 2012. "Inferring Correlation Networks from Genomic Survey Data." *PLoS Computational Biology* 8 (9).

https://doi.org/10.1371/JOURNAL.PCBI.1002687.

Gabor Csardi, and Tamas Nepusz. 2006. "The Igraph Software Package for Complex Network Research." *InterJournal* Complex Sy. https://igraph.org.

Gevers, Dirk, Subra Kugathasan, Lee A. Denson, Yoshiki Vázquez-Baeza, Will Van Treuren, Boyu Ren, Emma Schwager, et al. 2014. "The Treatment-Naive Microbiome in New-Onset Crohn's Disease." *Cell Host & Microbe* 15 (3): 382–92. https://doi.org/10.1016/j.chom.2014.02.005.

Glassner, Kerri L., Bincy P. Abraham, and Eamonn M.M. Quigley. 2020. "The Microbiome and Inflammatory Bowel Disease." *Journal of Allergy and Clinical Immunology* 145 (1): 16–27. https://doi.org/10.1016/j.jaci.2019.11.003.

Horvath, Steve. 2011. *Weighted Network Analysis*. *Weighted Network Analysis*. Springer New York. https://doi.org/10.1007/978-1-4419-8819-5.

Kurtz, Zachary D., Christian L. Müller, Emily R. Miraldi, Dan R. Littman, Martin J. Blaser, and Richard A. Bonneau. 2015. "Sparse and Compositionally Robust Inference of Microbial Ecological Networks." *PLOS Computational Biology* 11 (5): e1004226. https://doi.org/10.1371/JOURNAL.PCBI.1004226.

Langfelder, Peter, and Steve Horvath. 2008. "WGCNA: An R Package for Weighted Correlation Network Analysis." *BMC Bioinformatics* 9 (1): 1–13. https://doi.org/10.1186/1471-2105-9-559/FIGURES/4.

Li, Bing, Yingying Zhang, Yanan Yu, Pengqian Wang, Yongcheng Wang, Zhong Wang, and Yongyan Wang. 2015. "Quantitative Assessment of Gene Expression Network Module-Validation Methods." *Scientific Reports 2015 5:1* 5 (1): 1–14. https://doi.org/10.1038/srep15258.

Lin, Huang, and Shyamal Das Peddada. 2020. "Analysis of Compositions of Microbiomes with Bias Correction." *Nature Communications 2020 11:1* 11 (1): 1–11. https://doi.org/10.1038/s41467-020-17041-7.

Lloyd-Price, Jason, Cesar Arze, Ashwin N. Ananthakrishnan, Melanie Schirmer, Julian Avila-Pacheco, Tiffany W. Poon, Elizabeth Andrews, et al. 2019. "Multi-Omics of the Gut Microbial Ecosystem in Inflammatory Bowel Diseases." *Nature* 569 (7758): 655–62. https://doi.org/10.1038/s41586-019-1237-9.

Ma, Yongshun, Yao Zhang, Jianghou Xiang, Shixin Xiang, Yueshui Zhao, Mintao Xiao, Fukuan Du, et al. 2021. "Metagenome Analysis of Intestinal Bacteria in Healthy People, Patients With Inflammatory Bowel Disease and Colorectal Cancer." *Frontiers in Cellular and Infection Microbiology* 11 (February): 48. https://doi.org/10.3389/FCIMB.2021.599734/BIBTEX.

Mallick, Himel, Ali Rahnavard, Lauren J. McIver, Siyuan Ma, Yancong Zhang, Long H. Nguyen, Timothy L. Tickle, et al. 2021. "Multivariable Association Discovery in

Population-Scale Meta-Omics Studies." *PLOS Computational Biology* 17 (11): e1009442. https://doi.org/10.1371/JOURNAL.PCBI.1009442.

Mancabelli, Leonardo, Christian Milani, Gabriele Andrea Lugli, Francesca Turroni, Deborah Cocconi, Douwe van Sinderen, and Marco Ventura. 2017. "Identification of Universal Gut Microbial Biomarkers of Common Human Intestinal Diseases by Meta-Analysis." *FEMS Microbiology Ecology* 93 (12): 153. https://doi.org/10.1093/FEMSEC/FIX153.

Matsuoka, Katsuyoshi, and Takanori Kanai. 2015. "The Gut Microbiota and Inflammatory Bowel Disease." *Seminars in Immunopathology* 37 (1): 47. https://doi.org/10.1007/S00281-014-0454-4.

Metwaly, Amira, Andreas Dunkel, Nadine Waldschmitt, Abilash Chakravarthy Durai Raj, Ilias Lagkouvardos, Ana Maria Corraliza, Aida Mayorgas, et al. 2020. "Integrated Microbiota and Metabolite Profiles Link Crohn's Disease to Sulfur Metabolism." *Nature Communications 2020 11:1* 11 (1): 1–15. https://doi.org/10.1038/s41467-020-17956-1.

Mo, Zongchao, Peide Huang, Chao Yang, Sihao Xiao, Guojia Zhang, Fei Ling, and Lin Li. 2020. "Meta-Analysis of 16S RRNA Microbial Data Identified Distinctive and Predictive Microbiota Dysbiosis in Colorectal Carcinoma Adjacent Tissue." *MSystems* 5 (2). https://doi.org/10.1128/MSYSTEMS.00138-20/SUPPL_FILE/MSYSTEMS.00138-20-ST004.XLS.

Mori, Giorgia, Simone Rampelli, Beatrice Silvia Orena, Claudia Rengucci, Giulia De Maio, Giulia Barbieri, Alessandro Passardi, et al. 2018. "Shifts of Faecal Microbiota During Sporadic Colorectal Carcinogenesis." *Scientific Reports 2018 8:1* 8 (1): 1–11. https://doi.org/10.1038/s41598-018-28671-9.

Navas-Molina, José A., Juan M. Peralta-Sánchez, Antonio González, Paul J. McMurdie, Yoshiki Vázquez-Baeza, Zhenjiang Xu, Luke K. Ursell, et al. 2013. "Advancing Our Understanding of the Human Microbiome Using QIIME." *Methods in Enzymology* 531: 371. https://doi.org/10.1016/B978-0-12-407863-5.00019-8.

Nearing, Jacob T., Gavin M. Douglas, Molly G. Hayes, Jocelyn MacDonald, Dhwani K. Desai, Nicole Allward, Casey M.A. Jones, et al. 2022. "Microbiome Differential Abundance Methods Produce Different Results across 38 Datasets." *Nature Communications 2022 13:1* 13 (1): 1–16. https://doi.org/10.1038/s41467-022-28034-z.

Nearing, Jacob T., Gavin M. Douglas, Molly Hayes, Jocelyn MacDonald, Dhwani Desai, Nicole Allward, Casey M. A. Jones, et al. 2021. "Microbiome Differential Abundance Methods Produce Disturbingly Different Results across 38 Datasets." *BioRxiv*, May, 2021.05.10.443486. https://doi.org/10.1101/2021.05.10.443486.

Nistal, Esther, Nereida Fernández-Fernández, Santiago Vivas, and José Luis Olcoz. 2015. "Factors Determining Colorectal Cancer: The Role of the Intestinal Microbiota." *Frontiers in Oncology* 5 (OCT): 220. https://doi.org/10.3389/FONC.2015.00220/BIBTEX.

Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg

Peplies, and Frank Oliver Glöckner. 2013. "The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools." *Nucleic Acids Research* 41 (Database issue): D590-6. https://doi.org/10.1093/nar/gks1219.

Santoru, Maria Laura, Cristina Piras, Antonio Murgia, Vanessa Palmas, Tania Camboni, Sonia Liggi, Ivan Ibba, et al. 2017. "Cross Sectional Evaluation of the Gut-Microbiome Metabolome Axis in an Italian Cohort of IBD Patients." *Scientific Reports* 7 (1). https://doi.org/10.1038/S41598-017-10034-5.

Saus, Ester, Susana Iraola-Guzmán, Jesse R. Willis, Anna Brunet-Vega, and Toni Gabaldón. 2019. "Microbiome and Colorectal Cancer: Roles in Carcinogenesis and Clinical Potential." *Molecular Aspects of Medicine* 69 (October): 93. https://doi.org/10.1016/J.MAM.2019.05.001.

Sobhani, Iradj, Emma Bergsten, Séverine Couffin, Aurélien Amiot, Biba Nebbad, Caroline Barau, Nicola De'angelis, et al. 2019. "Colorectal Cancer-Associated Microbiota Contributes to Oncogenic Epigenetic Signatures." *PNAS*. https://doi.org/10.1073/pnas.1912129116.

Strehl, Alexander, and Ghosh Joydeep. 2002. "Cluster Ensembles---a Knowledge Reuse Framework for Combining Multiple Partitions." *Jmlr.Org* 3: 583–617. https://www.jmlr.org/papers/volume3/strehl02a/strehl02a.pdf.

Sultan, Salma, Mohammed El-Mowafy, Abdelaziz Elgaml, Tamer A.E. Ahmed, Hebatoallah Hassan, and Walid Mottawea. 2021. "Metabolic Influences of Gut Microbiota Dysbiosis on Inflammatory Bowel Disease." *Frontiers in Physiology* 12 (September): 1489. https://doi.org/10.3389/FPHYS.2021.715506/BIBTEX.

Yip, Andy M., and Steve Horvath. 2007. "Gene Network Interconnectedness and the Generalized Topological Overlap Measure." *BMC Bioinformatics* 8 (1): 1–14. https://doi.org/10.1186/1471-2105-8-22/FIGURES/7.

Zeller, Georg, Julien Tap, Anita Y Voigt, Shinichi Sunagawa, Jens Roat Kultima, Paul I Costea, Aurélien Amiot, et al. 2014. "Potential of Fecal Microbiota for Early-Stage Detection of Colorectal." *Molecular Systems Biology* 10 (11): 766. https://doi.org/10.15252/MSB.20145645.

Zhang, Qian, Huan Zhao, Dedong Wu, Dayong Cao, and Wang Ma. 2020. "A Comprehensive Analysis of the Microbiota Composition and Gene Expression in Colorectal Cancer." *BMC Microbiology* 20 (1). https://doi.org/10.1186/S12866-020-01938-W.

CHAPTER 4

An Eensemble Method For Better Phenotype Prediction From Microbial Data

1. Introduction

Over the past decades, there has been an increasing trend of microbiome research focusing on understanding the microbe-host interactions(Levy et al., n.d.; Rooks et al., n.d.). For certain phenotypes, such as colorectal cancer(Baxter et al., 2016; Zeller et al., 2014) and Crohn's disease(Gevers et al., 2014), the area under the curve (AUROC) can achieve over 0.9 using just microbial 16S rRNA sequencing data. At the same time, many studies invested a range of machine learning methods in improving the prediction of the phenotypic outcomes, and our previous publication also examined the usability of microbiome data for accurate disease prediction (Song et al., 2020; Zhou & Gallins, 2019).

However, there are two large limitations to these accuracies. Firstly, even though many models can achieve high accuracy using cross-validation within the same dataset, when we apply the same model to external datasets with a different group of subjects with the same phenotype, the high accuracy drops significantly. Secondly, when we examined the important features identified by the machine learning models, the highly important model does not agree with the key taxa identified by differential abundance analyses. In other words, it is impossible to infer the usefulness of these models in biological terms.

Microbial composition is known to be unstable as there are many environmental factors that can alter the microbial composition of an individual, such as changes in diet, disease progression, or geologic relocation(Bonder et al., 2016; Cheng et al., 2020). In addition, the microbe-disease relationship can be interpreted as an indirect reciprocal relationship, and microbiome data should be evaluated carefully and try to filter out the key microbial taxa, which have shown to be associated with certain diseases.

In this study, we examined the importance of feature selection prior to machine learning methods and demonstrated that while there is a reduction in overall AUROC, the external dataset shows promising results in terms of drastically improved AUROC. Initially, we examined random forest, support vector machine (SVM), extreme boost tree (XGB-Tree), extreme boost linear (XGB-Linear), k-nearest neighbor (KNN), and generalized linear model (E.g., LASSO). By examing the correlation between the methods and the individual AUROC, we optimally selected random forest and XGB-Linear to form an ensemble method.

To validate this novel approach, we examined three vastly different differential abundance taxa identification methods: Firstly, we utilized the Analysis of Compositions of Microbiomes with Bias Correction (ANCOM-BC)(Lin & Peddada, 2020), which is one of the most consistent approaches to identify differentially abundant taxa(Nearing et al., 2022). Next, we utilized the Zero-inflated Beta-binomial Model for Microbiome Data Analysis (ZIBB), which models the count matrix with phenotypes while taking account of the excess zeros(Hu et al., 2018). Lastly, we utilized the influential taxa identified by the Correlation and Consensus-Based Cross-Taxonomy Network Analysis, where these taxa represent significant changes in terms of network structure upon removal.

Finally, we utilize four datasets with two matching phenotype pairs: Colorectal Cancer Vs. Control and Crohn's Disease Vs. Control. Our analyses will evaluate the internal validation as well as external validation on the second dataset. The results will be evaluated in terms of the machine learning AUROC, sensitivity, and specificity, as well as the improvement or decline when using feature-selected taxa only.

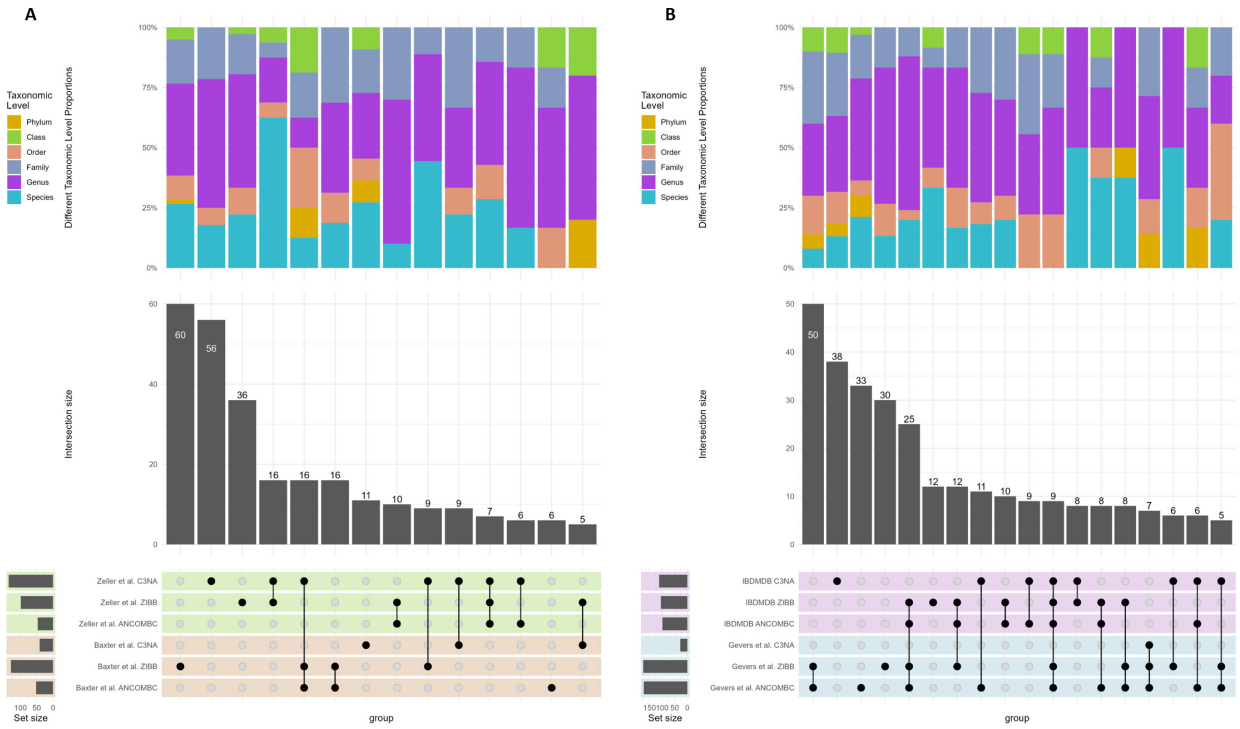


Figure 1. Differential abundance analysis results in comparison among colorectal cancer and Crohn's disease phenotypes and studies. **A.** Colorectal cancer from Bexter et al. and Zeller et al. **B.** Crohn's disease from Gevers et al. and IBDMDB. Each dot represents the presence of the number of taxa (bar height), and a link between the vertical dots represents interactions between the shared differential abundance analyses and study. The top bar chart illustrated the relative abundance of the corresponding taxonomic levels.

2. Results

2.1 Differential Abundance Analyses

2.1.1 Significant taxa identification

We utilized three different methods to identify important taxa that may contain critical information that differentiates between disease and control samples. There is a lack of consistent differentially abundant taxa between the disease and control samples for the four studies we have examined (Fig. 1). Hence, it is important to combine the taxa identification from multiple differential abundance analyses to gain different aspects of the microbial data.

