SUPERVISED LEARNING METHODS FOR ASSOCIATION DETECTION, BIOMARKER DISCOVERY, AND PATTERN RECOGNITION IN COMPOSITIONAL OMICS DATA

Andrew Lamont Hinton

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Curriculum of Bioinformatics and Computational Biology in the School of Medicine.

Chapel Hill
2022

Approved by:

Wesley Burks

Mike Kulis

Peter Mucha

Jeremy Purivs

William Valdar

ABSTRACT

Andrew Lamont Hinton: Supervised learning methods for association detection, biomarker discovery, and pattern recognition in compositional omics data
(Under the direction of Peter Mucha, Wesley Burks, and Mike Kulis)

Rapid advances and reduced cost in high throughput sequencing (HTS) technologies have enabled widespread profiling of microbial metagenomes and microbiomes in humans to better understand associations between microbial communities and disease. Data generated using these technologies are vast, high-dimensional, and nuanced, including limitations in instrument sequencing capacities and measurements that are inherently relative rather than absolute. Unlike absolute measurements, these relative counts — referred to as compositional data — require special methods for analysis and interpretation. Unfortunately, compositional data methodology are esoteric and generally not well adapted to high throughput sequencing data. Because of this, HTS data are often analyzed with traditional statistical methods that do not properly account for the underlying compositional sample space. This practice may result in spurious associations being reported which may limit study-to-study generalizations and reproducibility. In this thesis, building on existing literature in compositional data analysis and feature selection methodology, we develop a novel statistical association test and a powerful machine learning framework using robust pairwise logratios. Additionally, for each method, we developed freely available (GitHub) R packages (SelEnergyPermR & DiCoVarML) with functions to perform the core analysis of each method. In the first chapter we provide a basic overview of compositional data and its connection to HTS data. In the second chapter, we present the SelEnergyPerm method for detecting sparse associations in high dimensional metagenomic data. In the third chapter, building on the concept of differential compositional variation proposed in SelEnergyPerm, we present the DiCoVarML framework for supervised classification and biomarker discovery. In the final chapter, we apply the SelEnergyPerm method to test for an association between toxicant exposures and the composition of microbial communities in the nasal passage. Using a parsimonious logratio signature detected by SelEnergyPerm, we then perform integrative analysis, where we explore the connection between nasal microbiome dsybiosis and immune mediator expression in nasal lavage fluid.

I dedicate this dissertation to my grandparents *Ernestine* and *Jeremiah Hinton* who raised and groomed me to be the man I have become today. I also dedicate this work to my daughters *Gabrielle Leah* and *Madison Lauren Hinton*. May they know anything is possible if they put their mind to it. Finally, I dedicate this thesis to my wife *Tanisha Hinton* who stuck by me during my transition from career to graduate to career again. Without her none of this would be possible.

ACKNOWLEDGEMENTS

As I endeavored along the difficult path to become a scientist I encountered many people who were instrumental in helping ensure this journey was successful. First, I would like to thank Wesley Burks who answered my unsolicited email sent inquiring about food allergy research. This fortuitous interaction along with his mentorship helped shape my trajectory and success as a graduate student. I would also like to thank the Burks and Kulis Lab for their support and encouragement. Specifically, I would like to thank Kelly Orgel, Johanna Smeekins, Jada Suber, Lakeya Hardy, Erin Steinbach, and Suzanne Barshow for helping refine my ability to think scientifically in the basic science setting.I would also like to thank Edwin Kim for his support with career exploration, equipment, and mentorship throughout this experience.

Next, I would like to thank members of the Mucha lab that were instrumental in helping me refine my computational and mathematical thinking ability as it pertains to a range of biomedical and data science problems. In particular, I would like to thank Natalie Stanley, William Weir, Zachary Boyd, Austin Ferguson, Scott Emmons, and Ryan Gibson for many conversations that help shape the DiCoVarML and SelEnergyPerm methods. Additionally, I would like to thank Elise Hickman and Illona Jaspers for their collaboration which greatly influenced the development of the SelEnergyPerm project. Further, I would like to thank Laura Marks, John Robinson, and Jeff Henderson for their collaboration and support which greatly molded the final iteration of the DiCoVarML project. I would also like to thank the Howard Hughes Medical Institute for providing me with funding to pursue my research interest via the Gilliam Fellowship for advanced study.

Finally, I would like to thank my committee for their continued support throughout my graduate studies. I would like to thanks Mike Kulis for his guidance within the wet lab study and food allergy focused part of my research. I would like to thank Peter Mucha for his guidance along the data science and networks focused aspects of my research. I would also like to thank Will Valdar and Jeremy Purvis for their guidance along both the academic and research aspects of my graduate studies. In conclusion, I would like to thank Jon Cornett for his support with the important administrative aspects of being a professional graduate student.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

16S	16S ribosomal rRNA amplicon
ANOSIM	Analysis of Similarity
ALR	Additive Logratio
AUC	Area Under the Receiver Operating Characteristic Curve
BH	Benjamini-Hochberg
*c*F	Combined-F
CLR	Centered Logratio
CSF	Cerebral spinal fluid
DCV	Differential Compositional Variation
DiCoVArML	Differential Compositional Variation Machine Learning
FA	Food Allergy
IBD	Inflammatory bowel disease
LASSO	Least Absolute Shrinkage and Selection Operator
MCC	Matthews Correlation Coefficient
NPIH	Non-Post Infectious Hydrocephalus
OTU	Operational Taxonomic Unit
PERMANOVA	Permutational multivariate analysis of variance
PERMDISP2	Distance-Based Tests for Homogeneity of Multivariate Dispersions
PIH	Post Infectious Hydrocephalus
PLR	Pairwise Log Ratio
PLS-DA	Partial Least Squares Discriminate Analysis
RF	Random Forest
RFE	Random Forest Recursive Feature Elimination
SelEnergyPerm	Selection-Energy-Permutation
tarMFS	Targeted Multilevel Feature Selection
WGS	Whole-Genome Shotgun
ZINB	Zero Inflated Negative Binomial

CHAPTER 1: INTRODUCTION TO COMPOSITIONAL AND OMICS DATA

In this chapter we overview basic compositional data concepts important to methods developed in this work. Notably, this section is not meant to be an exhaustive review of compositional data literature. In the next section, we explain how high throughput sequencing derived omics data are a special case of compositional data. Finally, we briefly overview computational and statistical challenges associated with the application of standard compositional analysis techniques to omics data.

1.1 Compositional Data

1.1.1 Introduction

Relative data can take many forms including proportions (%, ppm, etc.), concentrations(μg/mL, mol/L, etc.), or data where the total sum (mass, number of reads, etc.) between samples are uninformative. For example, in geosciences, the chemical composition of rocks oxides can be measured using X-Ray Fluorescence, where variation in instrument sensitivity, specimen physical properties, and elemental compositions create uninformative scale difference between samples [Nakayama and Nakamura, 2014]. To account for this, the chemical composition of rock oxides are reported relative rather than independent of one another (i.e proportions). More formally, relative data are known as compositional data and can be defined as a vector (composition) $\boldsymbol{x}$ whose D-elements (parts) are strictly positive with a unit-sum constraint

$$x_1 + \ldots + x_D = 1. \tag{1.1}$$

While seemingly innocuous, the unit-sum constraint imposes significant limitations on the statistical modeling of compositonal data. The dangers of spurious correlations when analyzing correlations between parts of a composition were first noted by Karl Pearson in 1897 [Pearson, 1897]. While these difficulties were generally known, robust statistical methodology to analyze compositonal data were formally introduced in the 1982 paper titled *The Statistical Analysis of Compositional Data* by John Aitchinson [Aitchison, 1982]. Importantly, the analysis of compositional data are limited to relative rather than absolute interpretations. That is, single part interpretations are not justified in this context. For example, given a set of D-part compositions measured between two groups, one may be interested in making the following statement: “x_D is more abundant in group 1 than in group 2”. Unfortunately, in this setting, even after applying traditional statistical techniques (e.g. two sample t-test) “absolute” statements of this type can be spurious given the underlying compositional data are “relative”. While compositional data are rooted in rigorous statistical and geometric principles, the approaches presented in [Aitchison, 1986] have been subject, throughout history, to fierce resistance. While it’s beyond the scope of this section to extensively detail the historical dialogue, we refer the reader to here [Aitchison] for a general overview of types of opposition and confusion. Further, we refer the reader to [Scealy and Welsh, 2014] for a critical review of compositional data analysis methods. Notwithstanding, compositional data analysis methodology have been successfully used in many fields [Pawlowsky-Glahn and Buccianti, 2011], widely recognized to be important for the analysis of microbiome datasets [Gloor et al., 2017], and recently shown to be valuable when used to analyze various omics datasets [Greenacre et al., 2021].

1.1.2 Simplex Sample Space

Here we introduce the simplex sample space for modeling compositional data introduced in [Aitchison, 1982]. Importantly, the unit-sum constraint defined in Eq. 1.3 restricts the sample space to a simplex within real space. For example, using samples (n=23) from the 3-part chemical composition of aphyric Skye lavas[Aitchison, 1986], the unit-sum constraints can be visualized by plotting relative data in 3-dimensional coordinates (Figure 1.1). As mentioned above, the parts of $\boldsymbol{x}$ must be positive

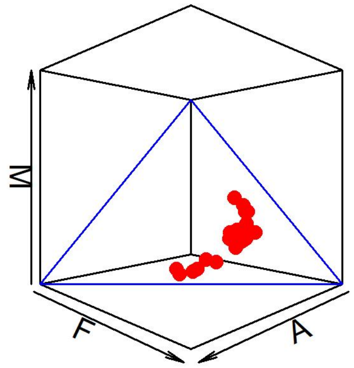

Figure 1.1: Visualization of a 2-Simplex (3-part Composition) embedded in 3-D real space. Here we show how the chemical compositions (n=23) from aphyric skye lavas (red points) are restricted to a simplex subspace (bordered by blue lines) in real space. The (x,y,z) axes (real space) represent the unit-sum normalized (A,F,M)-parts. Arrows show the proportions of parts and point from 0 to 1.

$$x_1 \geq 0, \ldots, x_D \geq 0. \tag{1.2}$$

Notably, only $d = D - 1$ parts are required to fully define a D part composition with a unit-sum constraint. This is true such that the value of the Dth part can be calculated by $x_D = 1 - \sum_{i=1}^{D-1} x_i$ From this the (d)-simplex space embedded in real space can be defined as

$$\mathcal{L}^d = \{x \in \mathbb{R}^D : \sum_{i=1}^{D} x_i = 1, x_1 \geq 0, \ldots, x_D \geq 0\}. \tag{1.3}$$

Given this, a 3-part composition which is a 2-simplex(triangle) can be easily visualized using barycentric ternary diagrams (Figure 1.2).

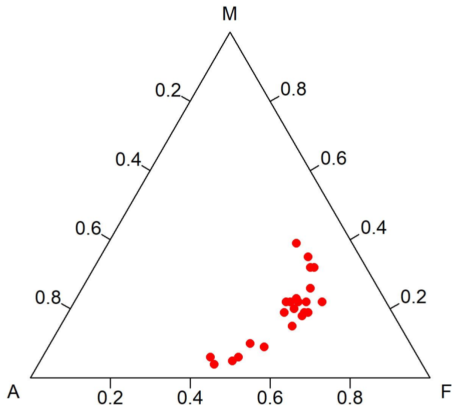

Figure 1.2: Ternary diagram of the 2-Simplex showing barycentric coordinate system. The chemical compositions (n=23) from aphyric skye lavas [Aitchison, 1986] (red points) are shown . Faces (triangle edges) of the simplex measure the proportion of each part where the intersection represents the overall composition.

1.1.3 Key Compositional Data Analysis Principles

A key admission when starting any compositional analysis is in stating the absolute sum of the composition being measured is uninformative. When this prerequisite is met, the samples are restricted to the simplex realm of the Euclidean real space where compositional data analysis is required. Based on [Aitchison, 1982] and important for work developed in this thesis, fundamental requirements of compositional data analysis must meet the scale invariance and subcompositional coherence principles.

1.1.3.1 Scale invariance

Scale invariance from a practical standpoint is such that the compositional data analysis methodology must yield the same conclusion independent of scale. That is, from a compositional standpoint, two compositions $\boldsymbol{j}$ and $\boldsymbol{k}$ are equivalent if there exist a constant p where $\boldsymbol{k} = p\boldsymbol{j}$ where $p > 0$. Functionally, the principle of scale invariance can be written such that a function f is scale invariant if $f(\boldsymbol{j}) = f(\boldsymbol{k})$. For compositional data this property is most conveniently met when $\boldsymbol{j}$ and $\boldsymbol{k}$ are expressed in terms of ratios between components.

1.1.3.2 Subcompositional Coherence

A subcomposition can be defined as any C-part composition formed from a full D-part composition. Notationally, we can define the subcomposition $\boldsymbol{s}$ from $\boldsymbol{x}$ (Eqn. 1.3) as

$$s_i = \frac{x_i}{\sum_{j=1}^{C} x_j} \text{ for } i = 1, \dots, C \tag{1.4}$$

While subcompositional analysis enables the independent study of a subset of parts from the full composition, it does not come without the risk of spurious interpretations. The principle of subcomposition requires, for subcompositons with shared parts, compositional data functions to yield the same conclusions independent of parts included in the subcompositon that are not shared. This is most easily demonstrated using a toy example where we use the 5-part (A,B,C,D,E) *Hongite* rock mineral composition dataset (n=25) described in [Aitchison, 1986] and accessed using the compositions R package. To demonstrate the importance of subcompositonal coherence, we examine correlation patterns among raw proportions for three subcompositions that share two parts (D,E). As observed in Figure 1.3, researchers who analyze the (A,D,E) subcomposition would infer a negative correlation between D-E while researchers who analyzed the (B,D,E) and (C,D,E) subcompositions would infer different but overall positive correlations. Surely, D-E cannot be both positively and negatively correlated simultaneously. To reconcile this incoherence, analysis of ratios between parts can be carried out. Here we are interested in the behavior of the ratio formed by D/E. As seen from the distribution of ratio values of D/E across subcompositions in Figure 1.4, ratios values are preserved across subcomposition. While the correlation in this context is no longer useful for measuring dependence between D-E the behavior of the D/E ratio can be studied (not discussed here). As can be seen from the D/E ratio coherence across subcompositon, compositional data analysis with ratios are subcompositional coherent.

1.1.4 The logratio transformation

In general, there are two popular geometric approaches to analyzing compostional data: (1) within-simplex via untransformed proportions, (2) out-of-simplex via logratio transformation. In this thesis we focus primarily on the out-of-simplex approach to compositional data analysis. Here we will describe

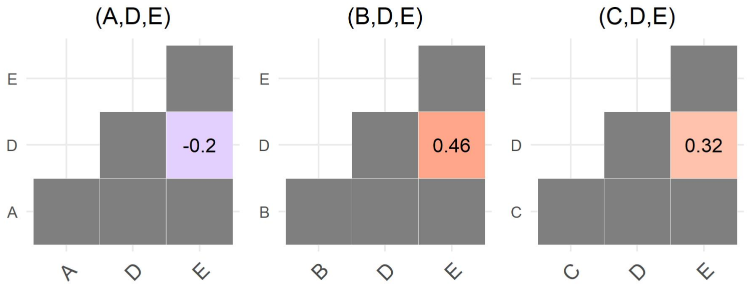

Figure 1.3: Toy example demonstrating subcompositonal incoherence when computing the product-moment correlations on raw compositions data Heatmap showing the product-moment correlation for three 3-part subcompositions sharing two parts (D,E). Data (n=25) for the full 5-part composition are from the mineral compositions of Hongite rock specimens [Aitchison, 1986].

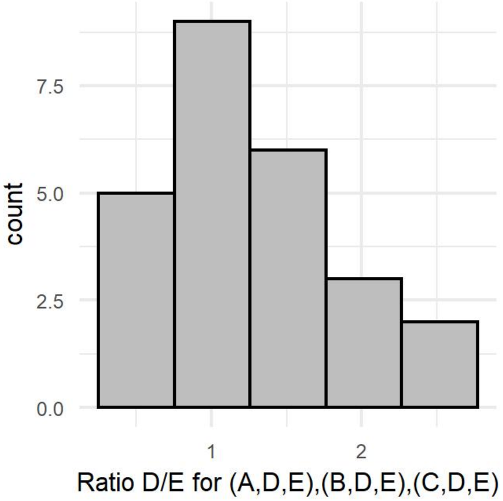

Figure 1.4: Demonstration of the principle of subcompositional coherence of ratio analysis Histogram of the ratio D/E across three 3-part subcompositions

the three important logratio transformations used in this thesis. Notably, the shift here from ratios to logratios is rooted in the fact that logarithms of ratios are antisymmetric such that the $\log a/b = C$ and the $\log b/a = -C$ leading to similar variance for logratios when $a > b$ or $a < b$.

1.1.4.1 Additive logratio transformation

One of the earliest logratio transformations proposed by John Aitchinson was the additive logratio (ALR) transformation to map $\mathcal{L}^d \to \mathbb{R}^d$. For the D-part compositional vector $\boldsymbol{x}$ (Eqn. 1.3) we define the ALR transformation with reference part x_D

$$\boldsymbol{z} = \left[\log \frac{x_1}{x_D}, \ldots, \log \frac{x_d}{x_D}\right]. \tag{1.5}$$

Importantly, the ALR satisfies both principles of compositonal data analysis and yield the same conclusion independent of which part is selected as the reference. Benefits of the ALR are in its simple interpretation and relative ease of computation given its $D - 1$ dimensionality. In contrast, the ALR is asymmetric in parts (different set of logratio depending on which reference is selected) and has geometric limitations. In [Egozcue et al., 2003], the ALR is described as non-isometric (inter-sample distance preserving) in the mapping from the simplex to real space. While these limitations exist, recent work in [Greenacre et al., 2021, Greenacre, 2019] suggested that with appropriate selection of a reference part the isometric limitation, albeit theoretical, may not be as severe in practice.

1.1.4.2 Centered logratio transformation

To address the asymmetry in parts, the centered logratio (CLR) transformation was proposed. The CLR, which maps compositonal data from the simplex to a D-dimensional hyper plane in real space can be described as

$$\boldsymbol{y} = \left[\log \frac{x_1}{\left(\prod \boldsymbol{x}\right)^{1/D}}, \ldots, \log \frac{x_D}{\left(\prod \boldsymbol{x}\right)^{1/D}}\right]. \tag{1.6}$$

Unlike the ALR, the CLR is isometric but a major drawback arises from its singular covariance matrix. This is a direct result of the denominator term in Eqn 1.8 which is the geometric mean of the full composition and thus is dependent on information from all parts of the composition.

1.1.4.3 Pairwise logratio transformation

The final transformation described here is the pairwise logratio (PLR) transformation. The PLR maps compositional data from the d-simplex to the $D(D - 1)/2$-dimensional real space. Letting

$p = D(D-1)/2$ we first define the PLR matrix for the compositional vector $\boldsymbol{x}$ as $\mathbf{P} \in \mathbb{R}^{D \times D}$ described as

$$\mathbf{P} = f(\boldsymbol{x}) = \begin{bmatrix} \log \frac{x_1}{x_1} & \cdots & \log \frac{x_1}{x_D} \\ \vdots & \ddots & \vdots \\ \log \frac{x_D}{x_1} & \cdots & \log \frac{x_p}{x_D} \end{bmatrix}. \tag{1.7}$$

Note, $\mathbf{P}$ is antisymmetric in the value of the logratio thus requiring only the lower or upper off-diagonals be computed to define the full frame of PLRs. From this, the PLR vector of $\boldsymbol{x}$ becomes

$$\text{PLR}(\boldsymbol{x}) = [p_{ij}, \ldots, p_{ij}] \text{ for all } (i = 2, \cdots, D); (j = 1, \cdots, D) \text{ where } i > j \tag{1.8}$$

The PLR is isometric where the primary limitations are computational given the quadratic ($D(D-1)/2$) increase of PLR dimensions as the number of parts increase. Additionally, for high dimensional compositonal data pre-transformation ($D > 10,000$ common in genomics data) the PLR transformation may be infeasible to compute without specialized computing clusters. Further high-dimensional PLR data may be particularly problematic for datasets with limited samples such that $p \gg n$.

1.2 Relative Nature of High Throughput Sequencing Data

In this work we focus on the analysis of high throughput sequencing (HTS) data with a specific focus on microbiome and metagenomic data generated with this technology. HTS technologies, in general, sequence a subset of extracted DNA fragments on an instrument yielding grouped read counts as parts. Depending on the experimental design and sequencing method (e.g. Metagenomic, 16S rRNA gene amplification analysis), parts, in this context, can be operational taxonomic units (16S) or microbes (Metagenomic). Importantly, HTS data are count data where the total counts for all parts observed for each sample is uninformative. This is due to finite upper bounds in the number of reads obtained by the sequencing instrument [Gloor and Reid, 2016, Quinn et al., 2019, Fernandes et al., 2014, Mandal et al., 2015a]. As described above, here the admission of uninformative total counts have profound consequences on the analysis of these data. Unlike traditional compositional data analysis described in [Aitchison, 1982, Greenacre, 2019, Pawlowsky-Glahn and Buccianti, 2011], HTS data are sampled count data (counts are random samples from true population), high-dimensional (in number of parts identified), and highly sparse (contain high proportions of zeroes). Indeed, one cannot compute logratios in the presence of zeros presenting at first glance an insurmountable challenge to the staple logratio transformation. Fortunately, various zero replacement strategies have been studied [Martín-Fernández et al., 2015, Martın-Fernandez et al., 2003] and can generally be implemented to overcome the problem of zeroes in compositional HTS data. We note however, despite this progress, zero imputation is still an open area of research. As will be detailed in the Chapters 2 and 3 of this work there has been significant progress made in adapting compositional data analysis principles to the exceptional HTS compositional datasets.

1.3 Overview of Thesis

With the required compositional data and HTS data perspective presented in this chapter we now briefly describe the content presented in this dissertation.

1.3.1 SelEnergyPerm

In Chapter 2 building on well-established compositional data analysis methods we develop a statistical association test for HTS compositional data. In particular, we describe our non-parametric group association test for metagenomic data that is designed to detect sparse association signals in settings where existing tests have reduced power. This method for multivariate analysis of metagenomic data performs feature selection on the set of all pairwise logratios, which are resistant to technical variation, to obtain signatures of association between groups or phenotypes of interest that are then directly interpretable in terms of relationships in the microbiome. The method as presented here has been implemented in a freely available R package available on GitHub.

The work presented in Chapter 2 has been adapted from our manuscript published here: Hinton AL, Mucha PJ. A Simultaneous Feature Selection and Compositional Association Test for Detecting Sparse Associations in High-Dimensional Metagenomic Data. Front Microbiol. 2022 Mar 21;13:837396. doi: 10.3389/fmicb.2022.837396. PMID: 35387076; PMCID: PMC8978828.

1.3.2 DiCoVarML

In Chapter 3, we develop a powerful machine learning framework for using metagenomic data to predict two or more phenotypes of interest. In particular, this framework enables researchers to target and identify predictive combinations of microbial biomarkers for various study objectives and is demonstrated through the two case studies. This is the first such method built directly in terms of pairwise logratios as the features in a machine learning framework with feature selection. In particular, this approach selects a predictive subset from the full frame of redundant pairwise logratios to be used for supervised classification of phenotypes of interest. Because of this, the resulting predictive signatures are inherently independent of overall signal magnitude, making them particularly appropriate for out-of-sample prediction between different studies, and the features selected tend to be more naturally interpretable compared to other methodologies. The method and key functionality presented here has been implemented in a freely available R package available on GitHub.

The work presented in this Chapter has been submitted for publication and is available as a preprint at: Hinton, A. L. & Mucha, P. J. Differential Compositional Variation Feature Selection: A Machine Learning Framework with logratios for Compositional Metagenomic Data. bioRxiv 2021.12.08.471758 (2021) doi:10.1101/2021.12.08.471758.