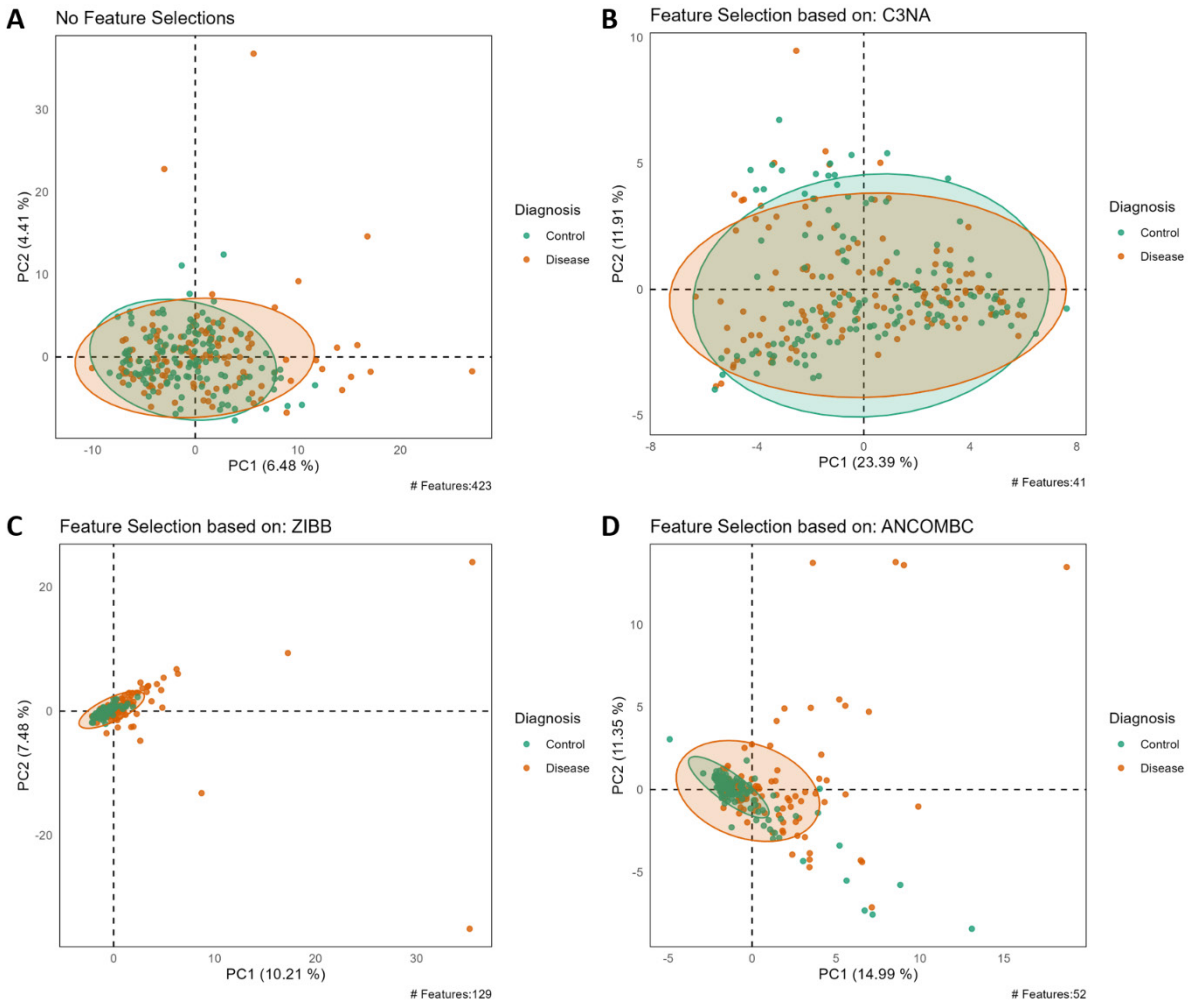


Figure 2. PCA analysis of colorectal cancer and control from the Baxter et al. study from feature selected and non-feature selected approaches. Disease: colorectal cancer; Control: healthy controls. A. No feature selections. B. C3NA. C. ZIBB. D. ANCOM-BC

2.1.2 Principle Component Analysis

We compared the feature selected and non-feature selected cross-taxonomy taxa relative abundance in principle component analysis (PCA) (Fig. 2). We utilized all relative abundance from all Phylum to Species-level taxa to establish the no feature selection PCA and the first two PCs' explain between 10 – 12% of the total variances. When we filter the taxa to only the differential abundance taxa, the sum of the first two PCs' can improve to as high as 40%. In Fig 2., we used Baxter et al. as an example. The total variance explained for two PCs' increased from 10.89% to 35.3%, 17.69%, and 26.34% for C3NA, ZIBB, and ANCOM-BC, respectively (Fig.

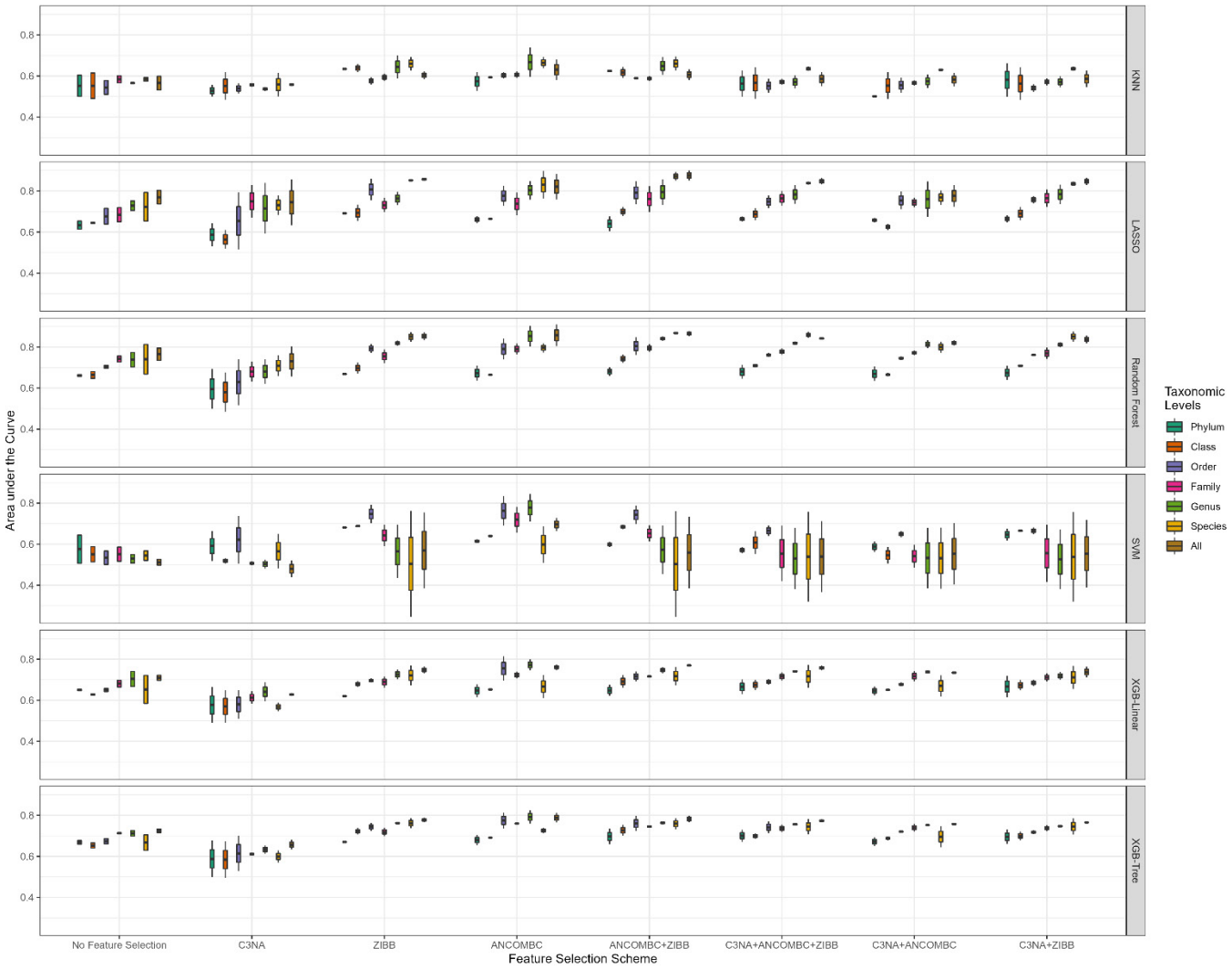


Figure 3. Feature selected taxa from different abundance taxa analyses and their corresponding AUROC across taxonomic levels. The x-axis represents "no feature selection" with different combinations of differential abundance methods. The y-axis represents the area under the curve (AUROC). The fill represents the Phylum, Class, Order, Family, Genus, Species, and All-taxa levels.

2A - 2D). We also observed similar trends in the other three datasets we have examined (Fig. S1 - S3). Overall, this improvement represents a potential for using differentially abundant taxa for the phenotype prediction model.

2.2 Machine Learning Results

2.2.1 Selection of the models for ensemble model

We examined a range of machine learning methods with different combinations of feature selection methods (Fig. 3). There is a clear trend of increasing AUROC with more refined taxonomic assignment methods as we used more refined taxonomic levels for LASSO, Random Forest, and both XGB methods. KNN and SVM did not perform well with inconsistent and low AUROC results. In addition, we applied the same model that trained on one colorectal cancer or Crohn's disease on the second independent dataset to evaluate the performance (Fig. S4); the results showed inconsistent AUROC for almost all machine learning methods. From the taxonomic level perspective, the optimal level is the Genus, Species, and full-taxonomic level.

To select the best model for ensemble methods, we evaluated the correlation among the machine learning methods (Fig. 4), and we want to select a method pair with low correlation and stable AUROC. We choose LASSO and random forest to form the ensemble methods due to the lower correlation between them, and their combination should capture different aspects of the data.

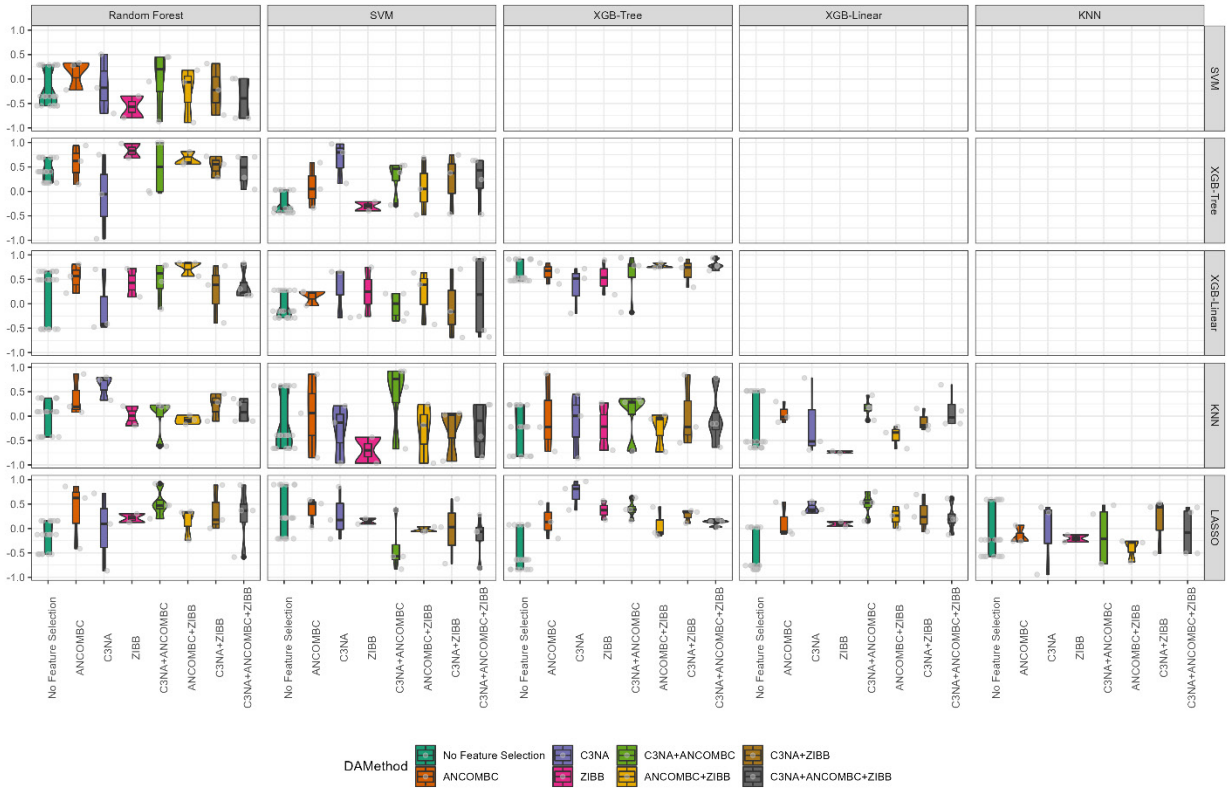


Figure 4. Correlation among the feature selected and unselected methods. The X-axis represents the combination of the differential abundance methods, and the y-axis is the correlation between the machine learning methods.

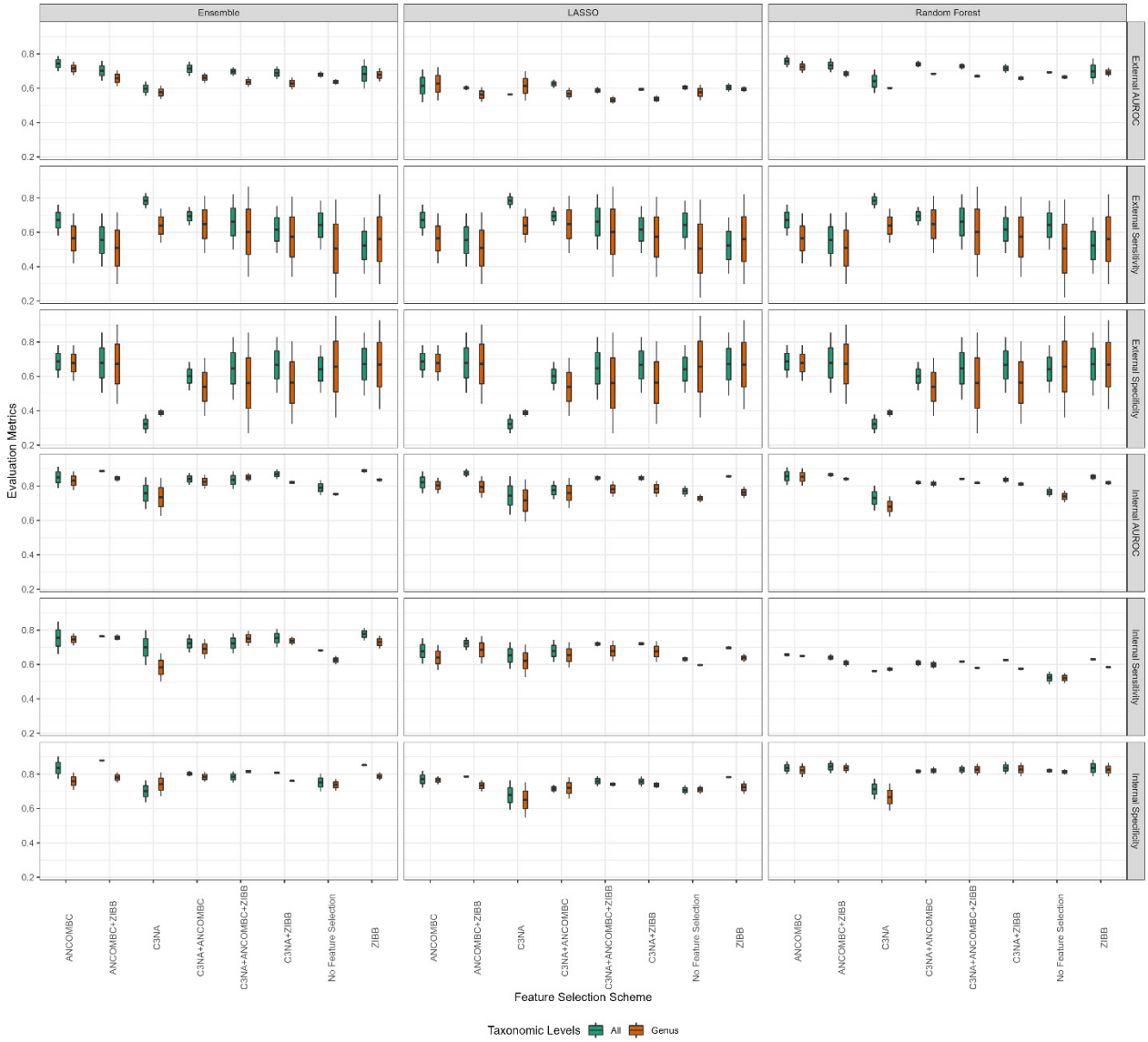


Figure 5. Ensemble methods of LASSO and Random Forest on internal and external testing. The area under the curve (AUROC) is on the y-axis, and the x-axis are the combination of machine learning models.