1.3.3 SelEnergyPerm Nasal Microbiome Analysis

In Chapter 4, the final chapter, we apply the SelEneryPerm method to detect and discover logratio signatures capable of explaining compositional differences in the nasal microbiome between healthy, e-cigarette, or cigarettes exposure groups. Using SelEnergyPerm we detected compositional difference in the nasal microbiomes of males and females. After controlling for differences in sex, we identify a sparse pairwise logratio signature capable of discriminating between exposure groups and integrate it with immune mediator data. This integrative analysis revealed an association between nasal microbiome dysbisos and immune mediator expression changes.

The work presented in this Chapter has been submitted for publication and is available as a preprint at: Elise Hickman*, Andrew Hinton*, Bryan Zorn et al. E-Cigarette Use, Cigarette Use, and Sex Modify the Nasal Microbiome and Nasal Host-Microbiota Interactions, 03 August 2021, PREPRINT (Version 2) available at Research Square [https://doi.org/10.21203/rs.3.rs-725763/v2]. (*indicates that these authors contributed equally to the work)

CHAPTER 2: SELECTION-ENERGY-PERMUTATION[1]

2.4 Introduction

Metagenomic studies have enabled unprecedented insight into connections between microbes, their functions, and human disease [Martín et al., 2014]. These insights are a direct result of rapid advances in next-generation sequencing technologies which are critical to metagenomic studies. Specifically, these technologies are leveraged in two popular approaches: 16S ribosomal rRNA amplicon (16S) and whole-genome shotgun (WGS) sequencing [Ranjan et al., 2016]. Application of these approaches are widespread and have been used to study associations between the gut microbiome composition and colorectal cancer [Gopalakrishnan et al., 2018], inflammatory bowel disease, obesity [Manichanh et al., 2012], cirrhosis [Qin et al., 2014], and anxiety/depression [Foster and McVey Neufeld, 2013] in humans via the gut-brain axis, to name a few. The skin [Kong et al., 2012], oral [Dewhirst et al., 2010], and nasal microbiomes [Wilson and Hamilos, 2014] among other sites have also been studied in connection to disease onset and progression. With an increasing number of putative associations between microbial communities from various sites of the human body and disease being reported, microbial compositions are now being explored as diagnostic and screening tools [Zackular et al., 2014, Schlaberg, 2020]. While exciting, appropriate statistical methods are still needed to overcome methodological challenges in these exceptional data, so that robust microbial biomarkers and true associations can be discovered among noisy high-dimensional metagenomic data, especially when sample sizes in observational studies are smaller than the number of features discovered.

Before metagenomic data can be used to test for associations, raw sequencing data must be appropriately processed. Taxonomic count tables are created by processing raw 16S or WGS sequencing data through bioinformatics pipelines such as Quantitative Insights Into Microbial Ecology (QIIME) [Caporaso et al., 2010] or mothur [Schloss Patrick D. et al., 2009] for amplicon sequencing data and Metagenomic Phylogenetic Analysis 2.0 (MetaPhlAn2) [Truong et al., 2015] or Kraken [Wood et al., 2019]

[1] This chapter was adapted from our published manuscript available here: Hinton AL, Mucha PJ. A Simultaneous Feature Selection and Compositional Association Test for Detecting Sparse Associations in High-Dimensional Metagenomic Data. Front Microbiol. 2022 Mar 21;13:837396. doi: 10.3389/fmicb.2022.837396. PMID: 35387076; PMCID: PMC8978828.

for WGS data. Sequencing reads are assigned to taxonomic units where the resulting count tables are then used to profile and analyze the association between groups under study at various taxonomic levels (Phylum - Species). These data are often sparse and summarize the total number of reads for each taxonomic assignment within each sample. In current practice, total counts in these settings have been widely recognized as being uninformative due to limitations within sequencing technology [Gloor et al., 2017, Gloor and Reid, 2016, Weiss et al., 2017]. That is, these data carry only relative information, requiring special statistical techniques and considerations. In particular, these relative data have a unit-sum simplex sample space where traditional Euclidean-based statistical methods have limited applicability due to geometrical differences between sample spaces. Ignoring these constraints has been shown to increase type I error [Weiss et al., 2017] and the chance of reporting spurious associations [Pearson, 1897], thus limiting the ability to generalize beyond studies.

A direct way to address simplex sample space constraints imposed by relative data is through a log-ratio transformation. Such transformations, which emerged from the statistical analysis of compositional data [Aitchison, 1982], function by mapping relative data from the unit-sum simplex to traditional Euclidean space. Importantly, log-ratio transformations are sub-compositionally coherent [Aitchison, 1982, Greenacre and Lewi, 2009], independent of the number of dimensions (Taxa, Operational Taxonomic Units (OTUs), etc.) observed in a cohort whereby true associations in the log-ratio form are preserved. This is not true for relative abundance where proportions change as new dimensions are considered, discovered, or removed. Sub-compositional coherence is of practical importance in biomedical studies where biomarker discovery, disease prediction, and beyond-study generalization are paramount. While log-ratio transformations are well-known and routinely applied in some fields [Pawlowsky-Glahn and Buccianti, 2011], their use in metagenomic datasets has been limited. Indeed, significant challenges exist when applying a log-ratio transformation to metagenomic data, including properly handling zeroes [Martín-Fernández et al., 2015, Martın-Fernandez et al., 2003], selecting and interpreting various log-ratio forms [Aitchison, 1982, Egozcue et al., 2003, Greenacre, 2019], and scale differences in counts [Lovell et al., 2020].

While the importance of the compositional nature of metagenomic data has recently been recognized [Gloor et al., 2017, Quinn et al., 2019], relatively few multivariate statistical methods have been developed directly for such data. The current state of the art methods for detecting differential abundance in compositional metagenomic data include ANOVA-like differential expression2 [Fernandes et al., 2014], Analysis of Compositions of Microbiomes [Mandal et al., 2015b], and Analysis of Compositions of

Microbiomes with Bias Correction [Lin and Peddada, 2020]. However, these univariate methods, while powerful, are unable to detect multivariate structure within complex interconnected microbial communities [Layeghifard et al., 2017]. In contrast, appropriate network and multivariate statistical methods — which are appropriate when there exist relationships between a set of variables (i.e., microbial composition) and two or more groups are to be analyzed — can be used to discover complicated microbial patterns, even in settings where there are significantly more variables than samples, and have better control over type I error [Obuchowski, 2005].

Currently, several multivariate statistical methods to detect between-group distributional differences or associations in metagenomic data can be used. A subset of these methods require a suitable beta diversity or between-sample distance (Euclidean, Manhattan, Mahalanobis, etc.) or dissimilarity (Bray-Curtis, weighted/unweighted Unique Fraction, Jaccard, etc.) metric be specified before analysis. Nonparametric tests such as permutational multivariate analysis of variance (PERMANOVA) [Anderson, 2017], Analysis of Similarity (ANOSIM) [Clarke, 1993], and the energy distance [Rizzo and Székely, 2016] can then be applied to test distributional differences between groups. Between-group association signals in metagenomic data may be sparse, i.e., resulting from differences among only a few features (OTUs, taxa, etc.), or they may be densely formed by differences between many features. Importantly, the above-mentioned nonparametric tests lack embedded feature selection and thus may have limited statistical power for detecting sparse signals in high-dimensional data.

Feature selection, which is essential to detecting sparse association signals in high-dimensional metagenomic data, requires sophisticated methods and care to simultaneously select features and test associations while maintaining reasonable type I error control [Baumann, 2003, Lindgren et al., 1996]. Indeed, for this reason, the adaptive microbiome-based sum of powered score (aMiSPU) [Wu et al., 2016] and microbiome higher criticism analysis (MiHC) [Koh and Zhao, 2020] methods were developed to test sparse associations in ultra-high-dimensional OTU-based 16S data (without taxonomic aggregation requiring phylogenetic analysis of sequences). Inspired from concepts put forth in the Direction-Projection-Permutation (DiProPerm) method for assessing statistical significances in high-dimensional settings [Wei et al., 2016], we introduce here the Selection-Energy-Permutation (SelEnergyPerm) method for testing and understanding sparse associations in both 16S and WGS data at the taxonomic level. SelEnergyPerm is the first method to our knowledge to utilize robust pairwise logratios to detect and understand parsimonious logratio signatures from all types of metagenomic data through simultaneous feature selection and association testing. We first show that our novel approach selects smaller subsets of non-redundant logratios that better maximize between-group associations when compared to other popular feature selection methods. Next, we show through an extensive simulation study using synthetic and empirical 16S/WGS data distributions that SelEnergyPerm has, on average, better combined power and false discovery control via the Matthews Correlation Coefficient (MCC) when compared to existing beta-diversity-based approaches. Finally, to demonstrate the utility of SelEnergyPerm in detecting and understanding differences between metagenomic distributions, we apply our method in four case studies utilizing publicly available metagenomic datasets where we test associations between: (1) cerebrospinal fluid microbiomes and post-infectious hydrocephalus in Ugandan infants, (2) delivery mode and the composition of infant gut microbiomes over the first three months of life, (3) adult gut microbiomes and abnormal fecal calprotectin levels, and (4) the gut microbiome composition of infants within the first 6 months of life and future food allergy to egg, milk, or peanuts.

2.5 Methods

2.5.1 Selection-Energy-Permutation (SelEnergyPerm)

In this section, we explain the SelEnergyPerm framework in detail. First, we describe our DCV scoring measure applied to each element of the full set of pairwise logratios (PLR) and then detail the construction of the weighted DCV network representations of these quantities. We next discuss the removal of redundant ratios using a maximum spanning tree that simultaneously maximizes log-ratio variance. After this, we introduce our network-based approach to feature selection and the two multivariate test statistics utilized to measure the strength of the association. We then detail our between-group association maximization algorithm with pseudocode. Finally, we describe the approach for assessing statistical significance via permutation testing using Monte Carlo sampling.

2.5.1.1 Differential Compositional Variation Scoring

For a given metagenomic study, let $\mathbf{M} \in \mathbb{R}^{n \times d}$ be the taxa count table for n samples and d taxa. Before working in the set of all $p = \binom{d}{2} = d(d-1)/2$ PLR of M, we must first address the problem of zero counts. While there are numerous strategies with various drawbacks to model and impute zeros based on type/cause [Martín-Fernández et al., 2015, Palarea-Albaladejo and Martín-Fernández, 2015], there is in general no consensus on which strategy should be used in metagenomic data. Notwithstanding, here we treat zero taxa counts as being below the detection level, and we adopt a corresponding multiplicative replacement strategy for imputing zeros proposed in [Martín-Fernández et al., 2015] that preserves the essential logratio and covariance structure. Specifically, we apply the closure operator to $\mathbf{M}$ to map the count data onto the unit-sum simplex, defining the matrix $\mathbf{X}$ with elements x_{ij} as

$$x_{ij} = (C[\mathbf{M}])_{ij} = \frac{m_{ij}}{\sum_{k=1}^{d} m_{ik}}\,. \tag{2.9}$$

Importantly, we set δ to be a constant equal to the smallest nonzero value across all $\mathbf{X}$ and then replace zeros to obtain $\mathbf{R}$ with elements

$$r_{ij} = \begin{cases} \delta\,, & x_{ij} = 0 \\ x_{ij}\,, & \left(1 - \sum_{k|x_{ij}=0} \delta\right) x_{ij} > 0 \end{cases} \qquad \text{for } i = 1, \ldots, n\,. \tag{2.10}$$

In this way, the interpretation of zeroes is consistent across samples which may not be the case strictly following the Bayesian approach. We then compute all PLRs from $\mathbf{R}$ to obtain $\mathbf{Z} \in \mathbb{R}^{n \times p}$ including all p PLRs (up to a sign). Because feature selection is critical to maximizing power and identifying sparse signals hidden within noisy high-dimensional data, we seek to reduce the dimensionality through feature selection. Notably, this setting is distinct from traditional logratio analysis [Aitchison, 1982] where dimensionality reduction using PCA is applied to all PLR transformed features to reduce dimensionality. Importantly, the set of p different PLRs are not independent of one another and require careful treatment to select ratios that are independent of each other. Here we propose Differential Compositional Variation (DCV), a scoring measure that enables efficient screening and ranking of PLR features within compositional data. Like the screening concept in [Fan and Lv, 2008] for ultra-high-dimensional feature spaces, DCV is motivated by Aitchison's compositional variation array [Aitchison, 1982] where patterns of compositional variability for a group of data can be expressed in terms of the logratio means $\xi_j = E[\mathbf{Z}_{*j}]$ and variances $\tau_j = \text{var}[\mathbf{Z}_{*j}]$ where $j = 1, \ldots, p$. Similarly, let $\zeta_j = \text{median}[\boldsymbol{Z}_{*j}]$.