Table 1. Selected example of ensemble model performance

Cross-Validation Dataset: Baxter et al. (Colorectal Cancer Vs. Healthy Control)

External Dataset: Zeller et al. (Colorectal Cancer Vs. Healthy Control)

DA Methods	Random Forest					
	Interal Testing			External Testing		
	AUROC	Sensitivity	Specificity	AUROC	Sensitivity	Specificity
No Feature Selection	0.795	0.560	0.835	0.697	0.380	0.902
ANCOMBC+ZIBB	0.854	0.623	0.881	0.772	0.600	0.756
ANCOMBC	0.805	0.647	0.801	**0.791**	0.800	0.756
C3NA+ANCOMBC+ZIBB	0.847	0.611	0.853	0.745	0.600	0.780
C3NA+ANCOMBC	0.808	0.629	0.804	0.758	0.760	0.610
C3NA+ZIBB	0.855	0.617	0.873	0.742	0.600	0.707
C3NA	0.657	0.567	0.654	0.572	0.760	0.244
ZIBB	**0.872**	0.622	0.886	0.771	0.560	0.780

DA Methods	LASSO					
	Internal Testing			External Testing		
	AUROC	Sensitivity	Specificity	AUROC	Sensitivity	Specificity
No Feature Selection	0.803	0.648	0.735	0.621	0.720	0.463
ANCOMBC+ZIBB	0.851	0.685	0.792	0.616	0.460	0.732
ANCOMBC	0.758	0.604	0.721	**0.710**	0.640	0.683
C3NA+ANCOMBC+ZIBB	0.863	0.705	0.789	0.568	0.520	0.707
C3NA+ANCOMBC	0.723	0.614	0.693	0.650	0.660	0.610
C3NA+ZIBB	**0.864**	0.710	0.786	0.584	0.520	0.659
C3NA	0.633	0.575	0.592	0.561	0.700	0.390
ZIBB	0.850	0.688	0.787	0.630	0.420	0.756

DA Methods	Ensemble					
	Internal Testing			External Testing		
	AUROC	Sensitivity	Specificity	AUROC	Sensitivity	Specificity
No Feature Selection	0.834	0.684	0.803	0.698	0.500	0.780
ANCOMBC+ZIBB	0.895	0.770	0.882	0.759	0.400	0.854
ANCOMBC	0.789	0.660	0.772	**0.788**	0.580	0.780
C3NA+ANCOMBC+ZIBB	0.887	0.783	0.817	0.722	0.500	0.829
C3NA+ANCOMBC	0.806	0.671	0.786	0.756	0.640	0.683
C3NA+ZIBB	0.898	0.807	0.817	0.727	0.480	0.829
C3NA	0.666	0.598	0.637	0.556	0.740	0.268
ZIBB	**0.900**	0.814	0.856	0.769	0.360	0.854

Cross-Validation Dataset: Gevers et al. (Crohn's Disease Vs. Healthy Control)

External Dataset: IBDMDB (Crohn's Disease Vs. Healthy Control)

DA Methods	Random Forest					
	Interal Testing			External Testing		
	AUROC	Sensitivity	Specificity	AUROC	Sensitivity	Specificity
No Feature Selection	**0.935**	0.992	0.453	0.652	0.870	0.293
ANCOMBC+ZIBB	0.912	0.979	0.493	0.653	0.891	0.293
ANCOMBC	0.818	0.923	0.435	**0.790**	0.435	0.915
C3NA+ANCOMBC+ZIBB	0.911	0.980	0.494	0.646	0.870	0.305
C3NA+ANCOMBC	0.835	0.932	0.443	0.777	0.739	0.756
C3NA+ZIBB	0.914	0.980	0.508	0.639	0.891	0.293
C3NA	0.787	0.897	0.394	0.726	0.674	0.634
ZIBB	0.915	0.981	0.512	0.640	0.891	0.305

DA Methods	LASSO					
	Internal Testing			External Testing		
	AUROC	Sensitivity	Specificity	AUROC	Sensitivity	Specificity
No Feature Selection	**0.914**	0.914	0.695	**0.743**	0.630	0.756
ANCOMBC+ZIBB	0.891	0.903	0.658	0.725	0.435	0.805
ANCOMBC	0.749	0.769	0.514	0.698	0.283	0.902
C3NA+ANCOMBC+ZIBB	0.892	0.914	0.641	0.731	0.435	0.805
C3NA+ANCOMBC	0.767	0.812	0.511	0.704	0.196	0.927
C3NA+ZIBB	0.899	0.898	0.668	0.729	0.435	0.805
C3NA	0.768	0.937	0.206	0.634	0.130	0.902
ZIBB	0.899	0.895	0.670	0.722	0.435	0.805

DA Methods	Ensemble					
	Internal Testing			External Testing		
	AUROC	Sensitivity	Specificity	AUROC	Sensitivity	Specificity
No Feature Selection	**0.946**	0.954	0.786	0.703	0.891	0.280
ANCOMBC+ZIBB	0.927	0.933	0.751	0.698	0.891	0.293
ANCOMBC	0.832	0.896	0.566	**0.790**	0.739	0.829
C3NA+ANCOMBC+ZIBB	0.915	0.936	0.722	0.694	0.891	0.317
C3NA+ANCOMBC	0.841	0.888	0.584	0.779	0.761	0.683
C3NA+ZIBB	0.935	0.943	0.744	0.695	0.891	0.305
C3NA	0.807	0.871	0.454	0.729	0.717	0.610
ZIBB	0.927	0.944	0.720	0.696	0.891	0.305

2.2.2 Ensemble Model Evaluations

We examined the AUROC, sensitivities, and specificities for all taxonomic levels and feature selections from differential abundance analysis combinations (Supplementary Table 1). The results highlight the improvement of AUROC from using ensemble methods compared to a single Random Forest or LASSO model (Fig. 5, Table 1). For example, when comparing the two cancer cases, the internal testing AUROC for the ensemble model has an AUROC of 0.834, whereas the random forest and LASSO are 0.795 and 0.834, respectively. The ensemble methods do not seem to perform well when we use all the features from the microbiome dataset; when we apply at least one of the feature selection methods, there seems to be an improvement of AUROC by roughly 10% for some of the cases.

2.2.3 Variable Importance Evaluations

We investigated the ranking of the features identified by the machine learning model from feature-selected and unselected models (Fig. X). Our previous investigation has shown the high-ranked important features are not necessarily the features identified by biomarker discovery, and our investigation of the four datasets in this study supported this inconsistency. Both the leftmost and rightmost bar plots from Fig. XA and XB are connected to features across the middle green bar range, where the ensemble model variable importance are ranged from highest to lowest from top to bottom. This highlights the disagreement between the machine learning feature selection and differential abundance analysis results. In addition, we also noticed a portion of the non-significant features from one of the studies become important in the second study and vice versa; this change is likely due to study-dependent variation among the population that impacts the machine learning model when testing on external datasets.

When examining how differentially abundant taxa ranked in the all feature model, it seems almost half of all differential abundance models will likely be linked to low important features from machine learning models. This highlights the importance of using more than one differential abundance analysis to gain different types of differential patterns among the microbial data. In our investigations, it is clear that when we combine at least two differential abundance methods, we can double the number of taxa we look into, and the external prediction accuracy will subsequently be improved.

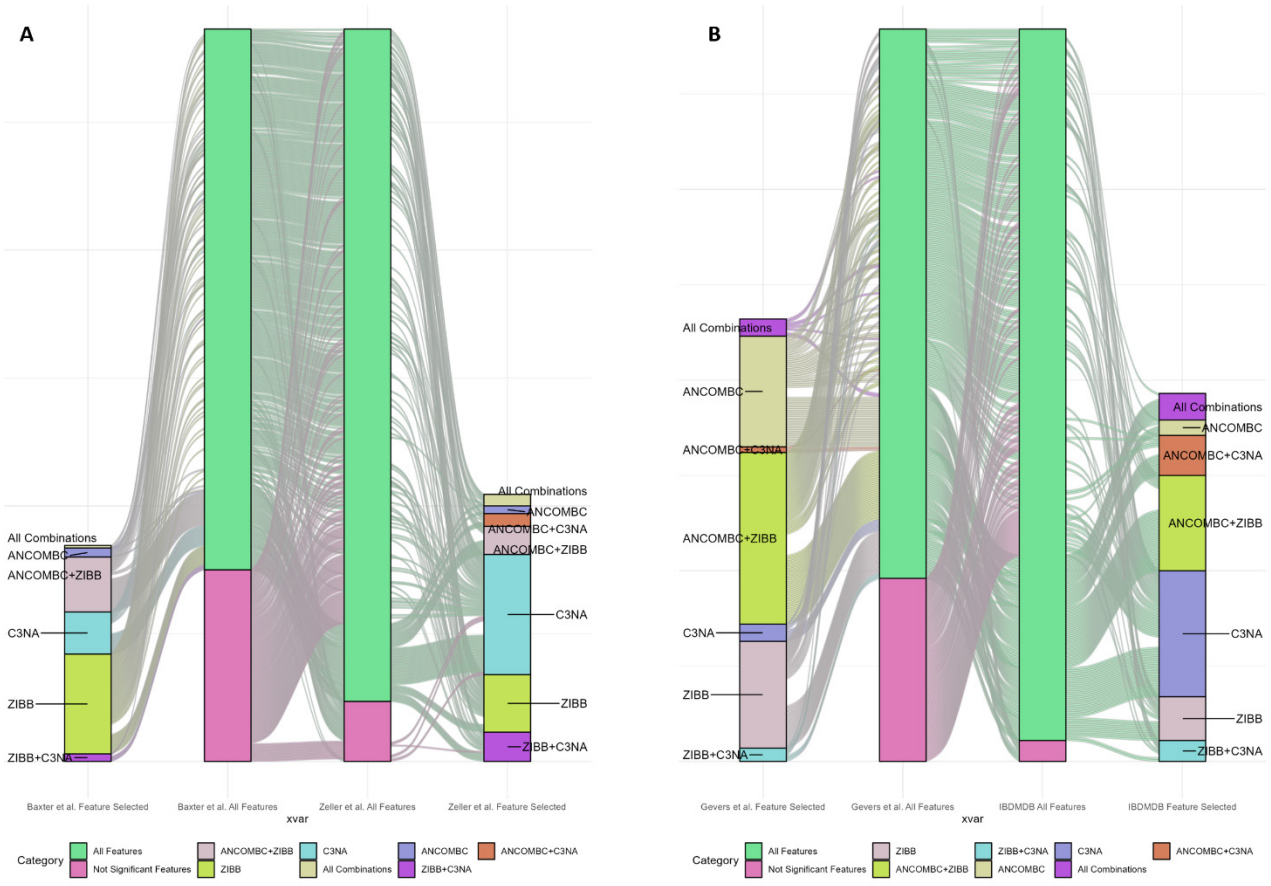


Figure 6. Sankey plot of comparing the ranking of the taxa features between the feature selected and unselected models. **A.** Colorectal cancer datasets from Baxter et al. and Zeller et al. **B.** Crohn's disease datasets from Gevers et al. and IBDMDB. Each bar plot is ranked based on the overall significance of the features by averaging the ranks between all feature models (middle two bar plots).

3. Methods

3.1 Dataset processing

There are four datasets used for this study, including two colorectal cancer datasets and two Crohn's disease datasets, and all datasets were extracted from SRA or designated study web servers. The taxonomic profiling was performed after quality control with the Kraken2/bracken, QIIME2 for 16S rRNA sequencing data. The two 16S rRNA colorectal cancer data are PRJNA290926(Baxter et al., 2016) and PRJEB6070(Zeller et al., 2014), and the two 16S rRNA Crohn's disease datasets are PRJEB13679(Gevers et al., 2014) and IBDMDB(Lloyd-Price et al.,

2019). These 16S rRNA datasets were processed using the DADA2(Callahan et al., 2016) taxonomic assignment method under QIIME2 (QIIME2, n.d.) (version 2021.4) environment with SILVA 138(Quast et al., 2013) reference. Lastly, all taxa that did not present for more than 10% of the samples (disease and control) are removed, as filtered taxa have shown to have better prediction ability.

3.2 **Differential Abundance Analyses**

In our analyses, we selected three differential abundance analysis methods: Analysis of Compositions of Microbiomes with Bias Correction (ANCOM-BC), Zero-inflated Beta-binomial Model (ZIBB), and Correlation and Consensus-Based Network Analysis (C3NA). For the ANCOM-BC, we used the Benjamini-Hochberg (BH) adjusted p-value instead of the default Benferroni-Holm methods. The parameter will include the detection of structural zero, which is used better to suit the unique structure of the microbiome data. All taxa with adjusted p-values below 0.05 were considered differential taxa. For ZIBB, we run by using the default settings with differential taxa defined as BH-adjusted p-value below 0.05. In the C3NA, we compared the phenotype-specific network between the disease and control, and the influential taxa were extracted with BH-adjusted p-values below 0.01.

3.3 **Machine Learning Model Settings**

All the machine learning methods are performed using the caret and caretensemble R package, which calls the “randomForest” package for the random forest model, “glmnet” package for the LASSO model, “class” package for the KNN model, “kernlab” for the SVM model, and “xgboost” for the XGB-Linear and XGB-Tree models. Each model undergoes 100 repeated 5-fold cross-validation. External validation was performed by adding missing/undetected taxa to the testing dataset with all zero’s to enable prediction.

4. Discussion

In this article, we explore an ensemble method aiming to study the usability of combining machine learning methods with differential abundance analyses to gain better external validations. The findings highlight the importance of considering feature selections prior to using machine learning models for phenotype prediction, as this practice will allow better prediction functionality when testing on a second independent dataset. This practice can be seen as removing potentially study-specific taxa that are not associated with the disease directly.

Firstly, we examined three different types of differential abundance analyses, which identified different taxa as significantly different between the disease and control samples. These differences, which are validated by a larger validation study, highlight the importance of using multiple differential abundance analyses by selecting the union of the differential taxa in order to gain more perspectives on the data across all biologically inferable levels.

Next, we examine the performance of the machine learning models across six of the commonly used models. The results were consistent with previous publications and our own study with Genus and all-taxa version having the best predictive values. The study of the correlation between the models suggested pairing LASSO with the random forest as these two methods have the lowest correlation while maintaining high prediction values. The ensemble methods were subsequently generated and evaluated by examing the same phenotype on a second dataset. The result suggested the importance of conducting feature selection prior to the use of a machine learning model to gain prediction value on a secondary dataset. When it comes to feature selection methods, we recommend the use of multiple differential abundance analyses and examine the combination and their resulted prediction values. In our study, the ANCOM-BC

model mostly obtains the highest external AUROC, but the internal AUROC will be lower compared to taxa selected from multiple methods.

Overall, we established a new perspective on how to combine differential abundance analyses with machine learning to discover which taxa can be selected for a consistent phenotype prediction.

References

Baxter, N. T., Ruffin, M. T., IV, Rogers, M. A. M., & Schloss, P. D. (2016). Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Medicine*, *8*(1). https://doi.org/10.1186/S13073-016-0290-3

Bonder, M. J., Kurilshikov, A., Tigchelaar, E. F., Mujagic, Z., Imhann, F., Vila, A. V., Deelen, P., Vatanen, T., Schirmer, M., Smeekens, S. P., Zhernakova, D. V., Jankipersadsing, S. A., Jaeger, M., Oosting, M., Cenit, M. C., Masclee, A. A. M., Swertz, M. A., Li, Y., Kumar, V., … Zhernakova, A. (2016). The effect of host genetics on the gut microbiome. *Nature Genetics*, *48*(11), 1407–1412. https://doi.org/10.1038/ng.3663

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

Cheng, Y., Ling, Z., & Li, L. (2020). The Intestinal Microbiota and Colorectal Cancer. *Frontiers in Immunology*, *11*, 3100. https://doi.org/10.3389/FIMMU.2020.615056/BIBTEX

Gevers, D., Kugathasan, S., Denson, L. A., Vázquez-Baeza, Y., Van Treuren, W., Ren, B., Schwager, E., Knights, D., Song, S. J., Yassour, M., Morgan, X. C., Kostic, A. D., Luo, C., González, A., McDonald, D., Haberman, Y., Walters, T., Baker, S., Rosh, J., … Xavier, R. J. (2014). The Treatment-Naive Microbiome in New-Onset Crohn's Disease. *Cell Host & Microbe*, *15*(3), 382–392. https://doi.org/10.1016/j.chom.2014.02.005

Hu, T., Gallins, P., & Zhou, Y.-H. (2018). A Zero-inflated Beta-binomial Model for Microbiome Data Analysis. *Stat (International Statistical Institute)*, *7*(1), e185. https://doi.org/10.1002/STA4.185

Levy, M., Blacher, E., microbiology, E. E.-C. opinion in, & 2017, undefined. (n.d.). Microbiome, metabolites and host immunity. *Elsevier*. Retrieved April 28, 2022, from https://www.sciencedirect.com/science/article/pii/S1369527416301497?casa_token=qb8APwjZsdEAAAAA:rKg1NrEwm76MlLimasczRDuqPQvFL8o2uc61Z3fMFVntH1-Hck9-iZvIACL7-27wooTMD8RRew

Lin, H., & Peddada, S. Das. (2020). Analysis of compositions of microbiomes with bias correction. *Nature Communications 2020 11:1*, *11*(1), 1–11. https://doi.org/10.1038/s41467-020-17041-7

Lloyd-Price, J., Arze, C., Ananthakrishnan, A. N., Schirmer, M., Avila-Pacheco, J., Poon, T. W., Andrews, E., Ajami, N. J., Bonham, K. S., Brislawn, C. J., Casero, D., Courtney, H., Gonzalez, A., Graeber, T. G., Hall, A. B., Lake, K., Landers, C. J., Mallick, H., Plichta, D. R., … Huttenhower, C. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, *569*(7758), 655–662. https://doi.org/10.1038/s41586-019-1237-9

Nearing, J. T., Douglas, G. M., Hayes, M. G., MacDonald, J., Desai, D. K., Allward, N., Jones, C. M. A., Wright, R. J., Dhanani, A. S., Comeau, A. M., & Langille, M. G. I. (2022).

Microbiome differential abundance methods produce different results across 38 datasets. *Nature Communications 2022 13:1*, *13*(1), 1–16. https://doi.org/10.1038/s41467-022-28034-z

QIIME2. (n.d.). *OTU picking strategies in QIIME — Homepage*. Retrieved August 10, 2020, from http://qiime.org/tutorials/otu_picking.html

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, *41*(Database issue), D590-6. https://doi.org/10.1093/nar/gks1219

Rooks, M., immunology, W. G.-N. reviews, & 2016, undefined. (n.d.). Gut microbiota, metabolites and host immunity. *Nature.Com*. Retrieved April 28, 2022, from https://idp.nature.com/authorize/casa?redirect_uri=https://www.nature.com/articles/nri.2016.42&casa_token=pgLCRtZ_qXQAAAAA:ZilzhmMR5V4lqkIAUw1WNpSU5aJU7k2gF_d63vPmlYtpBF5eYnSa4FBi8ysKYR3bN3f7hww1NpH4t7Ns

Song, K., Wright, F. A., & Zhou, Y.-H. H. (2020). Systematic Comparisons for Composition Profiles, Taxonomic Levels, and Machine Learning Methods for Microbiome-Based Disease Prediction. *Frontiers in Molecular Biosciences*, *7*, 610845. /pmc/articles/PMC7772236/

Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kultima, J. R., Costea, P. I., Amiot, A., Böhm, J., Brunetti, F., Habermann, N., Hercog, R., Koch, M., Luciani, A., Mende, D. R., Schneider, M. A., Schrotz-King, P., Tournigand, C., Nhieu, J. T. Van, Yamada, T., … Bork, P. (2014). Potential of fecal microbiota for early-stage detection of colorectalcancer. *Molecular Systems Biology*, *10*(11), 766. https://doi.org/10.15252/MSB.20145645

Zhou, Y.-H., & Gallins, P. (2019). A Review and Tutorial of Machine Learning Methods for Microbiome Host Trait Prediction. *Frontiers in Genetics*, *10*(JUN), 579. https://doi.org/10.3389/fgene.2019.00579

APPENDICES

Appendix A: Supplementary Materials for Chapter 2

DATA PROCESSING

Stage 1 – OTUs, ASVs, and K-mers Generation

The raw sequences first undergo FastQC to check the read qualities; there were no bad reads with quality scores below 20. Thus, all reads were subsequently imported to QIIME2. These demultiplexed reads undergo chimeric sequences removal as part of the QIIME2 (Version 2019.10) 16S rRNA analysis pipeline. (Bolyen et al., 2019) The *de novo*, open-reference, and closed-reference were performed all with a 97% similarity using QIIME2 version 2019.10. (Bolyen et al., 2019) Similarly, DADA2 was also performed using QIIME2 using the demultiplexed reads directly. (Bolyen et al., 2019; Callahan et al., 2016) All these four methods used the same reference provided by the SILVA rRNA database project Version 132 release. (Quast et al., 2013) The final output from these procedures is four OTUs/ASVs-level tables, and we subsequently extract the higher level OTUs/ASVs count matrices. We have also extracted seven levels of OTUs/ASVs matrices based on their taxonomic assignments: phylum, class, order, family, genus, species, and OTUs/ASV. We also removed three taxonomy assignments from these count matrices: missing, ambiguous taxa, or unassigned. The full processing workflow is shown in Figure **S1**.