The DCV score utilizes 5 different statistics to score the contained variation of each logratio; each component of DCV provides unique insight, enabling efficient screening of uninformative logratios for downstream multivariate analysis. Let $\mathbf{y}$ contain the labels for the binary classes/groups c_1 and c_2 under consideration, with n_c indicating the number of samples in class c. In terms of ξ_j and τ_j, the first component of DCV, which measures differences in group means, is Welch's t-statistic:

$$\Delta_j^1 = \frac{\xi_j^{c_1} - \xi_j^{c_2}}{\sqrt{\frac{1}{n_1}\tau_j^{c_1} + \frac{1}{n_2}\tau_j^{c_2}}} \,,$$

where superscripts on ξ_j^c and τ_j^c indicate the mean and variance, respectively, are computed over samples in class c, and we use superscripts on Δ to indicate the different components of DCV (not powers).

Next, we decompose the compositional variability of each $\mathbf{Z}_{*,j}$ using the classical F-statistic to again measure differences of means:

$$\Delta_j^2 = \frac{n_1\left(\xi_j^{c_1} - \xi_j\right)^2 + n_2\left(\xi_j^{c_2} - \xi_j\right)^2}{\tau_j^{c_1} + \tau_j^{c_2}} \,.$$

The third component of DCV is the Brown-Forsythe F-Statistic, measuring heterogeneity of variances, computed as follows. We collect the values for the jth logratio in the array a_{ci}, indexed as the ith sample in class c. From this, let $b_{ci} = |a_{ci} - \zeta_c|$, where ζ_c indicates the median of class c.

$$\Delta_j^3 = \frac{\sum_c n_c(\bar{b}_{c\cdot} - \bar{b}_{\cdot\cdot})^2}{\sum_c \sum_i (b_{ci} - \bar{b}_{c\cdot})^2 / \sum_c (n_c - 1)} \,,$$

where $\bar{b}_{c\cdot}$ indicates the group means and $\bar{b}_{\cdot\cdot}$ is the overall mean of the b_{ci} values.

For the fourth component, we first define the empirical distribution function for each ordered logratio, notated simply here for the j logratio as

$$F_j^c(x) = \frac{1}{n_c}\sum_i \mathbf{1}_c(y_i)\mathbf{1}(Z_{ij} < x)$$

where the $\mathbf{1}_c(y)$ indicator selects out samples in class c and the second indicator indicates whether the Z_{ij} logratio value is less than x, with the sum thus counting the number of samples that satisfy both criteria. We then set the fourth component of DCV to be equal to the Kolmogorov–Smirnov statistic between the different empiricial distributions for the j logratio:

$$\Delta_j^4 = \underset{x}{sup}\left|F_j^1(x) - F_j^2(x)\right|$$

The fifth component of DCV measures the importance of the logratios as attributes in terms of an entropy reduction when splitting by class, as implemented using the information_gain function in the R

FSelectorRcpp package with default settings on the logratio attributes and class response variable. The scores output from this function are organized into Δ_j^5.

We aggregate the different components into the DCV matrix (logratios by DCV components):

$$\mathbf{V} = \begin{bmatrix} \Delta_1^1 & \cdots & \Delta_1^5 \\ \vdots & \ddots & \vdots \\ \Delta_p^1 & \cdots & \Delta_p^5 \end{bmatrix} .$$

To account for differences in scale between the DCV components, we z-score standardize each component (column) to define the standardized DCV matrix $\hat{\mathbf{V}}$: $\hat{v}_{ij} = (v_{ij} - \bar{v}_{*j})/SD(v_{*j})$. The final set of DCV scores, $\breve{\mathbf{V}} \in \mathbb{R}^{p \times 1}$, which contains a score for each logratio, is then defined as

$$\breve{v}_j = \sum_{k=1}^{5} \hat{v}_{jk} \quad \text{where} \quad j = 1, \ldots, p \,.$$

2.5.1.2 DCV network and maximum spanning tree construction

Here we leverage the inherent network structure of logratios [Greenacre, 2019] to form our DCV network, defined as a directed graph where edges point from numerator vertices to denominator vertices. We then define $G = (V, E, W)$ to be the DCV network where V is the set of d taxa vertices, E is the edge set formed by all p pairwise logratios between taxa, and edge weights W are the corresponding DCV scores in $\mathbf{V}$ between classes. In the initial phase of feature selection on $\mathbf{Z}$, we require the logratio subsets to meet three important properties: 1) explain maximum logratio variance, 2) form a linearly independent set, and 3) contain maximum DCV. Notably, by construction the column rank of $\mathbf{Z}$ is $(d - 1)$ and thus any single-component connected network containing all d taxa explains 100% of the logratio variance contained in $\mathbf{Z}$. The second property requires the undirected version of the logratio subset to be acyclic, as may be achieved with a spanning tree. However, the number of spanning trees from G can be expressed by Cayley's formula: $T_{|V|} = |V|^{|V|-2}$. To circumvent considering this unmanageably large number of spanning trees, we utilize the weights imposed from the DCV scoring to enable efficient selection of a suitable spanning tree from G, as described next.

We sort the logratios of $\breve{\mathbf{V}}$ in descending order by DCV score to form $\breve{\mathbf{V}}'$ and retain the first set of q logratios that contain all d taxa to form $\breve{\mathbf{V}}''$. We then redefine the logratio network $G = (V, E, W)$ where V is the set of d taxa vertices and E is the edge set corresponding to these q pairwise logratios, with edge weights W from the values in $\breve{\mathbf{V}}''$. In practice, we have always found that the resulting network at this stage is a single connected component — in the event that the network is not, additional logratios from $\breve{\mathbf{V}}'$ should be added to make it connected. Finally, from G we compute the maximum spanning tree G_{MST} using the R igraph package and define $\mathbf{Z}' \in \mathbb{R}^{n\times(d-1)}$ to be the subset of logratios corresponding to the edge set of G_{MST}.

2.5.1.3 Multivariate Test Statistics

SelEnergyPerm considers two multivariate test statistics to determine the statistical significance of retained subsets of logratios. The first multivariate test statistic, the Distance Components F-ratio (discoF) is utilized when between-group dispersion effects are not detected in $\mathbf{Z}'$. The discoF statistic, proposed by [Rizzo and Székely, 2010], is like the traditional Analysis of variance ‘F’ ratio (but does not follow an F-distribution) where the total dispersion is partitioned into between- and within-group components derived from an inter-sample Euclidean distance matrix computed from $\mathbf{Z}'$. Computation of the discoF statistic is done here using the energy R package. As described by [Rizzo and Székely, 2010], the discoF test statistic for binary groups is of the form

$$F_{n,\alpha} = \frac{S_{n,\alpha}}{W_{n,\alpha}/(n-2)}$$

where $S_{n,\alpha}$ is the between-sample energy statistic, $W_{n,\alpha}$ is the within-sample dispersion statistic and $0 < \alpha \leq 2$ is the exponent on the pairwise between-sample norm. See [Rizzo and Székely, 2010, 2016] for specific details on computing the between- and within-group components of the discoF statistic. Here we use the energy R package to compute the discoF statistic.

The second statistic, used by SelEnergyPerm when dispersion effects between groups are detected in $\mathbf{Z}'$, is a scaled combined-F (cF) statistic which is distribution-free and attempts to jointly account for differences in both location and scale between distributions. The unscaled cF statistic is the sum of F-ratios obtained from PERMDISP2 with spatial medians [Anderson, 2006] and PERMANOVA [Anderson, 2017], computed using the R vegan package. We partition the variation of $\mathbf{Z}'$ and define the unscaled combined-F statistic as

$$\widetilde{cF} = F_{\text{location}} + F_{\text{dispersion}} = \left(\frac{SS_\alpha}{SS_w/(n-2)}\right) + \left(\frac{SS_T}{SS_E/(n-2)}\right)$$

where SS_α and SS_T are the between-group sum of squares components, and SS_w and SS_E are the within-group sum of square components of variation from the PERMANOVA (F_{location}) and PERMDISP2 ($F_{\text{dispersion}}$) procedures, respectively. See [Anderson, 2017] and [Anderson, 2006] for specific details on computing these between- and within-group components. Likewise, the scaled combined-F statistic that we use is computed in the same way but with z-score standardization relative to the permutation distribution. Let $\boldsymbol{n}\mathbf{F}_{\text{loc.}}$ and $\boldsymbol{n}\mathbf{F}_{\text{disp.}}$ be m-dimensional vectors of null $F_{\text{loc.}}$ or $F_{\text{disp.}}$ statistics sampled from the permutation distribution. We scale $\hat{F}_{\text{loc.}} = \frac{F_{\text{loc.}} - E[\boldsymbol{n}\mathbf{F}_{\text{loc.}}]}{SD(\boldsymbol{n}\mathbf{F}_{\text{loc.}})}$ and $\hat{F}_{\text{disp.}} = \frac{F_{\text{disp.}} - E[\boldsymbol{n}\mathbf{F}_{\text{disp.}}]}{SD(\boldsymbol{n}\mathbf{F}_{\text{disp.}})}$ and define the scaled combined-F statistic as

$$cF = \hat{F}_{\text{loc.}} + \hat{F}_{\text{disp.}}\,, \tag{2.11}$$

taking care to note that cF is approximate and thus the estimate has variability based on the number of samples drawn from the permutation distribution. We consider $m = 10^6$ samples here as a balance between computational cost and minimizing this variation.

2.5.1.4 Association Maximization and Greedy Forward Selection

In this step, we focus on the multivariate structure formed by a subset of logratios. Specifically, we are interested in maximizing the between-group variation induced by a subset of logratios in a low-dimensional multivariate space. To find a minimal, statistically-significant subset of logratios that maximizes $F_{n,\alpha}$ (location effects only) or cF (dispersion and location effects) between classes, we utilize a greedy forward stepwise feature selection procedure (see Appendix A: Algorithm 1). This procedure is notated here as selectionEnergy().

2.5.1.5 Association Significance testing

To assess the statistical significance of the observed association $F^{obs} = \text{selectionEnergy}(\mathbf{Z}^{obs}, \mathbf{y})$ we compute the null distribution by permutation testing via Monte Carlo sampling [Ernst, 2004]. Letting the number of permutations be k and $\boldsymbol{\pi}$ be the set of random permutations of $\mathbf{y}$, we obtain samples from the null distribution by $\mathbf{F}^{null} = \text{selectionEnergy}(\mathbf{Z}^{obs}, \boldsymbol{\pi})$. We then test if the F^{obs} is more extreme than what is expected at random given the data using the one-sided estimated p-value

$$\hat{p} = \frac{1 + \sum_{i=1}^{k} \mathbf{1}(F^{null} > F^{obs})}{k+1}$$

2.5.2 Simulation Strategy

We adapted several simulation settings to investigate and highlight key association detection characteristics of SelEnergyPerm when compared to ANOSIM, PERMANOVA, and the energy test. Additionally, to detect the presence of heterogeneity of multivariate dispersion between groups and understand its impact on association detection, we utilized the PERMDISP2 method as an indicator. The empirical association detection ability of each method was assessed within a binary classification framework. To do this, we measured the rate of each statistical test to correctly reject (Power) or accept (Type I Error) the null hypothesis (no difference between groups) at significance $\alpha = 0.05$. Further, to truly assess detection capabilities, we presented each method with binary instances drawn from either the same (Null Case) or different (True Case) distributions for each scenario using Monte Carlo simulations. The Matthews Correlation Coefficient (MCC), which effectively summarizes the binary confusion matrix, was then used to measure the overall accuracy of each method's ability to detect associations across various simulation scenarios. MCC was computed as

$$MCC = \frac{(\mathrm{TP})(\mathrm{TN}) - (\mathrm{FP})(\mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

where $\mathrm{TP} =$ true positive (reject the null hypothesis for True Case), $\mathrm{TN} =$ true negative (accept null hypothesis for Null Case), $\mathrm{FP} =$ false positive (reject the null hypothesis for Null Case), and $\mathrm{FN} =$ false negative (accept null hypothesis for True Case). For each simulation scenario, we generated 100 simulated datasets with 40 samples each in class 1 and class 2 for the balanced binary design and 20/60 (class 1/2) samples for the unbalanced design. Given we rely on permutation testing for significance of all methods, we generate a common set of 150 permutations per dataset to consistently compute significance for each method across all scenarios and settings.

2.5.2.1 Simulation Scenarios (Synthetic Data)

For all synthetic data scenarios, we consider datasets with $d =$ 50, 150, and 250 taxa, yielding a total of $p =$ 1225, 11175, and 31125 pairwise logratios, respectively. We note, based on our experience, that the sizes d tested, while modest, are in general reflective of the actual number of taxa typically analyzed for 16S or WGS datasets after sparse taxa are removed. Each of the following simulation scenarios is available in our SelEnergyPermR R package available at https://github.com/andrew84830813/selEnergyPermR using the function scenarioN() where N = [1,5]. All synthetic scenarios are inspired by settings considered in [Wei et al., 2016].

In Scenario 1, for the true case, we consider both multivariate location (in all dimensions) and dispersion effects that grow with increased numbers of dimensions. The increase in dispersion with dimension is similar to settings studied in [Wei et al., 2016]. Here, data from each sample are generated from the Dirichlet distribution $\mathbf{Dir}(\alpha)$, commonly used to model compositional data whereby data are naturally constrained within the unit-sum simplex. Data from class 1 are simulated with $\alpha_1 = 3$. Data from class 2 are generated with $\alpha_2 = \frac{3}{5} \log d$ where the $\log(d)/5$ factor shifts the overall location and increases dispersion as the dimensionality increases. For the null case, data from both classes are generated from $\mathbf{Dir}(\alpha_1)$.

In Scenario 2, for the true case, we generate sparse count data from two Dirichlet distributions that differ in the location of the first component only and overall dispersion. To generally mimic real library size or total counts per sample, we use a negative binomial (NB) distribution to model the total counts for each sample and simulated as $C_i \sim NB(s, s/(s+\mu))$ where $s = 1$ and $\mu = 10^7$. Notably, other discrete distributions can be used to achieve user specified library size characteristics. Count data for class 1 were generated by rounding $C_i \cdot \mathbf{Dir}(\boldsymbol{\alpha}_1)$ where α_1 elements are drawn from uniform distributions as

$$\boldsymbol{\alpha}_1 = \left(x_1 \sim U_{[3000,5000]}, x_{i\in[2,10]} \sim U_{[500,1500]}, x_{i\in[11,d]} \sim U_{[1,5]}\right).$$

Count data for class 2 were generated after rounding $C_i \cdot \mathbf{Dir}(\boldsymbol{\alpha}_2)$ where the $\boldsymbol{\alpha}_2$ elements are drawn as

$$\boldsymbol{\alpha}_2 = \left(x_1 \sim U_{[12500,17500]}, x_{i\in[2,10]} \sim U_{[500,1500]}, x_{i\in[11,d]} \sim U_{[1,5]}\right).$$

Notably, we use the $x_{i\in[11,d]} \sim U_{[1,5]}$ terms here to model random sparsity. For the null case, data from both classes are generated from $C_i \cdot \mathbf{Dir}(\boldsymbol{\alpha}_1)$.

In Scenario 3, for the true case, we generate compositional data with a large location effect that increases while the dispersion effects decrease with dimensionality. These settings are similar to settings considered for association benchmark comparisons in [Wei et al., 2016]. We simulate data from the additive logistic normal distribution on the simplex [Aitchison, 1982]. To do this we first let $\boldsymbol{S}_1 = N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\boldsymbol{S}_2 = N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ be samples drawn from multivariate normal distributions. We set $\boldsymbol{\mu}_1 = (0, \ldots, 0)$ and

$\boldsymbol{\mu}_2 = (1/\sqrt{d}, \ldots, 1/\sqrt{d})$ in the first 25% of dimensions and 0 in the remaining dimensions. The covariance structure was defined in the same way as in [Wei et al., 2016] where $\boldsymbol{\Sigma}$ was defined with 1's along the main diagonal and 0.2 along the two diagonals off the main. From this, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma} + \delta I_d$ and $\boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma} + \mathbf{U} + \delta I_d$ where $\mathbf{U} \in \mathbb{R}^{d \times d}$ is a matrix with $U_{[0,32/d^2]}$ entries and

$\delta = |\min(\text{eigenvalues}(\boldsymbol{\Sigma}), \text{eigenvalues}(\boldsymbol{\Sigma} + \mathbf{U}))| + 0.05$. Here row vectors from $\boldsymbol{S}$ represent additive logratio (ALR) vectors and are subsequently projected onto the simplex using the inverse additive logratio transformation defined in terms of the closure operator as $\text{ALR}^{-1} = C[\exp([s, 0])]$. For the null case, data for both classes were simulated from $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$.

In Scenario 4, for the true case, we generate compositional data with sparse location effects in the first dimension that grow stronger while dispersion effects grow weaker as the dimensionality increases. That is, $\boldsymbol{S}_1 = N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\boldsymbol{S}_2 = N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ are defined as in scenario 3 except we set $\boldsymbol{\mu}_2$ to $\log \frac{d}{3}$ in the first dimensions and 0 in the remaining dimensions. The simplex projection and null case are done as described in scenario 3.

Finally, in Scenario 5 for the true case, we generate compositional data from the additive logistic normal distribution with a small location shift and large dispersion difference that increases with dimensionality. Let $\boldsymbol{S}_1 = N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\boldsymbol{S}_2 = N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ be defined in as in scenario 3 except for μ_2 set to $\frac{1}{\sqrt{n_1 + n_2}}$ in all dimensions and $\mathbf{U} \in \mathbb{R}^{d \times d}$ with entries drawn from $U_{[0,32]}$. The simplex projection and null case are done as described in scenario 3.

2.5.2.2 Simulation Scenarios (Experimental Data)

For all experimental data scenarios, we used publicly available taxa count tables where sequencing data were already pre-processed. Each of the following simulation scenarios are available in our SelEnergyPermR R package available at https://github.com/andrew84830813/selEnergyPermR using the functions simFromExpData.covarianceShift() or simFromExpData.largeMeanShft(). Notably, the simulation scenarios below first convert count data into compositonal data. To control simulation parameters, the compositonal data are modeled using the additive logistic normal distribution [Aitchison, 1982]. After adjusting the mean/covariance structures in a controlled way, the compositional data are then converted back to count data for analysis.

For general 16S data characteristics, we utilized the *ob_goodrich_results.tar.gz* dataset from the microbiomeHD database [Goodrich et al., 2014, Duvallet et al., 2017]. We aggregated the taxa to the genus level (distinct genera = 247) and extracted the 428 healthy samples from the goodrich16S dataset for our

16S data simulations. For WGS data characteristics, we utilized the *ZeeviD2015* [Zeevi et al., 2015] dataset from the curatedmetagenome [Pasolli et al., 2017] database. We aggregated taxa counts by species (distinct species = 1,776) and extracted the 900 control samples for our WGS data simulations. Here we model the 16S and WGS count data using zero-inflated negative binomial (ZINB) models which have been shown to be a reasonable choice for modeling microbiome count data [Calgaro et al., 2020]. ZINB models were fit to the 16S and WGS dataset described above using the ZINBWAVE R package with default settings. For all experimental data scenarios, we used the fitted 16S/WGS ZINB models to simulate new samples for each dataset. That is, we first simulated 428 samples from the ZINB model for the 16S datasets or 900 samples for the WGS datasets. We then randomly select 40 samples per class for the balanced design and 20/60 (classes 1/2) samples for the unbalanced design. To reduce the presence of rare features we only retained features present in at least 15% of all samples for all datasets.

For Scenario 1, for the true case in both 16S and WGS datasets, we consider settings where the percent $P = \{5, 20, 35, 50\}$ of dimensions with a location shift increases while the dispersion effect between classes remains fixed. To do this, we first simulate count data $\mathbf{M}$ from the ZINB model, map it onto the unit-sum simplex using Eq. 2.9 and impute zeros to obtain $\mathbf{R}$ as in Eq. 4.13. The ALR transformation is then applied to $\mathbf{R}$ to obtain $\boldsymbol{A}$ with elements $a_{ij} = \log(r_{ij}/r_{id})$ for $j = 1, \ldots, (d-1)$.

For each class we simulate data from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where

$$\boldsymbol{\mu} = \mathrm{E}[\boldsymbol{A}] = (\mathrm{E}[a_{*1}], \ldots, \mathrm{E}[a_{*d-1}])^T$$

and

$$\boldsymbol{\Sigma}_{ij} = \mathrm{cov}[a_{*i}, a_{*j}]$$

The variance (diag($\boldsymbol{\Sigma}$)) of each dimension is ranked in ascending order whereby $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are reordered accordingly to form $\boldsymbol{\mu}_r$ and $\boldsymbol{\Sigma}_r$. Of note, this is done to ensure the location shift occurs in features with minimal variance. We then simulate S_1 from $N(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$ with $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_1$ using as above. Letting $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_1$ we then shift the first $P_i\%$ of dimensions of $\boldsymbol{\mu}_2$ by a factor of 1.25. From this we simulate S_2 from $N(\boldsymbol{\mu}_2, \Sigma_1)$. Finally, S_1 and S_2, which are in Euclidean ALR form, are mapped back to the simplex (relative abundance) using the inverse ALR transformation. For the null case, data for both classes are simulated from $N(\boldsymbol{\mu}_r, \Sigma_r)$.

Finally, for Scenario 2, we consider settings for the true case (in both 16S and WGS datasets) with location shifts in the first 10% of dimensions that are confounded by increasing dispersion effects as the number of dimensions increase. Here we compute S_1 in Euclidean ALR form as described in Scenario 1 (Experimental Data) such that $S_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$. From this, $\boldsymbol{\Sigma}_{s_1} = \boldsymbol{\Sigma}_1 + \delta I_d$ and $\boldsymbol{\Sigma}_{s_2} = \boldsymbol{\Sigma}_1 + \mathbf{T} + \delta I_d$ where $\mathbf{T}$ is a $d \times d$ matrix with entries drawn from $U_{[0,\beta_i]}$ and $\delta = |\min(\text{eigenvalues}(\boldsymbol{\Sigma_1}), \text{eigenvalues}(\boldsymbol{\Sigma_1} + \mathbf{T}))| + 0.05$. For 16S data $\beta = (0.10, 1.40, 2.70, 4.00)$ and for WGS data $\beta = (0.10, 4.07, 8.03, 12.00)$. Additionally, letting $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_1$ we shift the first 10% of dimensions of $\boldsymbol{\mu}_2$ by a constant factor of 1.25 for WGS data and by a factor $F = (1.20, 1.17, 1.13, 1.10)$ for 16S data. In all, the final multivariate forms are $S_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{s_1})$ and $S_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{s_2})$. These distributions, which are in ALR form, are mapped back onto the simplex using $\text{ALR}^{-1}(s_{i*}) = C[\exp([s_{i*}, 0])]$. Lastly, for the null case, data for both classes are simulated from $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{s_1})$.

For both scenarios, counts could alternatively be obtained via a negative binomial distribution (or other suitable discrete distribution) using a similar process as described in Scenario 2 of the Synthetic Data simulation section above.