Raw k-mer calculations were done using R (Version 4.0.1) (R Core Team, 2013) by breakdown each of the FASTQ raw sequences and counts each of the fragments that to a raw counting matrix. We split the k-mers length in our example to very short k-mers and short k-mers, the very short k-mers are k-mers range from 4-mers, 5-mers, 6-mers, and 7-mers, and the short k-mers are 15-mers, 21-mers, and 30mers. Both k-mer categories have been used to study for their phenotype prediction properties. (Asgari et al., 2018; Molik et al., 2020) Since we are

trying to preserve as many unique reads and their associated counts as possible and at the same time, make the analyses computationally feasible, during the k-mer extraction process for the short k-mers, we have removed any reads less than five counts out of each sample. This process reduces the resulted dataset and dimension by at least half.

Through this stage, we have generated 21 OTUs count matrices, 7 ASV count matrices, 4 very short-chain k-mer matrices, and 3 short-chain k-mer matrices.

Stage 2 Normalization

Previous publications have investigated the potential usefulness of utilizing normalization technology on the count matrices. (Weiss et al., 2017) This idea stemmed from the normalization of RNA-Seq data, which is another large count matrix similar to OTUs/ASVs matrices. The key difference is the OTUs table is zero-inflated, which is not a common problem in the RNA-Seq data (except single-cell RNA-Seq). Thus, some of the methods used for RNA-Seq transformation might benefit from adjustment, which is an active research field. In this project, we chose DESeq2, which was recommended from a previous publication, though its usefulness from simulated data was controversial. (McMurdie & Holmes, 2014; Weiss et al., 2017) Briefly, DESeq2 models the counts with Negative Binomial to detect the differential abundance while accounting for the sampling depth and OTUs/ASVs composition. Due to the large dimension of the short-chain k-mers, which are the 15-mer, 21-mer and 30-mer and we removed the count lower than 5 reads per sample when generating the final count matrices.

After this stage, we have obtained 42 OTUs count matrices, 14 ASV count matrix, 8 very short-chain k-mer matrices, and 3 short-chain k-mer matrices.

Stage 3 Filtering the Samples

For the OTUs and ASVs count matrices undergo the three filtering criteria reported previously.(Duvallet et al., 2017; Goodrich et al., 2014; Zhou & Gallins, 2019) The first filter excludes the sample with less than 100 reads, and the second filter subtracts OTUs with less than 10 reads. (Duvallet et al., 2017) The third filter removes OTUs that present less than 5% of samples.(Goodrich et al., 2014) We detail the number of features prior to the machine learning algorithm is shown in table [!Sheet1 in the OTUsASVsTables.xlsx]. Generally, the filtering removed fewer features (OTUs/ASVs) with a more specific taxonomic level, i.e. more features were kept on the species level compared to class level. There are many ways that filtering can undergo, other studies have implemented a less rigorous third rule, which only removed OTUs that present less than 1% of samples (Ross et al., 2015; Singh et al., 2015; Vincent et al., 2013)

After this stage, we have obtained 84 OTUs count matrices, 28 ASV count matrices, 8 very short-chain k-mer matrices, and 3 short-chain k-mer matrices.

Stage 4 Machine Learning Methods

The details for most of the machine learning methods except logistic regression were part of our previous work and explained in great detail previously. (Zhou & Gallins, 2019) All the methods were tested against a binary outcome, Disease Vs. Control.

Among the 11 machine learning algorithms we have tested, ten methods were part of the supervised learning methods. The LASSO (Tibshirani, 1996), Ridge (Hoerl & Kennard, 1970), Elastic Net (Zou & Hastie, 2005) and Logistic (Nelder., 1989) are the regression aspects of the supervised learning. The Support Vector Machine (SVM)(Cortes & Vapnik, 1995), Gradient Boost (XgBoost) (Friedman, 1999), random forest (Breiman, 2001), K-nearest Neighbors, Hierarchical clustering, and Neural Network (Ditzler et al., 2015) represents different methods

within the supervised classification category. Lastly, we also used a K-means method to shed some light on unsupervised machine learning methods.

After this stage, our analyses have undergone 924 OTUs-based predictions, 308 ASV-based predictions, 88 very short-chain k-mer-based predictions, and 33 short-chain k-mer-based predictions. This sums up to a total of 1,353 combinations per disease type.

Stage 5 Evaluation

K-Fold cross-validation is a commonly used resampling protocol for evaluating machine learning methods. In our pipeline, we utilize a 5-fold cross-validation scheme with 100 iterations. Briefly, at the beginning of each iteration, we randomly break down the sample consisting of the diseased and control subjects to 5 roughly equal-size groups. Then each of these five groups was used as the testing set, while the remaining four groups were selected as the training set. At the end of each 5-fold validation, we extract all the predicted values from each of the 11 methods and save them prior to running the next iterations. For the TwinsUK dataset, we modified the sampling methods to ensure twins who came from the same family were kept in the same training/testing set.

Overall, we have just over 5,412 combinations tested.

References

Asgari, E., Garakani, K., McHardy, A. C., & Mofrad, M. R. K. (2018). MicroPheno: predicting environments and host phenotypes from 16S rRNA gene sequencing using a k-mer based representation of shallow sub-samples. *Bioinformatics (Oxford, England)*, *34*(13), i32–i42. https://doi.org/10.1093/bioinformatics/bty296

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., … Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, *37*(8), 852–857. https://doi.org/10.1038/s41587-019-0209-9

Breiman, L. (2001). *Random Forests* (Vol. 45).

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

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/bf00994018

Ditzler, G., Polikar, R., & Rosen, G. (2015). Multi-Layer and Recursive Neural Networks for Metagenomic Classification. *IEEE Transactions on Nanobioscience*, *14*(6), 608–616. https://doi.org/10.1109/TNB.2015.2461219

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

Friedman, J. H. (1999). Greedy Function Approximation : A Gradient Boosting Machine 1 Function estimation 2 Numerical optimization in function space. *North*, *1*(3), 1–10. https://doi.org/10.2307/2699986

Goodrich, J. K., Waters, J. L., Poole, A. C., Sutter, J. L., Koren, O., Blekhman, R., Beaumont, M., Treuren, W. Van, Knight, R., Bell, J. T., Spector, T. D., Clark, A. G., & Ley, R. E. (2014). Human genetics shape the gut microbiome. *Cell*, *159*(4), 789. https://doi.org/10.1016/J.CELL.2014.09.053

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics*, *12*(1), 69–82. https://doi.org/10.1080/00401706.1970.10488635

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

Molik, D. C., Pfrender, M. E., & Emrich, S. J. (2020). Uncovering Effects from the Structure of

Metabarcode Sequences for Metagenetic and Microbiome Analysis. *Methods and Protocols*, *3*(1), 22. https://doi.org/10.3390/mps3010022

Nelder., P. M. and J. A. (1989). *Generalized linear models*. London ; New York : Chapman and Hall, 1989. https://catalog.lib.ncsu.edu/catalog/NCSU4818332

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, *41*(Database issue), D590-6. https://doi.org/10.1093/nar/gks1219

R Core Team. (2013). *R: A language and environment for statistical computing*. http://cran.univ-paris1.fr/web/packages/dplR/vignettes/intro-dplR.pdf

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

Singh, P., Teal, T. K., Marsh, T. L., Tiedje, J. M., Mosci, R., Jernigan, K., Zell, A., Newton, D. W., Salimnia, H., Lephart, P., Sundin, D., Khalife, W., Britton, R. A., Rudrik, J. T., & Manning, S. D. (2015). Intestinal microbial communities associated with acute enteric infections and disease recovery. *Microbiome*, *3*(1), 45. https://doi.org/10.1186/s40168-015-0109-2

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

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

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E. R., & Knight, R. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, *5*(1), 27. https://doi.org/10.1186/s40168-017-0237-y

Zhou, Y.-H., & Gallins, P. (2019). A Review and Tutorial of Machine Learning Methods for Microbiome Host Trait Prediction. *Frontiers in Genetics*, *10*(JUN), 579. https://doi.org/10.3389/fgene.2019.00579

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320. https://doi.org/10.1111/J.1467-9868.2005.00503.X@10.1111/(ISSN)1467-9868.TOP_SERIES_B_RESEARCH

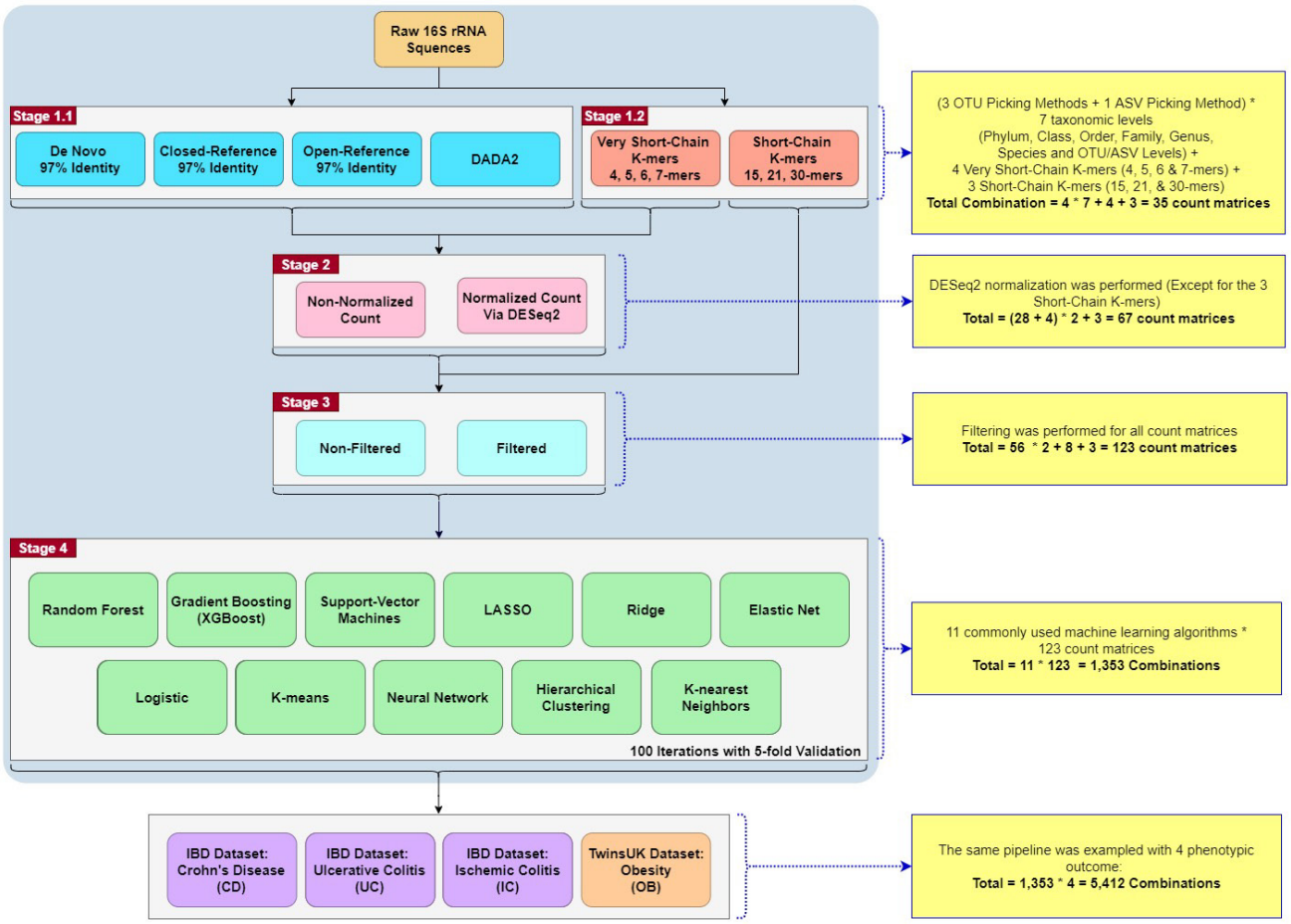


Figure S1. Workflow of the projects with the calculation of count matrices or combinations at each stage.

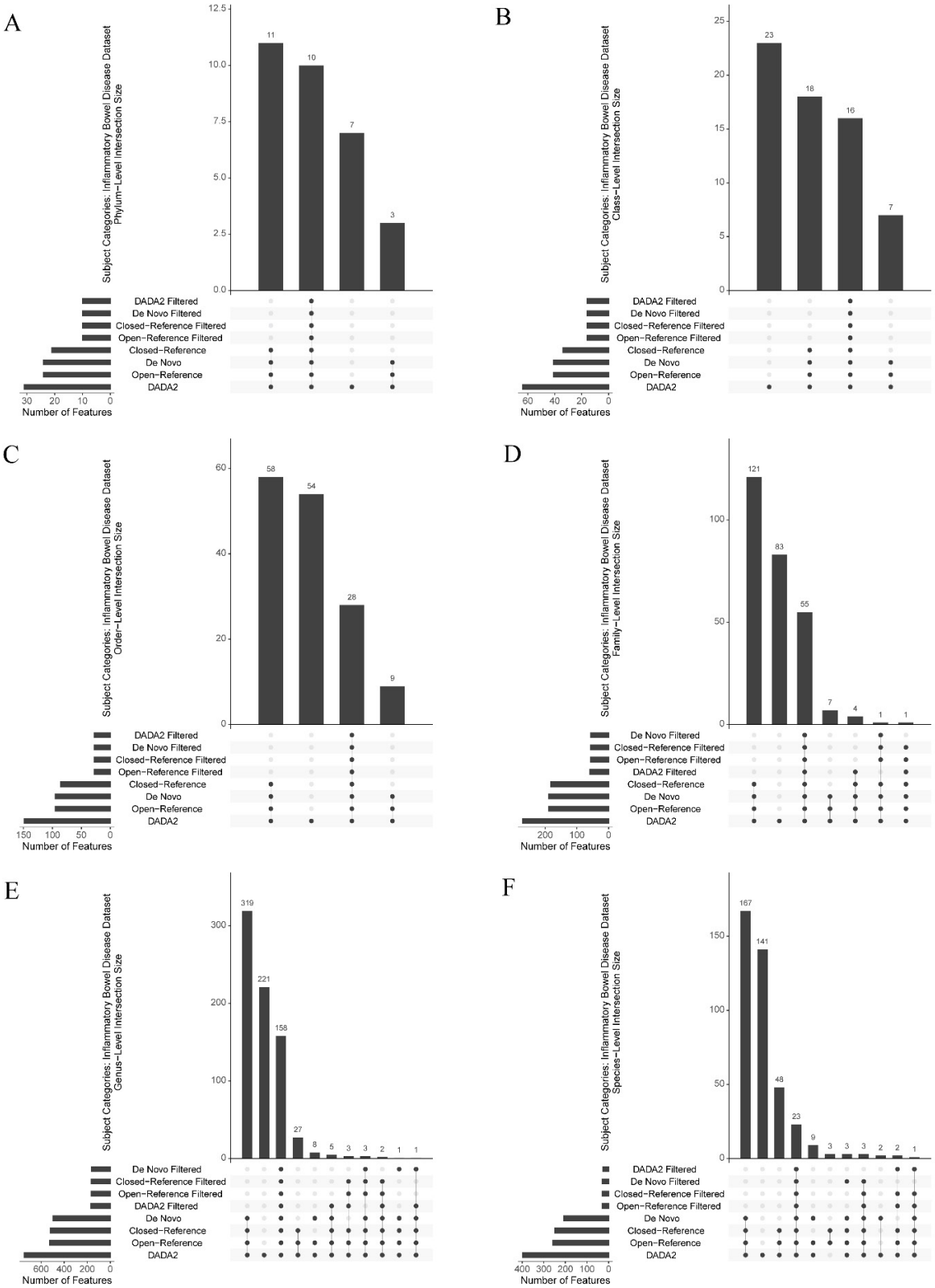


Figure S2. Upset plot for the interaction of features for all three Inflammatory Bowel Disease diagnoses and Control. Including the filtered and non-filtered OTU/ASV picking methods at different taxonomic levels. **(A-F)** Phylum, Class, Order, Family, Genus and Species-level.