2.5.3 Feature Selection Benchmarks

For the feature selection (FS) benchmark we used the Boruta R package with maxRun set to 100 and importance set to Gini for the Boruta FS. The glmnet R package was used for LASSO FS where alpha was set to 1 and lambda was optimized via cross-validation. The caret R package was used to implement RFE FS where 5-fold cross-validation was used to evaluate AUC and feature importance of sets $s = \{2^1, 2^2, \ldots, 2^n\}$, where $n = \text{floor}(\log_2 p)$. The FSelectorRcpp R package with default settings was used for the Information Gain Filter FS. For each Scenario (Synthetic Data), FS characteristics were evaluated on 200 synthetic datasets across feature space sizes of $p = \{1225, 4950, 11175, 19900, 31125\}$ logratios. Performance characteristics considered were the number of logratios selected, log-ratio network clustering coefficient, and the combined-F statistic. Here we use the number of logratios selected by each method as a proxy for model complexity. Specifically, higher model complexity or the number of features retained increases the risk of overfitting and unnecessarily reduces the biological interpretation corresponding to the logratios. Log-ratio networks were formed using the final subset selected by each method, defined as a graph where vertices represent taxa and edges connect taxa pairs to form a logratio. Redundancy in a log-ratio network of this type can be inferred from cycles in the network. While it does not detect all cycles, the clustering coefficient can be used here to detect cycles between three nodes (closed triangles versus triplets). Computation of the global clustering coefficient was done using the R igraph package. Finally, the cF statistic, measuring the strength of the overall association, was computed as in Eq. 2.11 for each subset. All performance characteristics were evaluated in both balanced and unbalanced sampling designs. Computational time was recorded in seconds for each simulation scenario, feature space, and sample design. The recorded time represents the CPU time required by each FS method to select the final logratio subset. All computations were run on UNC–Chapel Hill's Linux-based Longleaf cluster in R parallelized with 10 cores using the foreach R package with 5GB of RAM.

2.6 Results

To robustly uncover sparse microbial signatures while simultaneously testing multivariate group associations, we based our SelEnergyPerm framework on a novel network-based feature selection approach combined with permutation testing for sparse high-dimensional low-sample-size compositional metagenomic data. Our framework (Figure 2.1A), which selects from all pairwise logratios between features (Taxa, OTUs, etc.), first scores the between-group variation of individual logratios using our Differential Compositional Variation (DCV) scoring measure (see Methods). From this, a weighted DCV log-ratio network is formed and subsequently pruned to reduce redundancy and complexity via a maximum spanning tree. Final subsets are then selected by maximizing the between-group association using a greedy forward stepwise selection procedure. Multivariate test statistics, which measure the strength of the association between groups, are then computed on the final retained subset. Statistical significance is determined by repeating this process with permuted group labels to obtain the permutation distribution of the test statistic of interest under feature selection. In this way, we determine whether the observed association is larger than what would be expected by chance (Figure 2.1B). To this end, our framework tests the overall null hypothesis of no association between the metagenomic composition and group labels.

2.6.1 Feature selection comparison to other methods

We first benchmarked the multivariate characteristics of subsets selected by our feature selection approach against a set of other popular methods for feature selection: Boruta [Kursa et al., 2010], Least Absolute Shrinkage and Selection Operator (LASSO) [Tibshirani, 1996], Information Gain Filtering [KENT, 1983], and Random Forest Recursive Feature Elimination (RFE) [Granitto et al., 2006]. The benchmarks were carried out by varying the number of log-ratio dimensions in the full feature set using five simulation scenarios, considering both balanced and unbalanced sampling designs (see Methods). Specifically, for

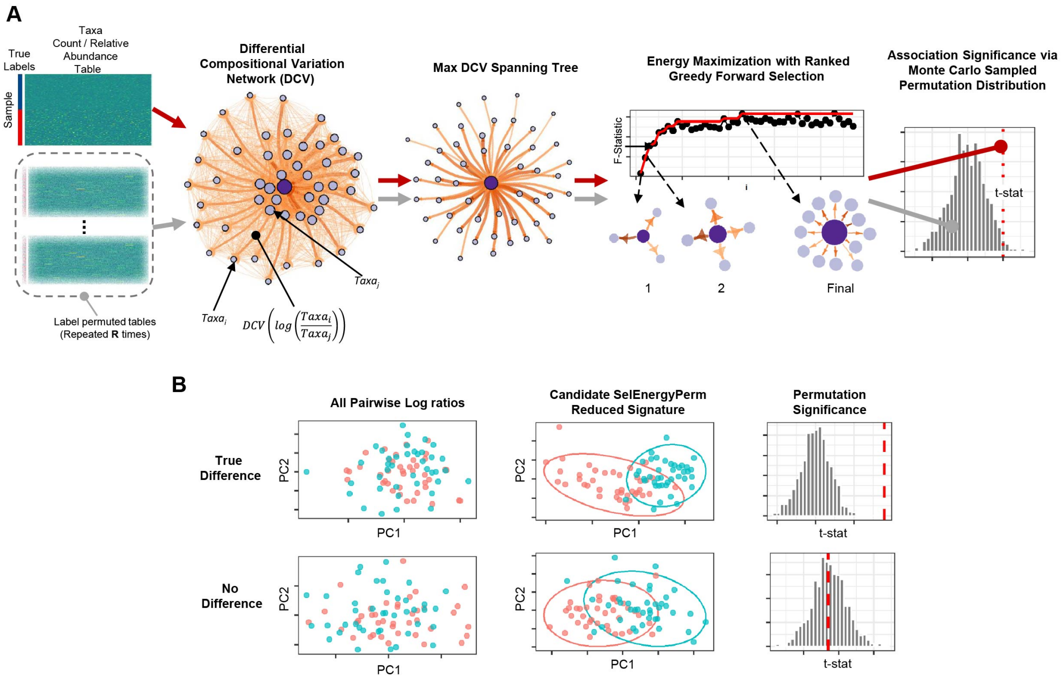

Figure 2.1: Overview of the SelEnergyPerm framework for non-parametric group association testing in metagenomic data A. Relative abundance/count data are transformed using all pairwise logratios. These logratios are subsequently scored (DCV) and used to efficiently select a subset that: (1) is independent via a maximum spanning tree and (2) maximizes the energy or association between groups via greedy optimization. The entire process is repeated using permutation testing to control false discovery and assess statistical significance. B. Detection/rejection of sparse associations hidden within high dimensional data via simultaneous feature selection and permutation testing.

subsets returned by each method, we studied the number of logratios selected (as a proxy for model complexity), the clustering coefficient of the log-ratio network (measuring log-ratio redundancy), and the combined F-statistic (strength of association, see Methods), and the computational time required to return the final subset (Appendix A: Figure A.1). In Figure 2, we present results from scenarios with a balanced sampling design. Notably, the results for the unbalanced sampling design scenarios are similar and do not change the overall comparative interpretation (Appendix A: Figure A.2). Examination of the clustering coefficient across all simulation scenarios/dimensions demonstrates that SelEnergyPerm consistently selects linearly independent subsets of logratios (Figure 2.2 and Appendix A: Figure A.2, clustering coefficient = 0), in contrast with the subsets observed in other methods tested. Of note, a clustering coefficient > 0 indicates a selected log-ratio subset contains at least one triple of linearly-dependent logratios (closing a triangle in the log-ratio network), thereby unnecessarily increasing dimensionality and model complexity. (We note that any cycle present in a log-ratio network indicates linear dependence, though we did not test for cycles larger than closed triangles. We emphasize that by construction the SelEnergyPerm-selected subsets do not include any such cycles.) Additionally, the number of logratios retained by each method across every scenario tested revealed subsets selected by SelEnergyPerm were, on average, 14 to 149 times smaller than other methods (Figure 2.2 and Appendix A: Figure A.2).

Next, the strength of the association measured by the combined-F statistic (see Methods) indicates SelEnergyPerm-selected subsets typically capture higher between-group variations than other methods tested. In Scenarios 2–4, SelEnergyPerm subsets were observed to have on average, higher combined-F values than all other methods across all dimensions tested (Figure 2.2 and Appendix A: Figure A.2). Meanwhile, in Scenarios 1 and 5, SelEnergyPerm subsets generally performed similarly to the other methods but better as the dimensionality increased. Notably, Scenarios 1 and 5 do not simulate sparse association signals and have strong between-group dispersion effects present. These results indicate SelEnergyPerm returned subsets better capturing sparse associations (Scenarios 2–4) than the other feature selection methods tested. Computational time experiments show, across all scenarios tested, SelEnergyPerm is on average faster than Boruta and RFE but slower than LASSO and Information Gain Filtering (Appendix A: Figure A.1). Overall, SelEnergyPerm subsets were non-redundant, significantly more parsimonious, and captured stronger associations than other methods tested, thereby enabling robust biological interpretation using logratios in high-dimensional feature spaces.

2.6.2 Detection of sparse associations in synthetic data

Here, we use data simulated from theoretical distributions to compare the ability of SelEnergyPerm, PERMANOVA, ANOSIM, and the energy test to detect associations in sparse high-dimensional data. That is, we are interested in determining how well each method accepts or rejects the null hypothesis (no difference between groups) when presented with two groups of data that, as ground truth, come from the same (Null Case; Type I error assessment) or different (True Case; power assessment) distributions. From this, we report the performance of each method in terms of the Matthews Correlation Coefficient (MCC) at $\alpha = 0.05$ for 4 simulation scenarios (see Methods) with balanced or unbalanced sampling designs (Figure 2.3). For brevity, we shall refer to the collection of PERMANOVA, ANOSIM, and energy tests as the standard methods.

In Scenario 1, where data are simulated from a Dirichlet distribution with between-group location and dispersion effects that grow as the number of dimensions increase (see Methods), we see for the balanced design that both SelEnergyPerm and the energy test perform well over all dimensions (number of logratios)

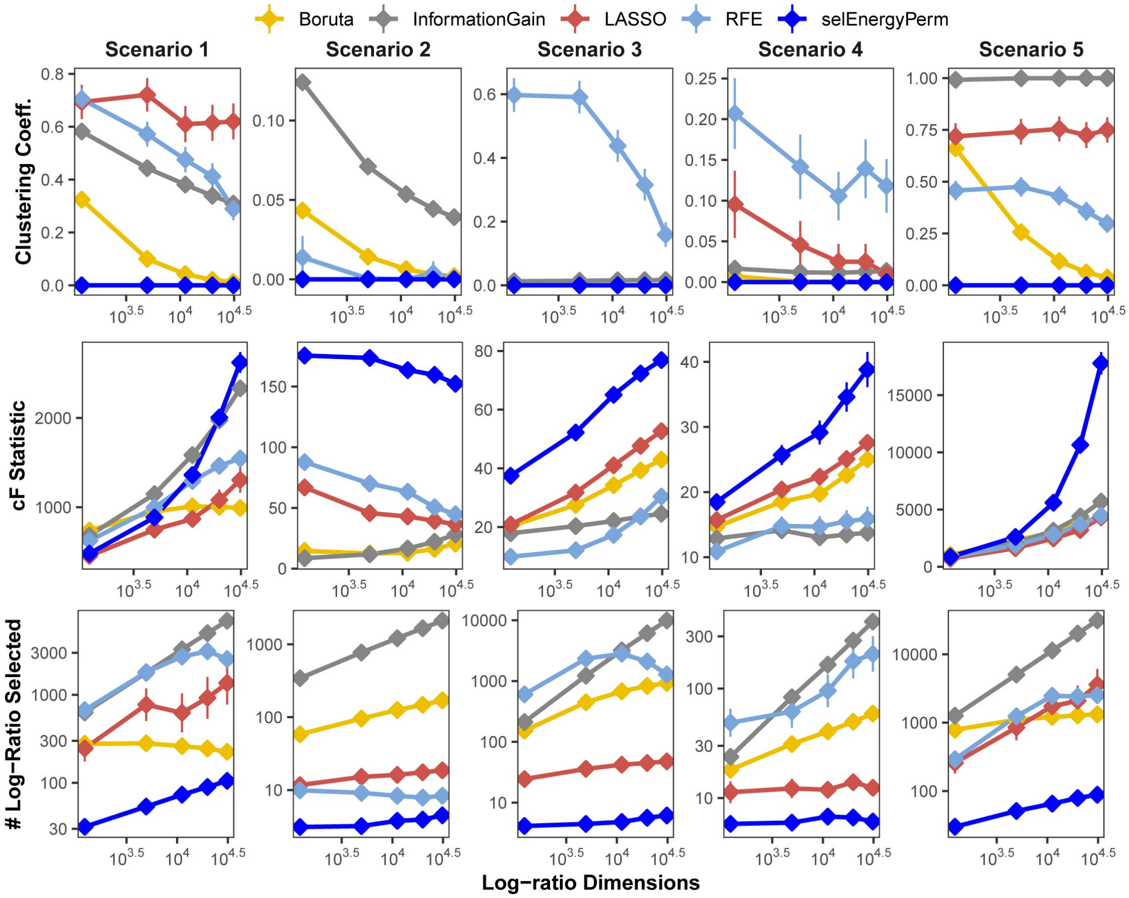

Figure 2.2: SelEnergyPerm-selected log-ratio subset characteristics compared with Boruta, Information Gain Filtering, LASSO, and RFE across five simulation scenarios for the balanced sampling design. Using 200 simulations for each scenario-dimension by method we assessed: (Top Row) the clustering coefficient of logratio networks formed by selected subsets returned from each method, (Middle Row) the magnitude of the association as measured by the combined-F (cF) statistic on selected subsets returned from each method, and (Bottom Row) the number of logratios returned by each method. Points are the mean for each experimental condition and error bars indicate 95% confidence interval.