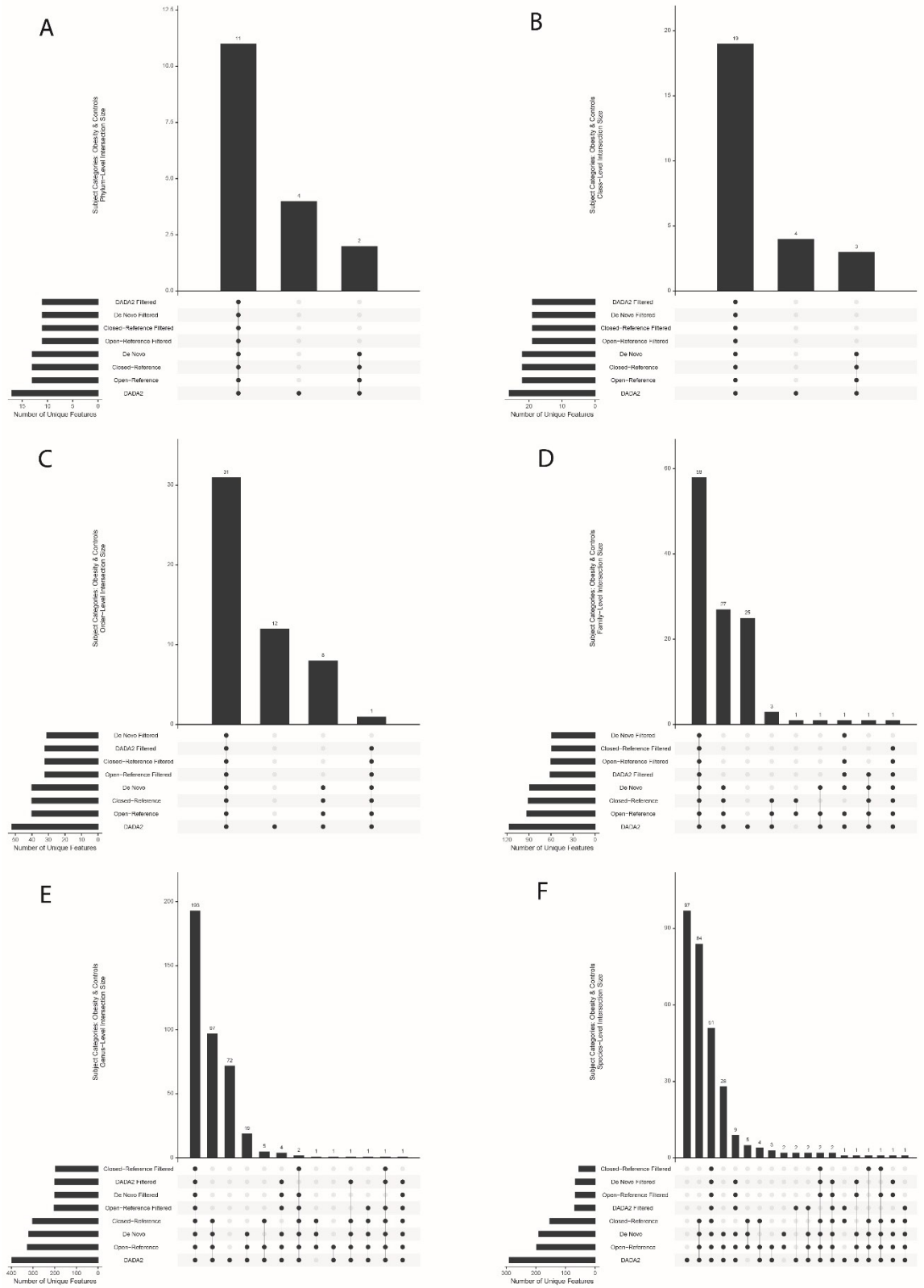


Figure S3. Upset plot for the interaction of features for Obesity and Control. Including the filtered and non-filtered OTU/ASV picking methods at different taxonomic levels. **(A-F)** Phylum, Class, Order, Family, Genus and Species-level.

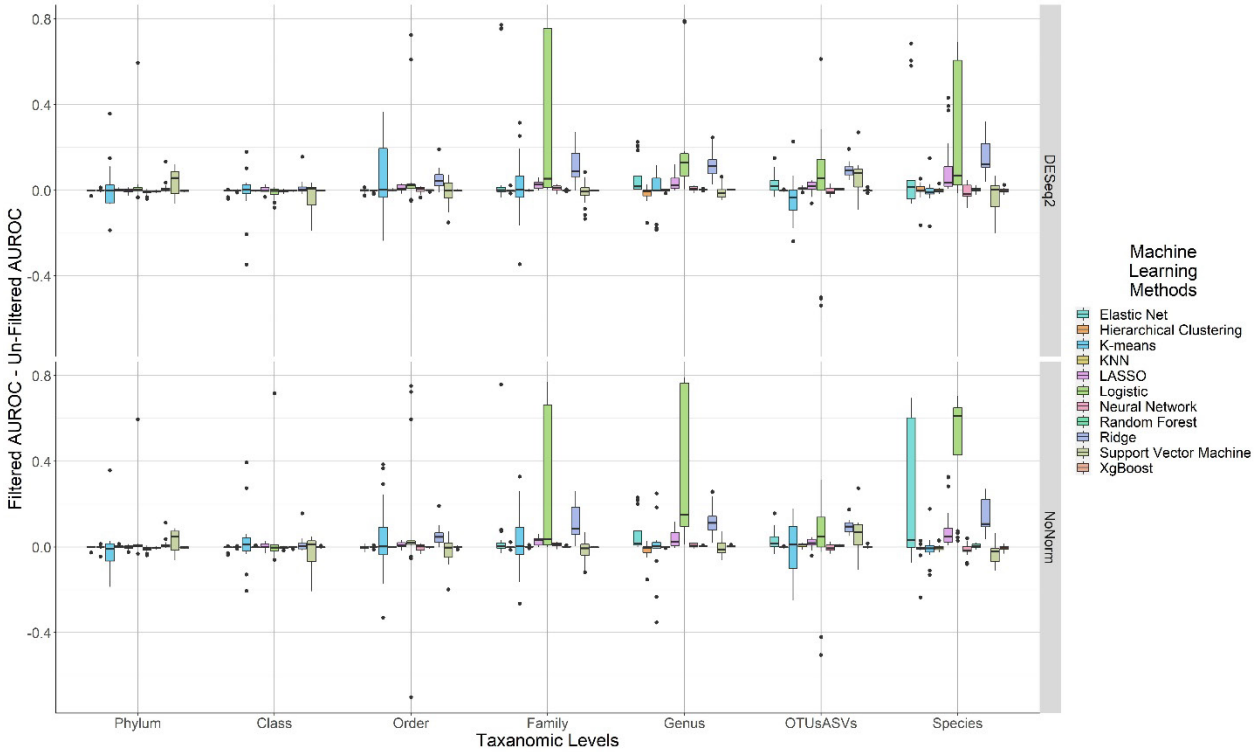


Figure S4. Difference in AUROC Caused by Filtering between Normalized and Unnormalized Combinations. The Y-axis is the difference between the filtered and unfiltered AUROC when holding other parameter at the same level. If the results are above 0, it means the filtered combination generated a better AUROC compared to the unfiltered combination. Top row is the results from DESeq2 normalization. Bottom row is the result from no normalization.

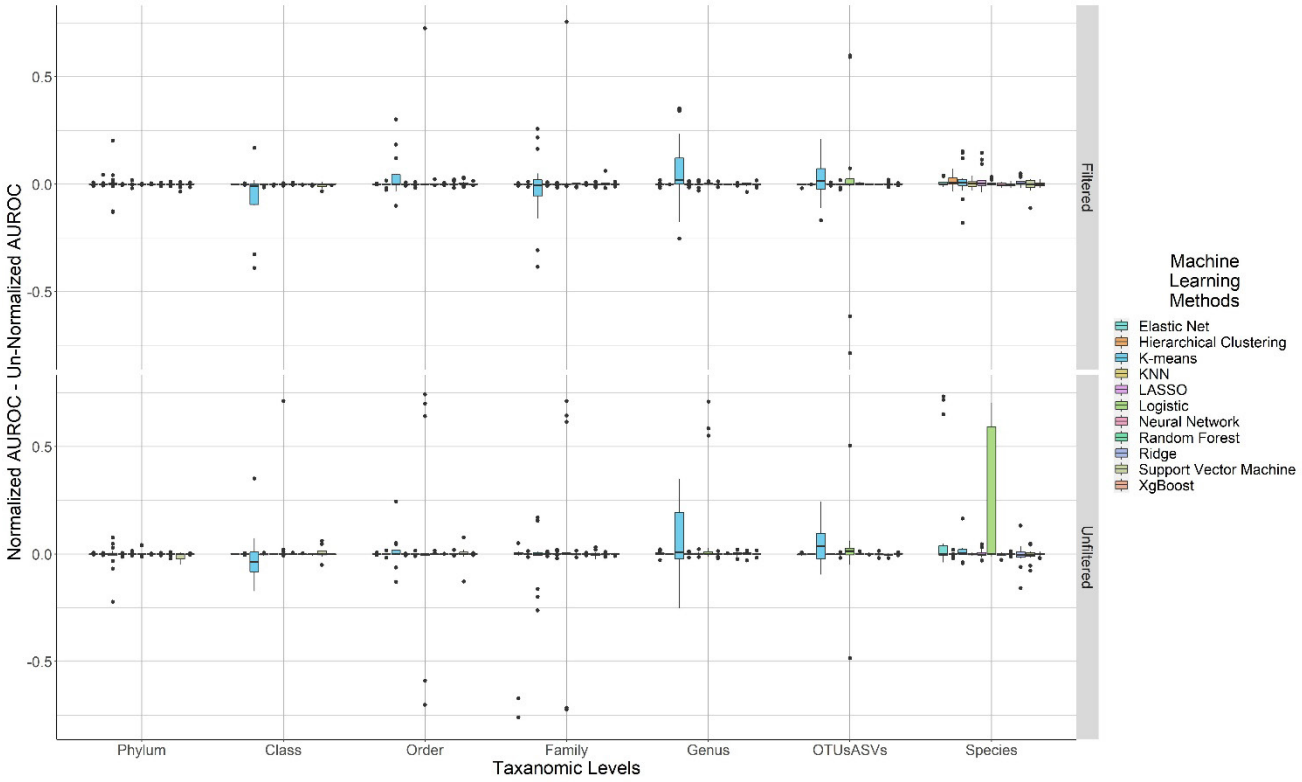


Figure S5. Difference in AUROC Caused by Normalization between Filtered and Unfiltered Combinations. The Y-axis is the difference between the normalized and unnormalized AUROC when holding other parameter at the same level. If the results are above 0, it means the normalized combination generated a better AUROC compared to the unnormalized combination. Top row is the results from DESeq2 normalization. Bottom row is the result from no normalization.

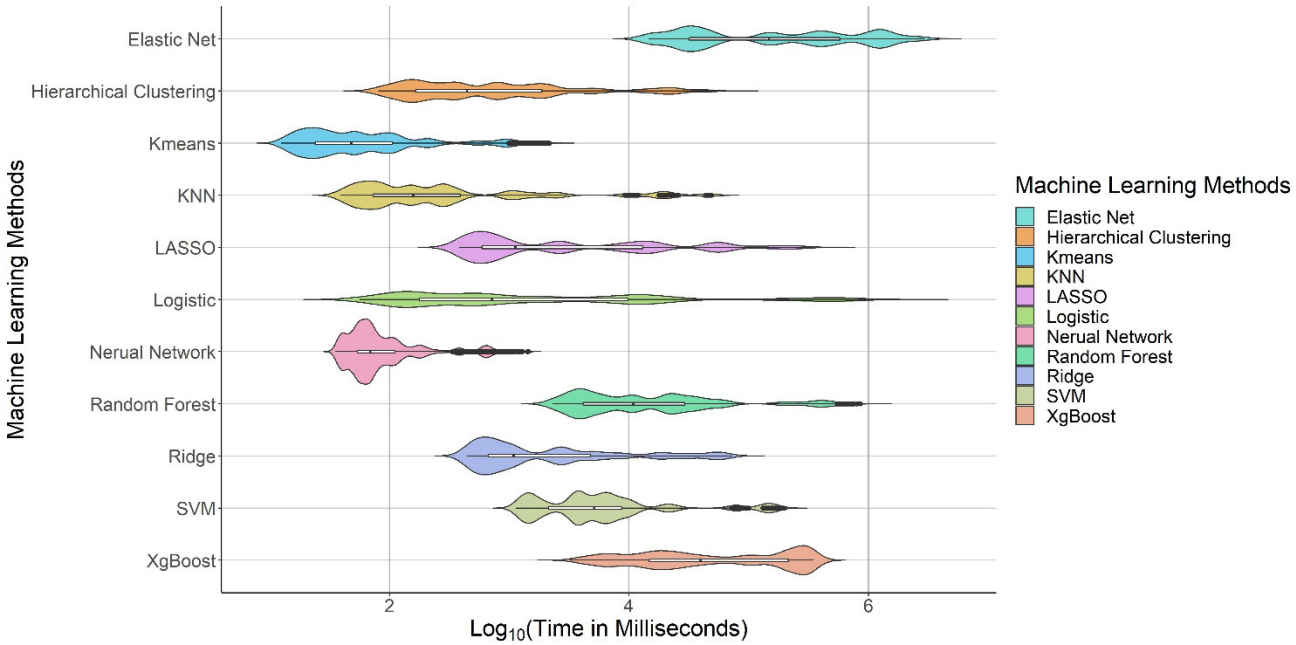


Figure S6. Benchmarking the Machine Learning Algorithms. The X-axis is the $\log_{10}$-transformed milliseconds.

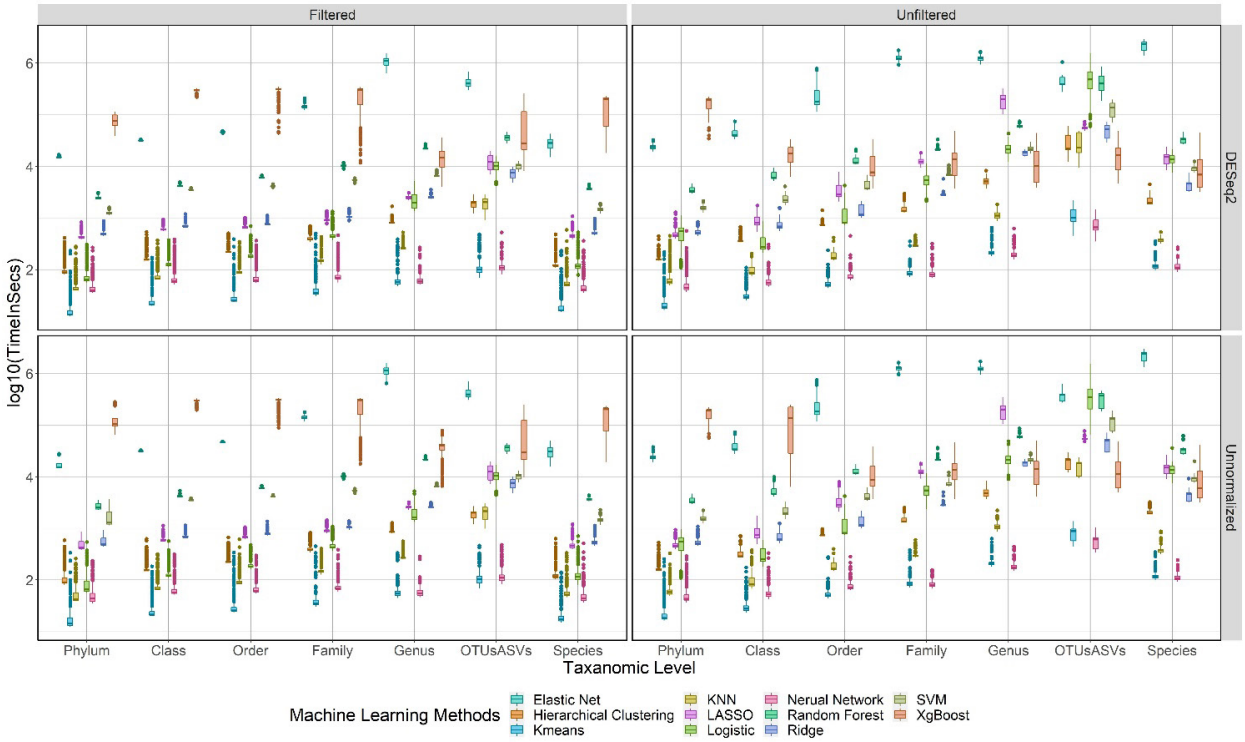


Figure S7. Benchmarking of machine learning algorithms faceted by the normalization and filtering status.

Appendix B: Supplementary Materials for Chapter 3

Supplementary Results

SparCC Results Comparison between Single-Taxonomic Level and Cross-Taxonomic Levels

C3NA filtered the SparCC results with the taxa-taxa correlations (BH-adjusted p-value ≤ 0.05) and has positive correlations of at least 0.2, which is determined by examining the SparCC's stability to find correlations from single-taxonomic level SparCC runs.

We benchmarked the comparisons using the Baxter et al. with the "Cancer" phenotype, and each of the runs will include 1,000 iterations of SparCC, and the results are extracted and compared. Firstly, we examine the correlation pairs detected by both single- and multi-taxonomic level runs; the differences are removed and shown in Fig. S1A. The correlation differences between them are primarily between ±0.1. Next, we examined the unique correlations found from either method, as shown in Fig. S1B, and these individual correlations are primarily below 0.2. As a result, we recommended a 0.2 correlation threshold for filtering the taxa-taxa correlation as this is a suitable threshold to detect significant correlations and not too high to omit essential correlations.

Differential Abundance among Studies

While there are shared taxa identified by different DA methods, there are still disagreements among them, with ANCOM-BC capturing more than the other two methods. C3NA, by definition, captures different taxa as the interpretation of C3NA influential taxa highlights a change in modular membership between the two phenotypes, not necessarily representing the differentially abundant taxa. Supplementary Fig. S2-5 represents the differential abundance results from Baxter et al., Zeller et al., Gevers et al., and IBDMBD datasets, respectively.

Impact of Filtering on Taxa Identification across Studies and OTUs/ASVs Assignment Methods

In our investigation, we used four different datasets, each undergoing taxonomic assignments via *de novo* and DADA2, two very different methodologies. There are taxa rare taxa across all six taxonomic levels identified and filtered out across studies and Crohn's disease and Colorectal Cancer phenotypes. When we compared the remaining taxa across all six levels, we found consistent patterns in which the majority (215) taxa were identified; in addition, the results also highlighted disease-specific and study-specific taxonomic assignments (Supplementary Fig. S6). We also investigated the removed taxa, and their patterns are unique to the study and the clustering methods (Supplementary Fig. Sd7). The number of filtered taxa and original taxa are shown in Supplementary Table 1, and this reduction of rare taxa corresponds to a decrease in computation time, which used 12 cores on Intel(R) Xeon(R) CPU E5-2670 0 @ 2.60GHz with 30 GB of RAM.

Effect of Consensus-based on Different pattern selections

When we examined the module membership changes from different minimal module sizes, there is a clear trend of decreasing changes in module memberships, and as we reduce the number of modules below ten, there will be more duplicated patterns (Fig. S1-S12). We selected all unique designs with minimal module sizes equal to or greater than ten for all our investigations, providing consistent results.

We examine the impact of different module patterns and the optimal number of clusters based on the consensus matrix by examining the result of remaining significant correlations after clustering. Using different selected module patterns and a range of optimal clusters, we evaluated how many key correlations remain. The results show consistency in important correlations, particularly for the essential correlations greater or equal to 0.2. We determine the final taxa

cluster through clustering of the consensus matrix (Fig. S12-S26), there is a clear trend of a quick increase of average silhouette widths with the increased number of clusters, and it gradually plateaued after 15 clusters. As Supplementary Fig. S9 show, we can generally categorize the silhouette plot into three categories, the first panel is the "base dynamic region" (minimal number of taxa per module from 3 to 12), where we include all the dynamically changing regions, and with more patterns, the silhouette width plot gradually forms a curve with a plateau region from linear trends. The second panel includes the "less dynamic region," which includes the region with a few duplicated patterns, but the pattern generally changes (small changes) with each increment (the minimal number of taxa per module ranges from 13 to 25). The third panel includes the "stable region" where the majority of repetitive patterns occur. C3NA recommends the user to pick any of the dynamic regions which can be identified by minimal repetitive region with the number of modules for each pattern between 10 to 20. As the first and third panel indicates selections, other regions might lead to an unstable silhouette plot for the optimal number of cluster determination.

Next, we evaluated the impact of the remaining number of significant intra-modular correlations greater or equal to 0.2 from each of the parameters we observed with a similar number of clusters with an optimal number of clusters greater than 15. This also indicates the stability of the consensus-based clustering; with enough patterns selected, slight differences in a few patterns and different numbers of optimal clusters will not drastically affect the downstream correlation results. For example, when we look at the minimal number of taxa per module from 15 to 22 (Supplementary Fig. S9B, x-axis), after the proportion of zeros per cluster dropped below 10% (uncrossed-off numbers), the number of intra-modular correlations between the

optimal number of clusters 15 – 22 are all around 1,300 ± 100. C3NA advises using an optimal number of clustering as low as possible once the silhouette plot enters the plateau region.

SparCC Stability with Different Number of Bootstraps

We use the DADA2 ASV assignment method with the Cancer samples from Baxter et al. to evaluate the impact of iterations on the stability of SparCC (Supplementary Fig. S11). For each of the 10, 25, 50, 100, and 500 iterations, we run six different bootstrap rounds. We investigate the difference in terms of taxa-taxa correlations greater or equal to 0.2 with BH-adjusted p-values less or equal to 0.05. The results show that most of the correlations are shared compared to 1,000 iterations. The most significant difference is the total number of these correlation pairs, with 1,000 iterations having roughly 16% fewer correlations compared to that with ten iterations, though most of the missed correlations are less than 0.3. Thus, when running a smaller number of iterations for preliminary investigation and evaluation under the Shiny application, the user should filter the correlations to 0.3. We also evaluated the time consumption while running these bootstraps. The computation time for 10, 25, 50, 100, and 500 using one core on Intel(R) Xeon(R) CPU E5-2670 0 @ 2.60GHz are 10 minutes, 1.25 hours, 3 hours, 10 hours, and 34 hours, respectively.

Optimal Number of Clusters Selection for Datasets included in the study

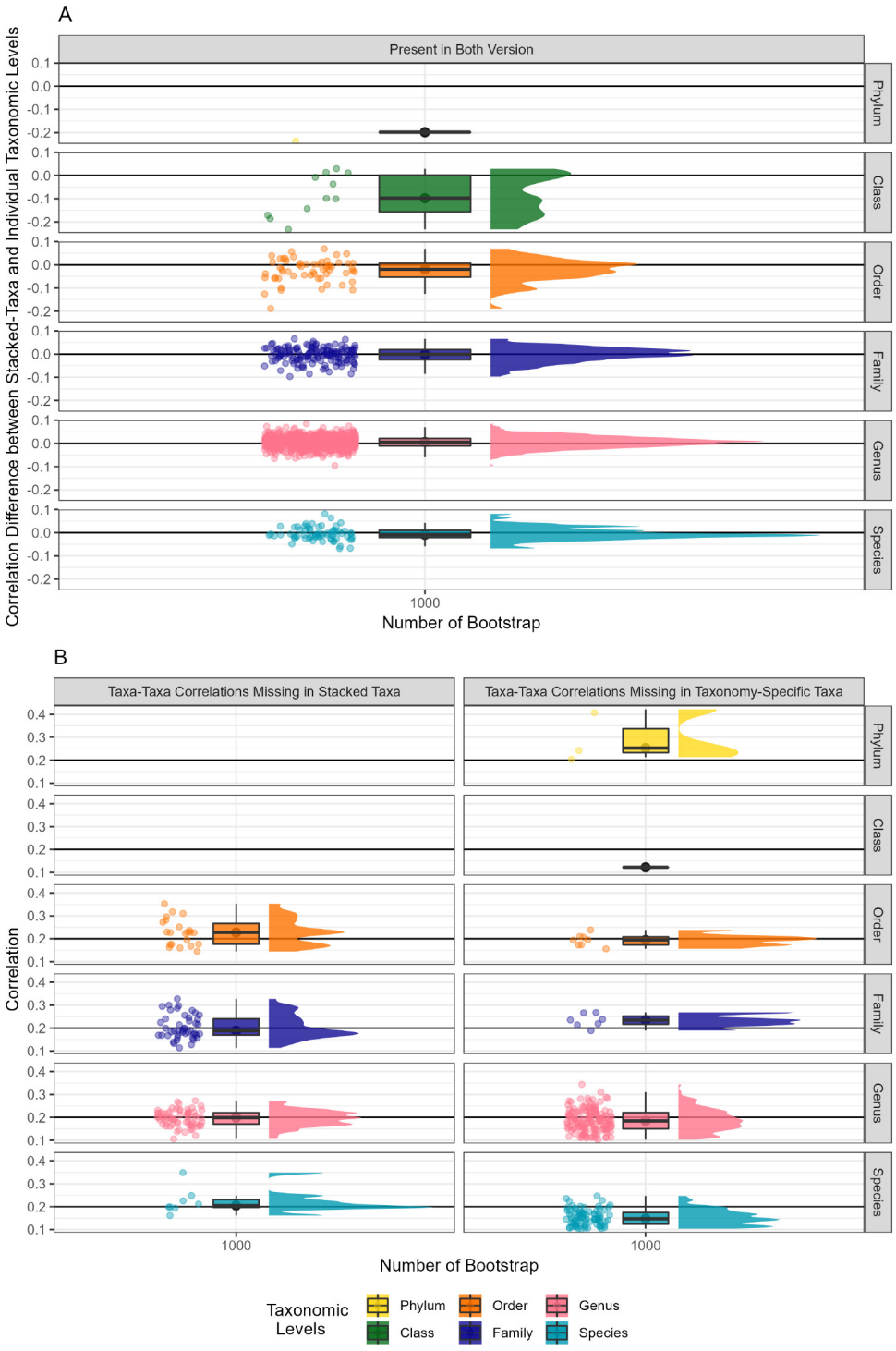
The patterns and silhouettes curves for the datasets are saved on Github (https://github.com/zhouLabNCSU/C3NA_ScriptsAndData) under the file Supplementary Figures.pdf with Fig. S11 – S26.

Consensus Matrices for the Datasets included in the study

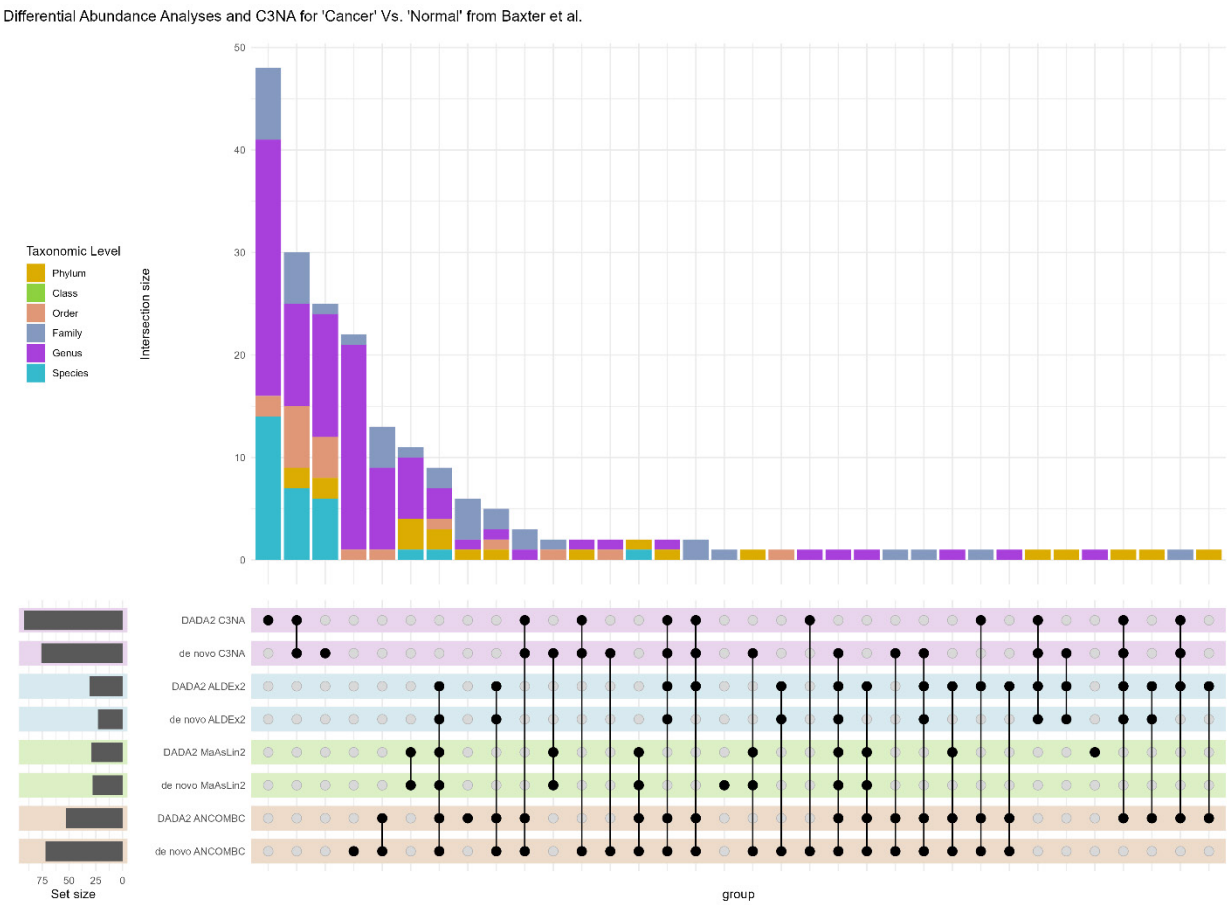
The consensus matrices based on the optimal number of clusters are saved on Github (https://github.com/zhouLabNCSU/C3NA_ScriptsAndData) under the file Supplementary Figures.pdf with Fig. S27 – S42.

Correlation Matrices for Datasets included in the study

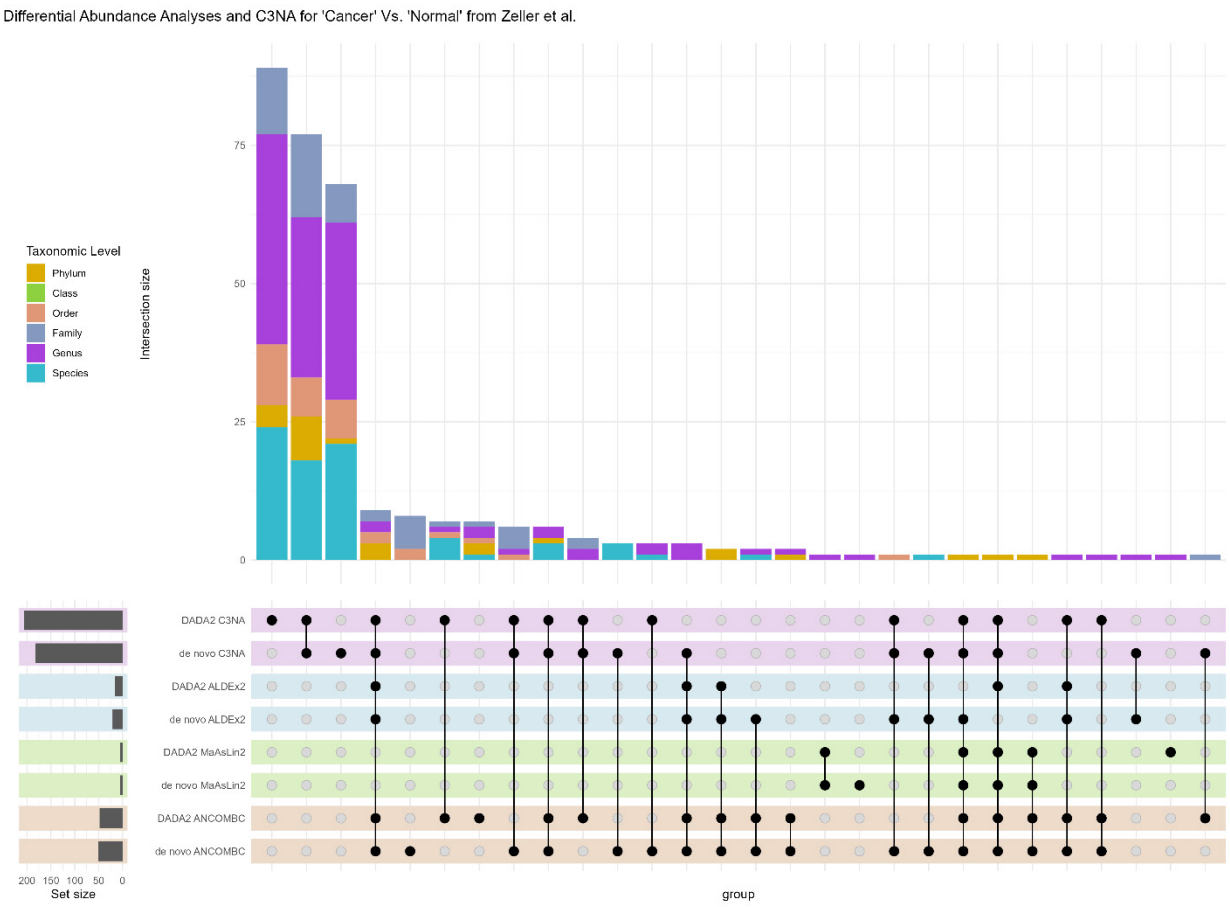
The correlation matrices based on the optimal number of clusters are saved on Github (https://github.com/zhouLabNCSU/C3NA_ScriptsAndData) under the file Supplementary Figures.pdf with Fig. S43 – S58.



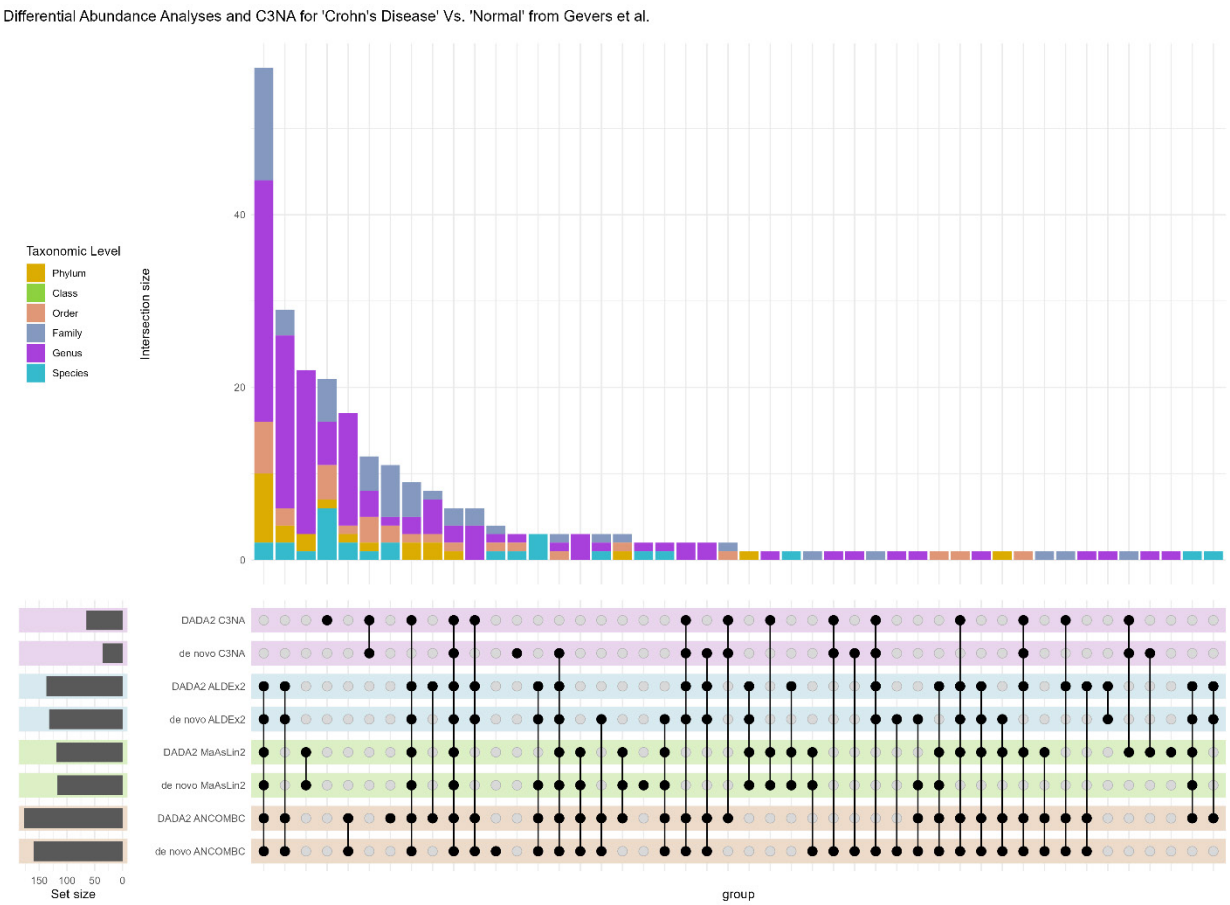
Supplement Figure 1. Comparison of Stacked-Taxa Correlation with Individual Taxonomic Correlation. The results used the Colorectal Cancer from Baxter et al., with the samples associated with the phenotype "Cancer." (**A**) Comparison of the difference between the stacked-taxa with the individual taxonomic correlations for the taxa-taxa pairs that are above 0.1 with adjusted p-values less or equal to 0.05. (**B**) Comparison of the stacked-taxa or individual taxonomic only correlations that do not present in the other results.