tested. Notably, ANOSIM loses the ability to detect associations as the number of dimensions increases while PERMANOVA performs poorly over all dimensions. The poor performance of ANOSIM and PERMANOVA is directly attributable to the underlying heterogeneity of variance present in the data generated in this scenario; these limitations of PERMANOVA and ANOSIM have been discussed previously [Anderson and Walsh, 2013]. The presence of dispersion effects is confirmed with the Distance-Based Tests for Homogeneity of Multivariate Dispersions (PERMDISP2) [Anderson, 2006] method and can be observed to be steady (Figure 2.3 - Scenario 1) and increasing across dimensions. For the unbalanced design, SelEnergyPerm and the energy test both retain strong performance and have comparable performance over most dimensions, whereas ANOSIM completely loses the ability to detect associations under the unbalanced design and PERMANOVA continues to perform poorly across all dimensions.

For Scenario 2 (Figure 2.3), the data distributions for each group are simulated from two Dirichlet distributions that differ in the location of the first component and overall variance. That is, this scenario embeds a sparse signal (location shift) in the first dimension with random noise in the remaining dimensions. The results for this scenario show that for the balanced case SelEnergyPerm performs significantly better than all other methods tested. For the unbalanced case, SelEnergyPerm performs better than all other methods for smaller numbers of dimensions, however, it performs similarly to ANOSIM as the number of dimensions increases. Notably, the performance of ANOSIM improves as the number of dimensions increases for both the balanced and unbalanced cases.

For Scenario 3 (Figure 2.3), the data distributions for the first class are simulated from the additive logistic normal distribution. Data for the second class are also generated from an additive logistic normal distribution with the same parameters (same covariance matrix) but with location shifts in the first 25% of the dimensions. Under this scenario, we observed the performance of SelEnergyPerm to be comparable to the standard methods for the balanced case and slightly worse than the standard methods for the unbalanced case. The reduced performance in the unbalanced case is attributable to the dense signal (25% of features) being in direct tension with the SelEnergyPerm objective of reduced feature selection.

Lastly, in Scenario 4 (Figure 2.3), a location shift only (same between-class covariance structure) was embedded in the first component of two additive logistic normal distributions, with the shift increasing with the number dimensions. Here, SelEnergyPerm outperformed the standard methods as the number of dimensions increased for both the balanced and unbalanced cases. While performing better overall relative to the standard methods, a notable decrease in performance from the balanced to the unbalanced case was observed for SelEnergyPerm. This decrease in performance was exacerbated among the standard methods where performance not only decreased between sampling designs but also generally declined as the number of dimensions increased in the unbalanced design.

Overall sparse association detection performance as measured by MCC, sensitivity, specificity, positive predictive value, negative predictive value, Youden index, and false-positive rate across all scenarios at an $\alpha = 0.05$ are shown in Appendix A: Figure A.3. These aggregate results demonstrate SelEnergyPerm generally outperforms the standard methods for detecting sparse associations under the synthetic data simulation scenarios considered here.

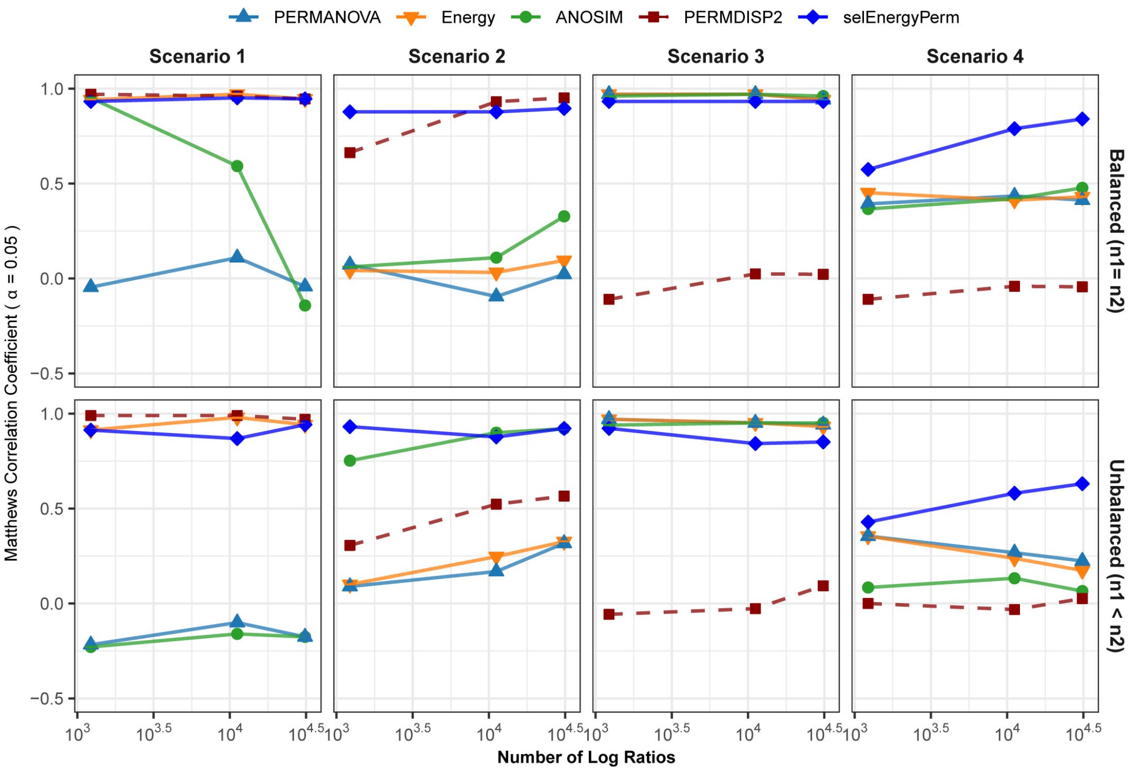

Figure 2.3: Comparison of the Matthews Correlation Coefficient measuring the ability of each method to properly detect/reject associations in data generated from synthetic distributions in both balanced and unbalanced sampling designs. For each scenario and logratio feature space size, test datasets were simulated to include data distributions that have either true between-group differences (n=100) or no between-group difference (n=100). Results from the PERMDISP2 procedure are displayed to indicate heterogeneity of variance between groups.

2.6.3 Detection of sparse associations in data simulated from empirical datasets

To further assess performance, we benchmarked our method against the standard methods on data simulated from properties observed in real metagenomic datasets. In this way, unique metagenomic data characteristics such as sparsity, over dispersion, and complex co-occurrence patterns are assessed synthetically. As above, MCC is used to assess the ability of each method to detect associations across these settings.

In the first setting, (Figure 2.4 – 16S/WGS: Increasing Covariance Diff.), an increasing covariance effect with a decreasing location effect between classes was simulated using healthy subsets of 16S and WGS samples. The increasing dispersion effect is confirmed with PERMDISP2 for both sampling designs (Figure 2.4). For 16S and WGS data with a balanced sampling design, SelEnergyPerm outperforms the standard methods across all effect sizes and has strong performance as the number of dimensions increases.

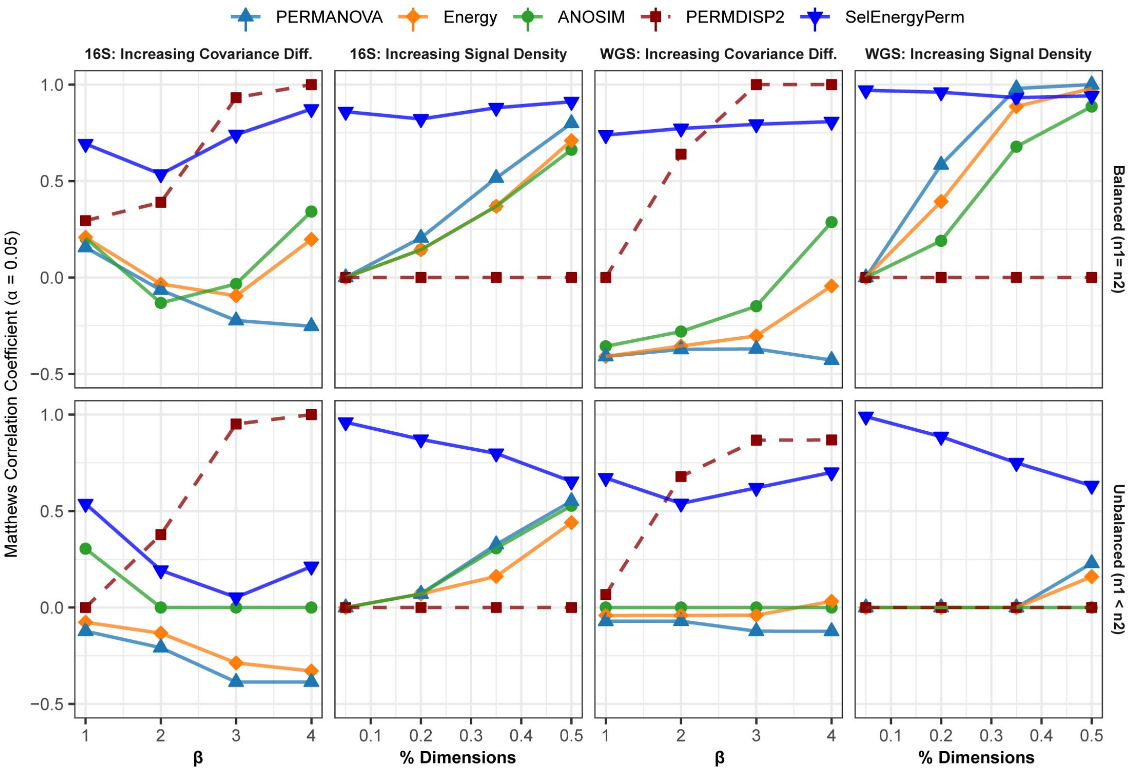

Figure 2.4: Comparison of the Matthews Correlation Coefficient measuring the ability of each method to properly detect/reject associations in data simulated from real 16S and WGS data distributions in both balanced and unbalanced sampling designs. For each data type and scenario, datasets were generated to include data distributions that have either true between-group differences (n=100) or no between-group difference (n=100). Results from the PERMDISP2 (dashed line) procedure are displayed to indicate heterogeneity of variance between groups.

For 16S data with an unbalanced design, all methods performed poorly as the location shift effect increases. This trend is traceable to the strong embedded covariance effect between classes, which is a known confounder in high-dimensional association settings [Anderson and Walsh, 2013]. Notably, only SelEnergyPerm and ANOSIM maintain positive MCCs on average, indicating these methods better control type I error (albeit with severely limited power) under this sampling design. For WGS data with an unbalanced design, SelEnergyPerm outperformed the standard methods and had better association detection across all effect levels.