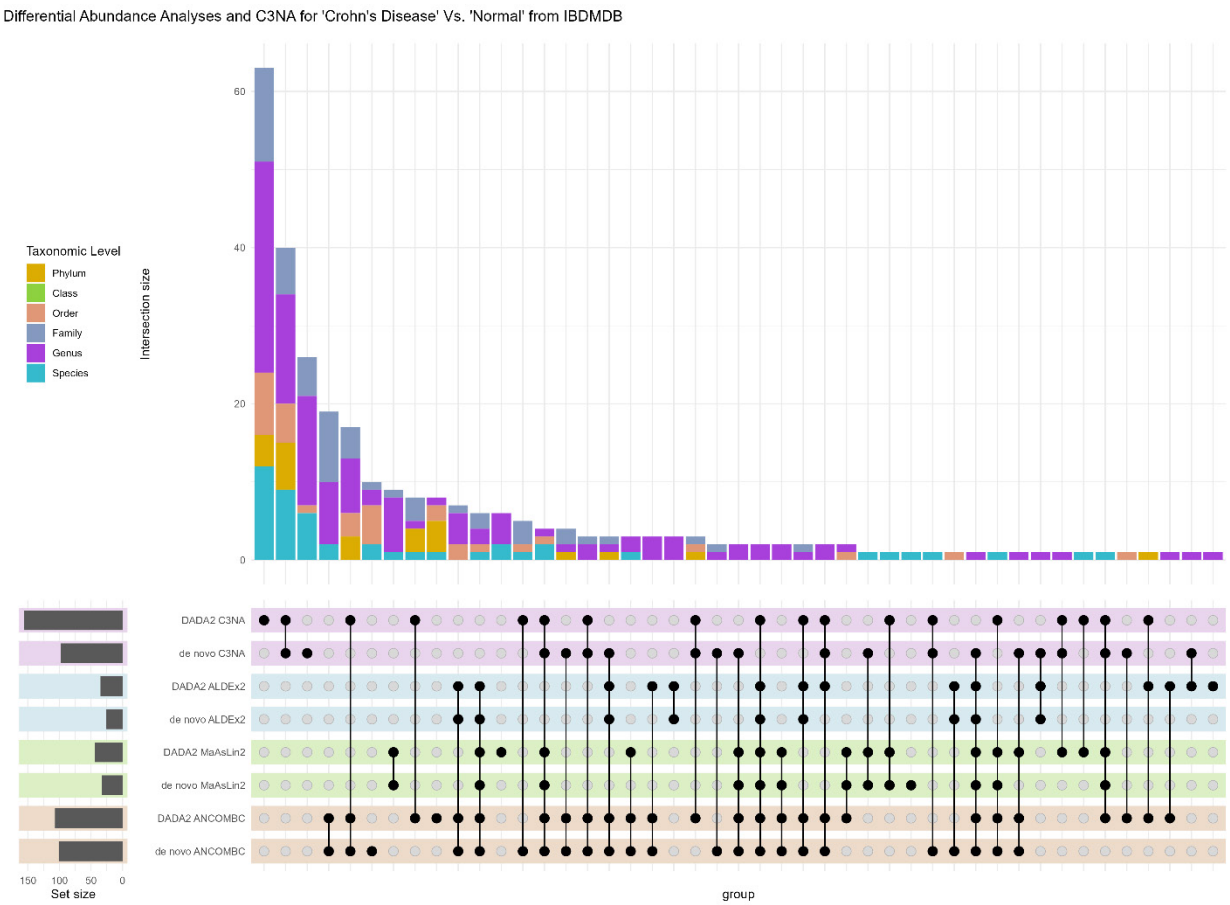
Supplement Figure 2. Compare differentially abundant taxa between the "Cancer" and "Normal" among two OTUs/ASVs assignment methods in Baxter et al. Orange filled intersect between the *de novo* and DADA2 methods under ANCOM-BC, green filled intersect between *de novo,* and DADA2 methods for under MaAsLin2, blue filled intersect between *de novo* and DADA2 methods for under ALDEx2, and purple filled intersect between *de novo* and DADA2 methods for under C3NA influential taxa.



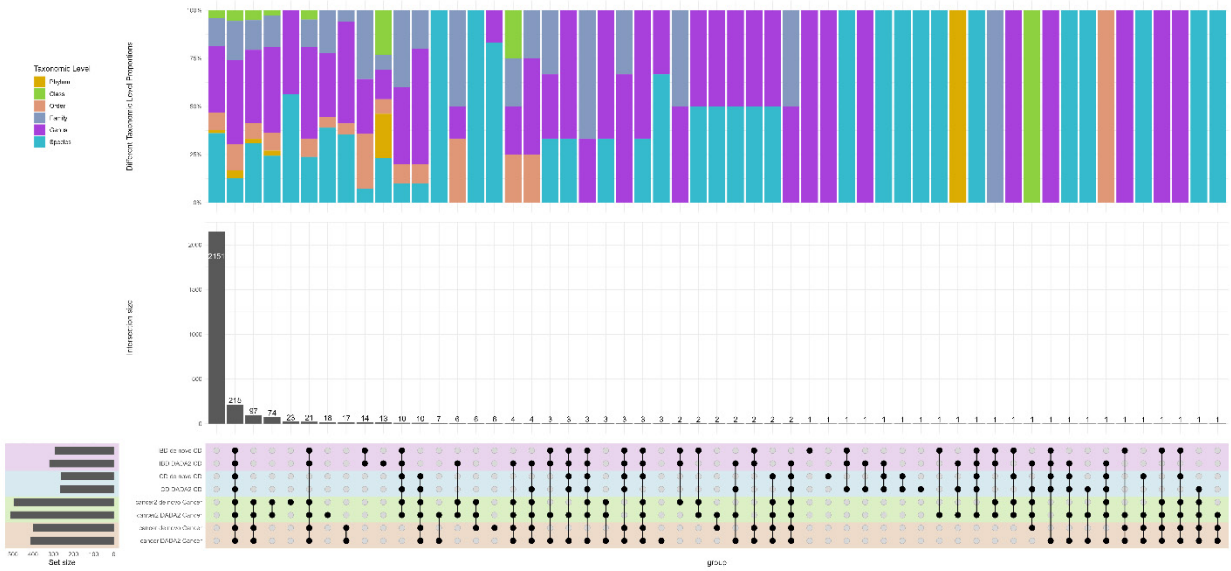
Supplement Figure 3. Compare differentially abundant taxa between the "Cancer" and "Normal" among two OTUs/ASVs assignment methods in Zeller et al. Orange filled intersect between the *de novo* and DADA2 methods under ANCOM-BC, green filled intersect between *de novo,* and DADA2 methods for under MaAsLin2, blue filled intersect between *de novo* and DADA2 methods for under ALDEx2, and purple filled intersect between *de novo* and DADA2 methods for under C3NA influential taxa.



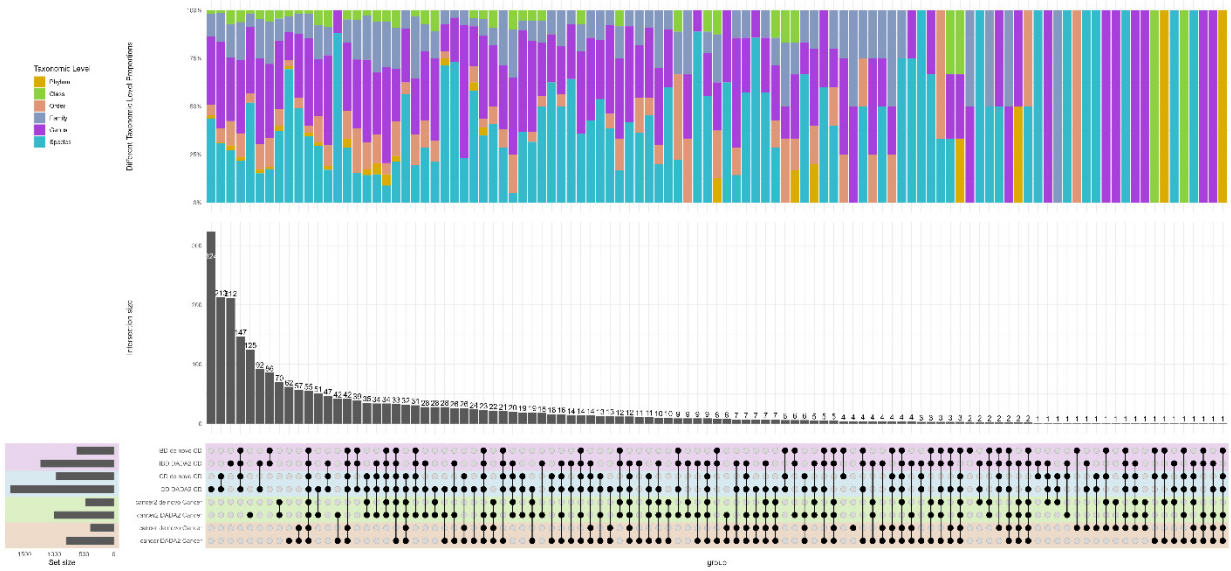
Supplement Figure 4. Compare differentially abundant taxa between the "Cancer" and "Normal" among two OTUs/ASVs assignment methods in Gevers et al. Orange filled intersect between the *de novo* and DADA2 methods under ANCOM-BC, green filled intersect between *de novo,* and DADA2 methods for under MaAsLin2, blue filled intersect between *de novo* and DADA2 methods for under ALDEx2, and purple filled intersect between *de novo* and DADA2 methods for under C3NA influential taxa.



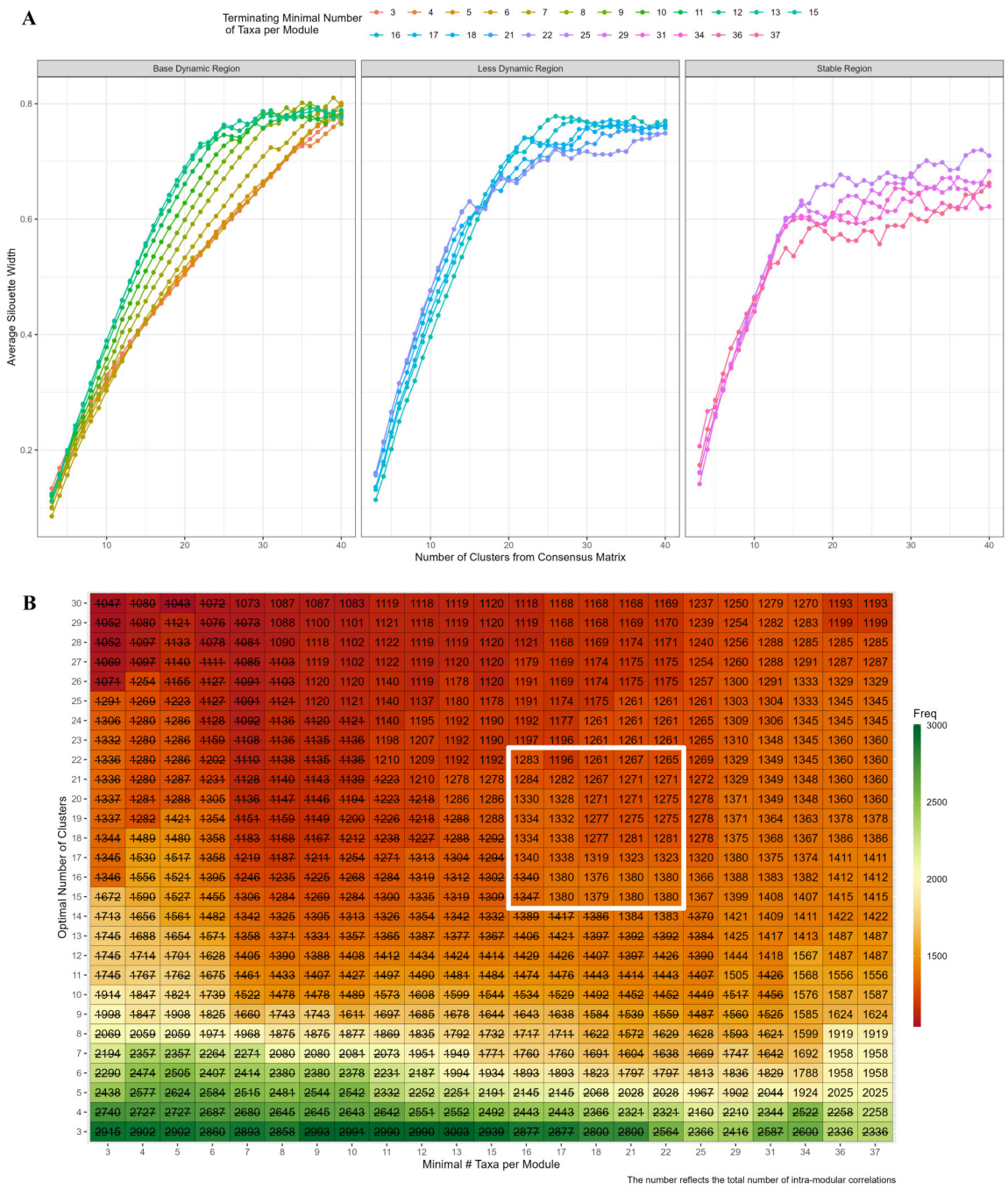
Supplement Figure 5. Compare differentially abundant taxa between the "Cancer" and "Normal" among two OTUs/ASVs assignment methods in IBDMDB. Orange filled intersect between the *de novo* and DADA2 methods under ANCOM-BC, green filled intersect between *de novo,* and DADA2 methods for under MaAsLin2, blue filled intersect between *de novo* and DADA2 methods for under ALDEx2, and purple filled intersect between *de novo* and DADA2 methods for under C3NA influential taxa.



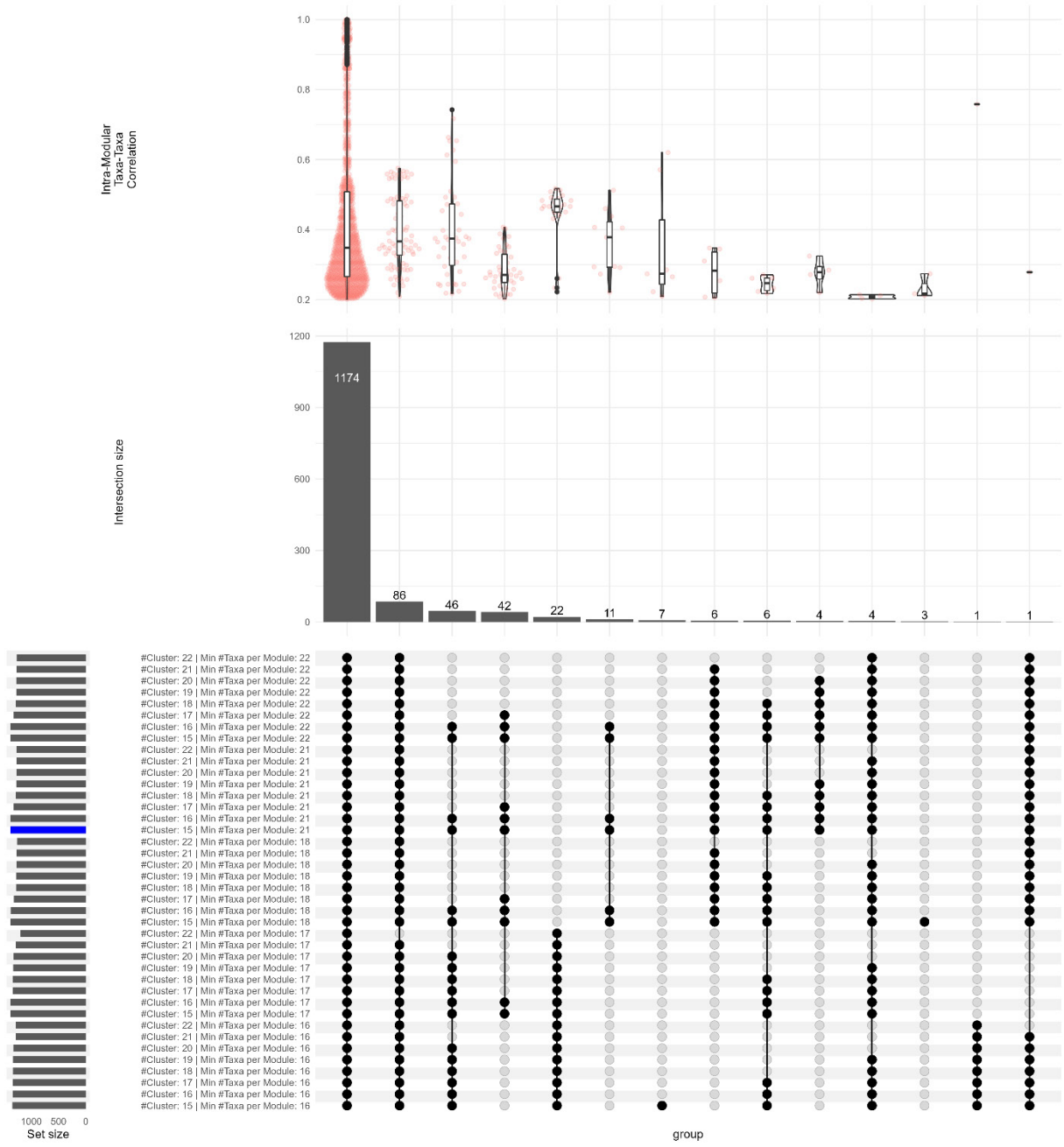
Supplement Figure 6. Impact of Filtering on Taxa Identification across Studies and OTUs/ASVs Assignment Methods from Disease Phenotypes. The top barplot highlights the proportion of each taxonomic level within the corresponding intersected taxa. The bottom bar plot highlights the intersected taxa with the solid point representing the presence of taxa in the corresponding row combination of study, clustering methods, and phenotype. Phenotypes: CD: Crohn's disease; Cancer: Colorectal Cancer.



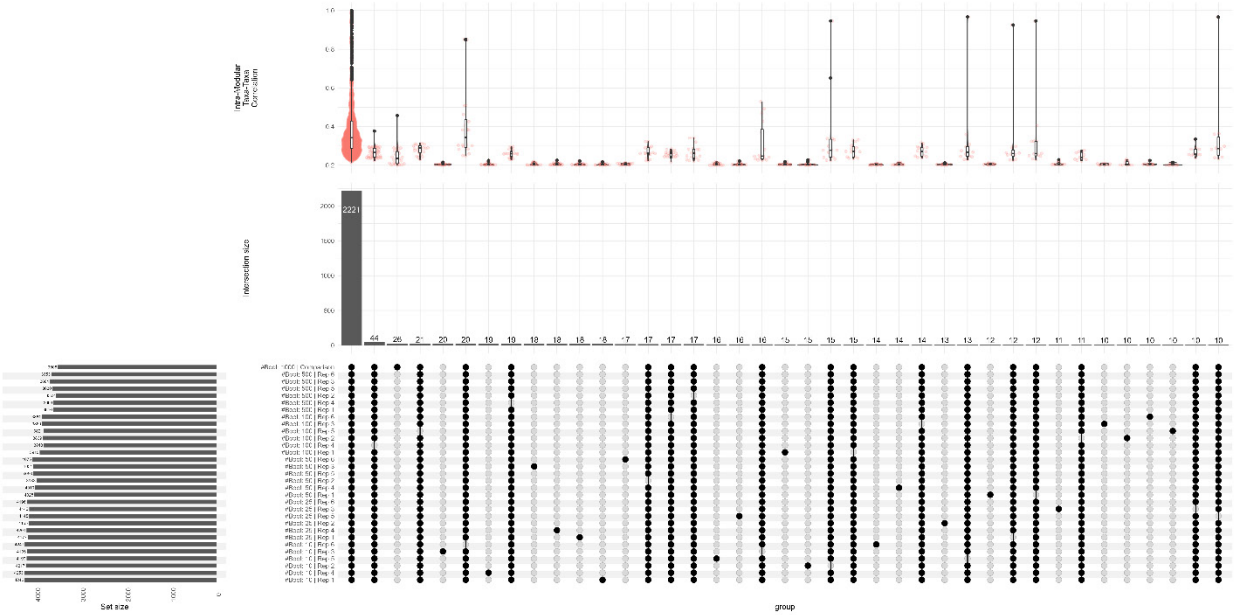
Supplement Figure 7. Removed Taxa from Six Levels across Studies and OTUs/ASVs Assignment Methods from Disease Phenotypes. The top barplot highlights the proportion of each taxonomic level within the corresponding intersected taxa. The bottom bar plot highlights the intersected taxa with the solid point representing the presence of taxa in the corresponding row combination of study, clustering methods, and phenotype. Phenotypes: CD: Crohn's disease; Cancer: Colorectal Cancer.



Supplement Figure 8. A. Average silhouette width plots with all Minimal Number of Taxa per Module and with Different Number of Clusters. **B.** Effects of Intra-modular Correlations from Different Minimal Number of Taxa per Module and Optimal Number of Clusters. The cross-out numbers represent the number of intra-modular correlations collected from clusters, including more than 10% intra-modular consensus proportions. The data is generated using the Baxter et al. with "Cancer" samples under the DADA2 ASV assignment method, and the minimal number of taxa per module chosen with ten modules is 21, and the optimal number of clusters chosen is 15. The corresponding consensus and correlation plots are Fig. S2 and S13, respectively. The white circle includes the ideal region for selection.



Supplementary Results Figure 9. Intra-modular taxa-taxa correlations are shared among the selected combination of Minimal Number Taxa per Module and Number of Clusters. The blue bar highlighted set represents the selected combination for C3NA analysis.



Supplementary Results Figure 10. Comparison of Significant Taxa-Taxa Correlations from Different Number of Bootstraps in sparCC. Each bootstrap undergoes six different replications. The number of bootstraps includes 10, 25, 50, 100, 500, and these are compared to the standard 1,000 iterations. Intersect taxa less than ten taxa are removed, the boxplot on the top represents the average correlation with each intersected taxa-taxa correlations among the different combinations.

									C3NA		Differential Abundance			
Dataset	**Clustering Methods**	**Phenotype**	**Number of Samples**	**Unfiltered Taxa**	**Number of Filtered Taxa**	**Taxa-Taxa Correlation**	**Taxa Remaining Percentage (%)**	**SparCC Computation Time (hours)**	**Number of Modules**	**Intra-Modular Taxa-Taxa Correlation**	**Influential Taxa**	**ANCOM-BC**	**ALDEx2**	**MaSaLin2**
Baxter et al.	DADA2	Cancer	127	954	414	3,885	43.40%	122	15	1,549	92	53	31	29
		Control	134	839	372	4,314	44.34%	96	20	1,014				
	de novo	Cancer	127	733	401	3,508	54.71%	107	19	1,361	76	72	23	28
		Control	134	636	355	3,779	55.82%	50	17	1,061				
Zeller et al.	DADA2	Cancer	41	921	512	3,219	55.59%	125	23	2,405	206	51	21	5
		Control	50	932	502	3,011	53.86%	125	15	2,580				
	de novo	Cancer	41	706	494	2,921	69.97%	66	21	2,306	182	48	16	5
		Control	50	706	469	2,711	66.43%	57	22	1,953				
Gevers et al.	DADA2	CD	731	1634	265	2,801	16.22%	47	13	1,045	65	177	137	119
		Control	335	1429	296	2,805	20.71%	104	22	839				
	de novo	CD	731	1089	261	2,390	23.97%	41	16	793	36	160	132	117
		Control	335	1020	287	2,407	28.14%	49	16	786				
IBDMDB	DADA2	CD	86	1148	320	4,383	27.87%	124	16	1,376	156	107	35	44
		Control	46	859	355	2,878	41.33%	115	15	1,742				
	de novo	CD	86	738	291	3,293	39.43%	36	18	930	98	101	26	33
		Control	46	590	318	2,381	53.90%	50	17	1,198				

Supplementary Table 1. Dataset summaries from the four studies, two phenotypes, and two OTUs/ASVs taxonomic assignment methods

Appendix C: Supplementary Materials for Chapter 4

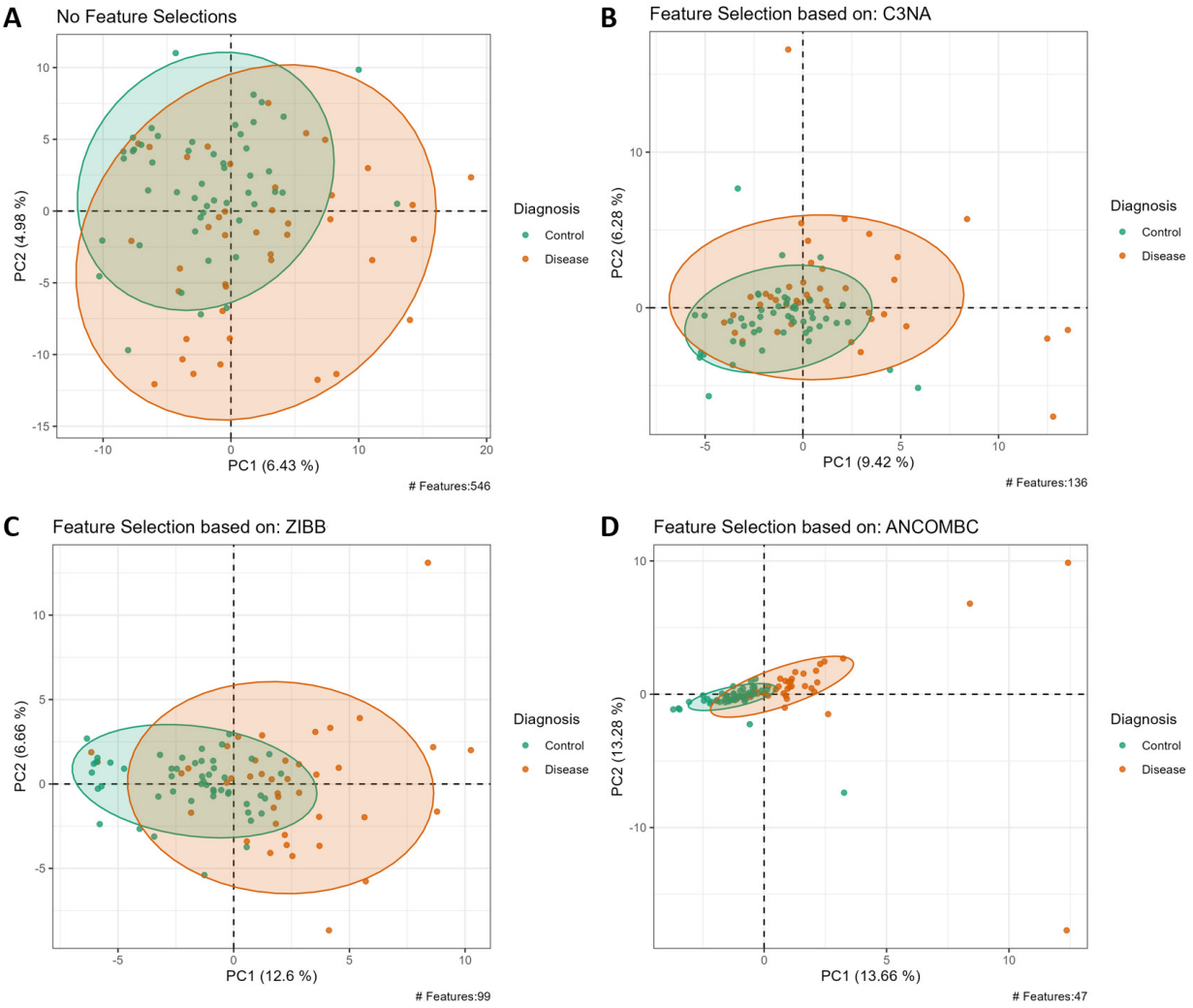


Figure S1. PCA analysis of the colorectal cancer and control from the Zeller et al. study from feature selected and non-feature selected approaches. Disease: colorectal cancer; Control: healthy controls. **A.** No feature selections. **B.** C3NA. **C.** ZIBB. **D.** ANCOM-BC

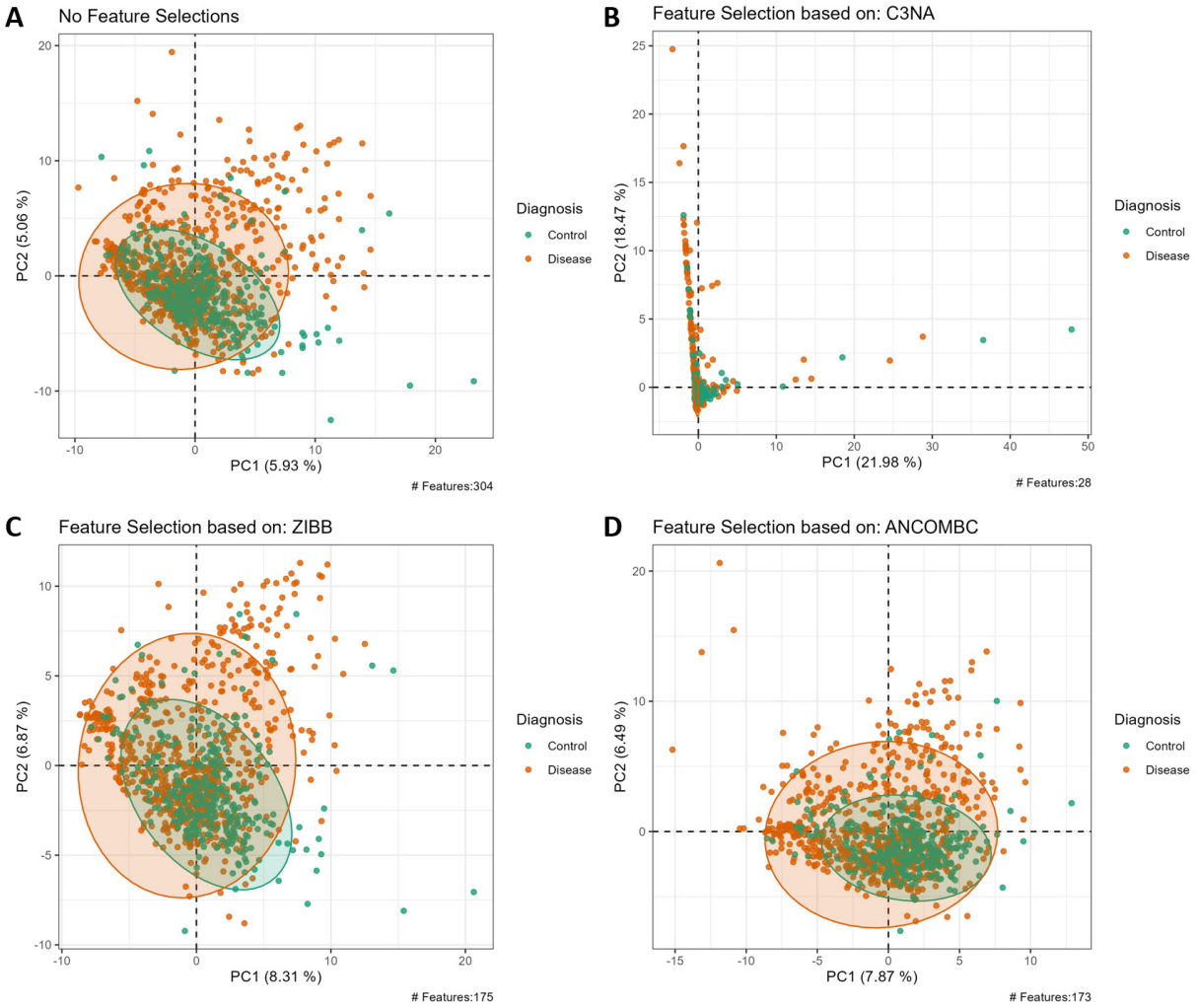


Figure S2. PCA analysis of the colorectal cancer and control from the Gevers et al. study from feature selected and non-feature selected approaches. Disease: Crohn's disease; Control: healthy controls. **A.** No feature selections. **B.** C3NA. **C.** ZIBB. **D.** ANCOM-BC

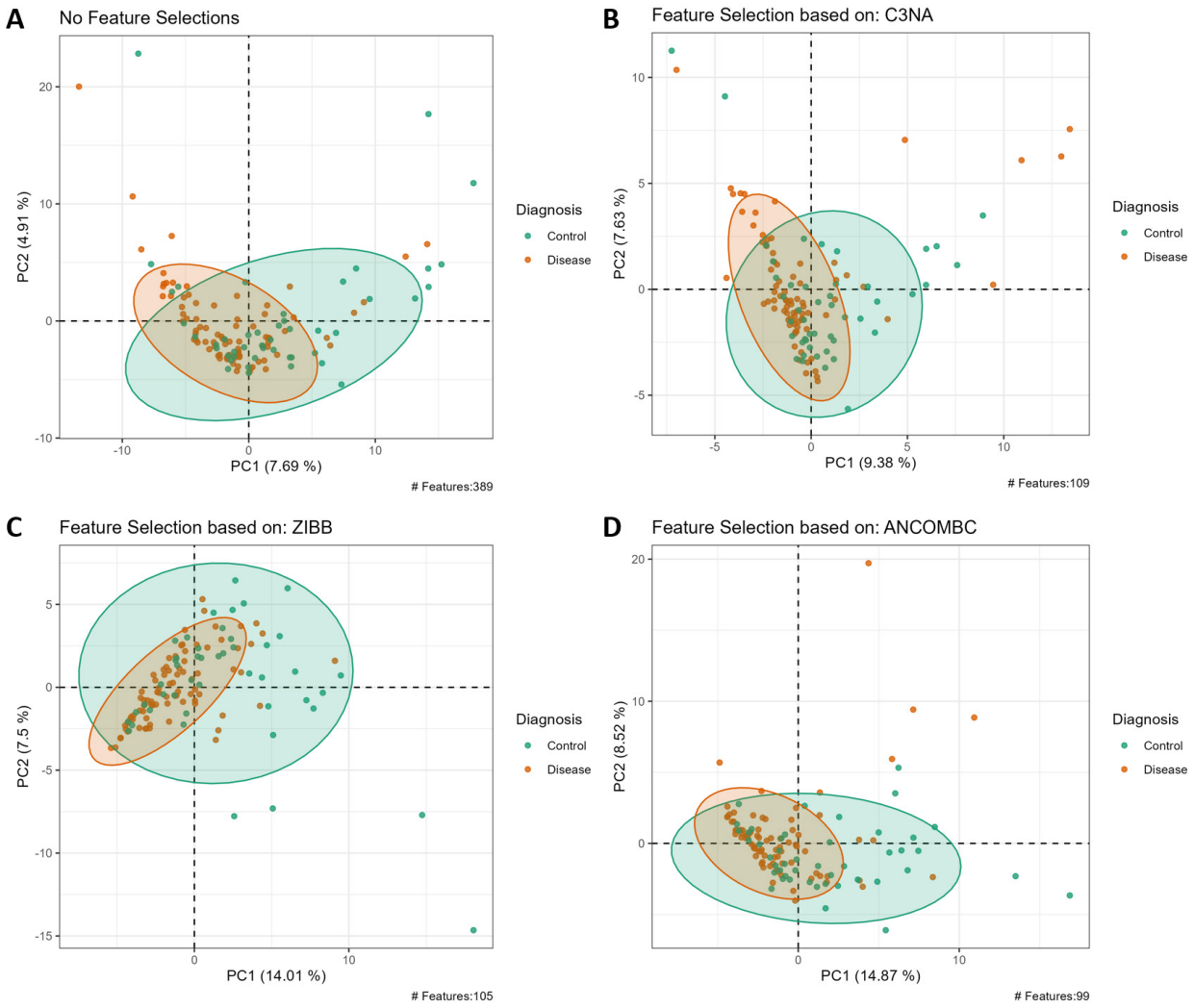


Figure S3. PCA analysis of the colorectal cancer and control from the IBDMDB from feature selected and non-feature selected approaches. Disease: Crohn's disease; Control: healthy controls. **A.** No feature selections. **B.** C3NA. **C.** ZIBB. **D.** ANCOM-BC

Figure S4. Feature selected taxa from different abundance taxa analyses and their corresponding AUROC across taxonomic levels. The x-axis no feature selection, with different combination of differential abundance methods. The y-axis represents the external data Area under the Curve (AUROC). The fill represents the Phylum, Class, Order, Family, Genus, Species, and All-taxa levels.