For the second simulation setting, (Figure 2.4 – 16S/WGS: Increasing Location Effects), we simulated large location shifts between classes by increasing the size of the association signal from 5% to 50% of all features with fixed covariance structures. These shifts were computed using synthetic subsets of WGS and 16S samples from publicly available healthy gut microbiomes. Indeed, PERMDISP2 analysis confirmed the absence of covariance effects. For both 16S and WGS data with a balanced sampling design, SelEnergyPerm outperformed all standard methods. As expected, in both WGS and 16S data, the performance of the standard methods increased as the association signal became less sparse. Again, for the unbalanced design in both WGS and 16S data, SelEnergyPerm outperformed all standard methods. Importantly, the detection ability of the standard methods improved as the association signal became less sparse.

Finally, overall sparse association detection performance metrics are shown in Appendix A: Figure A.4. These aggregate results demonstrate SelEnergyPerm has better overall sparse association detection performance when compared to standard methods using data simulated from real 16S and WGS datasets.

2.6.4 Cerebrospinal fluid microbiomes association with post-infectious hydrocephalus

The cerebral spinal fluid (CSF) of Ugandan infants was profiled by Paulson et al. using 16S sequencing to characterize microbial agents associated with Post Infectious Hydrocephalus (PIH) following neonatal sepsis [Paulson et al., 2020]. This processed gut microbiome dataset, retrieved from microbiomeDB [Oliveira et al., 2018], consisted of 369 distinct taxa measured on 92 samples (58 PIH and 34 Non-Post Infectious Hydrocephalus (NPIH) patients). Removing taxa not present in at least 10% of samples yielded 57 total distinct taxa (i.e., 1,596 logratios between taxa). We apply SelEnergyPerm to determine if there was an association between the microbiome composition in the CSF and PIH/NPIH disease status. We then utilize the reduced SelEnergyPerm log-ratio signature of PIH in CSF to gain insight into specific microbiome compositional differences.

We confirm with SelEnergyPerm a significant association (combined-F = 33.59817, empirical p = 0.007) exists between the composition of microbes in the CSF and PIH/NPIH (Figure 2.5A) and identify a reduced log-ratio signature of 12 ratios between 13 taxa as being significantly associated with PIH/NPIH (Figure 2.5B). Random forest (RF) models were then used to understand the capability of this SelEnergyPerm signature for discriminating between disease statuses. Using 50 repeats of 10-fold cross-validation, we computed an Area Under the Receiver Operating Characteristic Curve (AUC) = 0.906 (0.879-0.935 95% CI) (Figure 2.5C). We emphasize, however, that the more complex RF models with all 1,596 pairwise logratios yielded a comparable AUC = 0.892 (0.860-0.923 95% CI) (Figure 2.5C). For comparison, microbiome analysis carried out in Paulson et al. revealed *Paenibacillus* alone to be important for predicting PIH; but here using only the relative abundance of *Paenibacillus* with RF we observed an AUC = 0.830 (0.792-0.867 95% CI), significantly lower than that obtained using the logratios identified by

SelEnergyPerm. Combined, these results suggest the parsimonious SelEnergyPerm-derived log-ratio signature retains important disease interactions and better discriminates PIH vs. NPIH when compared to *Paenibacillus* alone.

To understand how the logratios in our signature work together to explain differences between the CSF microbiome of PIH vs. NPIH patients, we apply principal component analysis (PCA) (Figure 2.5D) and analyzed the means of the logratios. Examination of the distribution of samples shows the greatest separation between disease groups occurs along PC1 (Figure 2.5D), which explains 78.48% of the total variation. This separation indicates positive (negative) scores along PC1 are associated with NPIH (PIH) samples. Analyses of the logratio mean between groups for each logratio in the SelEnergyPerm signature indicate the abundance of *Paenibacillus* is significantly increased (Figure 2.5E) relative to taxa it is connected to (Figure 2.5B). Moreover, RF variable importance indicates the logratio between *Paenibacillus* relative to *Pseudomonas* to be most important for distinguishing between disease statuses. Indeed, analysis of Principal Component 1 loadings (Figure 2.5E) reveals increased abundance of *Pseudomonas* relative to *Paenibacillus* results in positive loadings (NPIH associated) along Principal Component 1. Overall, our results confirm, using pairwise logratios derived from SelEnergyPerm, the importance of *Paenibacillus* in PIH. Additionally, we show the interaction between the abundance of *Pseudomonas* relative to *Paenibacillus* is particularly important whereby more Pseudomonas is characteristic of NPIH and more *Paenibacillus* is associated with PIH.

2.6.5 Association between delivery mode and infant gut microbiome composition

Bokulich et al. monthly profiled the gut microbiome of infants with either a vaginal or cesarean delivery mode using 16S sequencing for the first two years of life [Bokulich et al., 2016]. The processed dataset was retrieved from the Qiita repository using study ID 10249 [Gonzalez et al., 2018]. Specifically, we extracted samples during the first 3 months of life, totaling 230 samples from 63 infants (Cesarean = 25, Vaginal = 38). We aggregated OTUs to the family-genus level which resulted in 140 distinct taxa (9,730 logratios) present in at least 10% of all samples by month. Here we apply SelEnergyPerm to determine if the gut microbiomes are different between the delivery modes of infants at any of the first 4 monthly time points collected (0–3 months). Secondarily, we studied our reduced logratio signatures to understand gut microbiome compositional differences between delivery modes at time points where significant differences were detected.

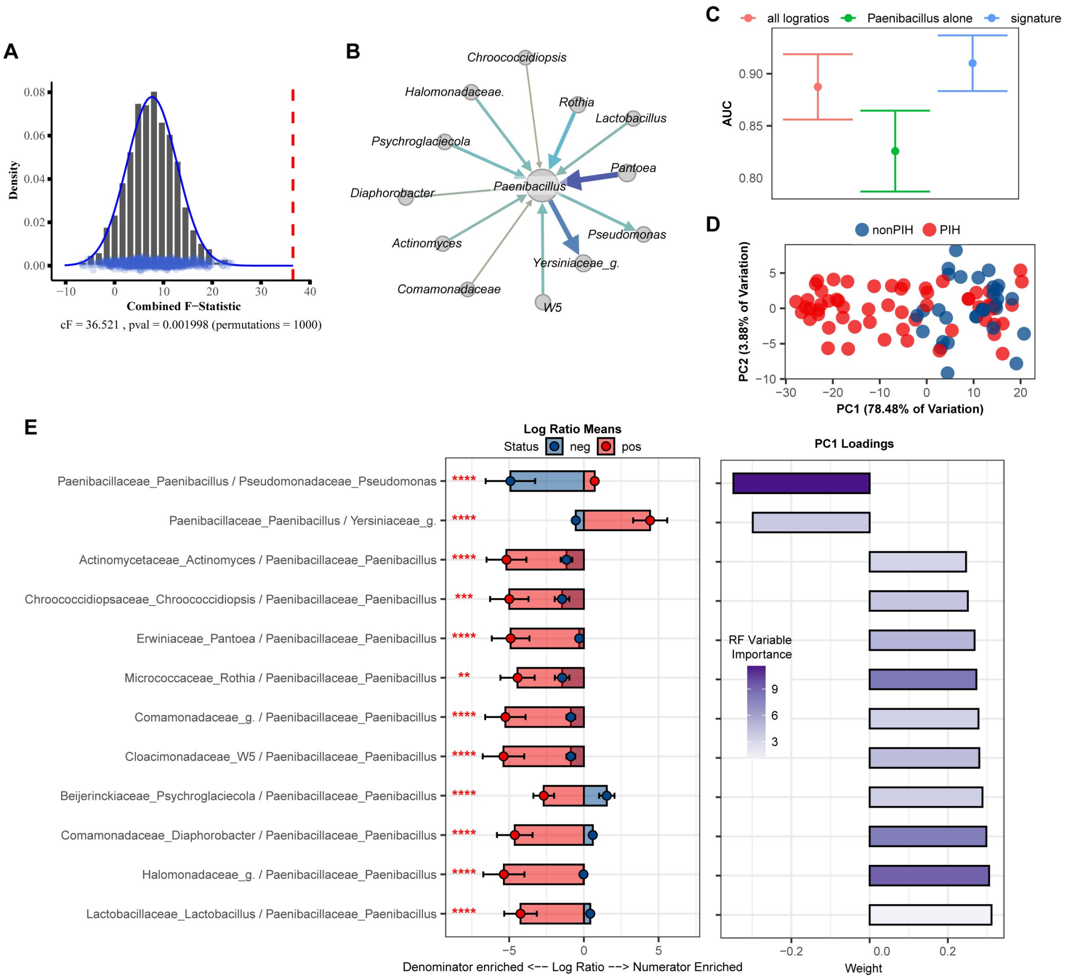

Figure 2.5: SelEnergyPerm case study examining the association between Ugandan infant's cerebrospinal fluid microbiomes and post-infectious hydrocephalus using 16S data. A. SelEnergyPerm permutation test results displaying the null distribution of the cF statistic (Histogram, Density, and Points) and the empirical cF statistic (dashed red vertical line). B. Random forest (RF) importance weighted directed logratio network (edges point from numerator to denominator) of the SelEnergyPerm selected signature (nodes = taxa, node size = weighted degree, edges = logratio, edge width/color = RF variable importance). C. ROC (Receiver Operator Characteristic) comparisons of disease status discrimination using RF. Models were trained with repeated (r = 50) 10-fold cross-validation using either the SelEnergyPerm Signature, all logratios, or *Paenibacillus* alone. D. Principal component analysis using the SelEnergyPerm signature. E. (Left) logratio means comparison (NPIH/PIH) of each logratio included in the SelEnergyPerm signature. (Right) Loading weights of the first principal component. Significance codes (*, **, ***, ****) indicate BH corrected p-value $<$ (0.05, 0.01, 0.001, 1e-4, 0) for NPIH versus PIH Wilcoxon Rank Sum Test. For the logratio means, positive values indicate numerator more abundant than the denominator and negative values indicate the denominator is more abundant numerator. Error bars indicate the 95% CI of the mean. Notably, error bars that do not span 0 indicate numerator/denominator is on average more abundant than the opposite.

Applying SelEnergyPerm to each time point with restricted permutation testing to account for repeated host microbiomes within a collection month and correcting for multiple comparisons using the Benjamini-Hochberg (BH) procedure, we found significant differences in the composition of the gut microbiomes between delivery modes during the collection periods in months 0–2 (Figure 6A). Notably, restricted permutation testing with PERMANOVA and ANOSIM using all taxa pairwise logratios (PLR) failed to detect differences between the gut microbiomes at $\alpha = 0.05$. Similarly, when using Partial Least Squares Discriminate Analysis (PLS-DA) with repeated cross-validation stratified by both delivery mode and host, we observed the AUC of the SelEnergyPerm-derived signatures to be higher across all time points compared to models trained using all PLR (Figure 6B). We next used the reduced log-ratio signatures and their PLS-DA variable importance scores to better understand which taxa are most important for discriminating between delivery modes. Indeed, aggregating to the family level for ease of interpretation, we found during months 0 and 1 that *Bacteroidaceae* were top contributors to compositional differences (Figure 6C). This pattern changed during month 2 where *Rikenellaceae* taxa were most important for discriminating between delivery modes (Figure 6C). Finally, to understand the direction of these differences (i.e., for a given logratio, is the numerator more abundant than denominator or vice-versa between groups), we analyze the directed log-ratio means network of the SelEnergyPerm signature relatively (i.e., taxa A more/less abundant than taxa B) between delivery modes (Figure 6D). Specifically, given the hub-spoke character of the observed network, with a single highly connected and central node in the directed maximum spanning tree formed by the SelEnergyPerm signature, we can see month 0 is dominated by differences between logratios that include *Lachnospira* and *Bacteroides*, which are more abundant relative to their network of taxa connections for infants with a vaginal delivery mode whereas the opposite is true for infants with a cesarean delivery mode. For month 1, *Bacteroides* are observed to be more abundant relative to its network of taxa connections for infants with a vaginal delivery mode. The opposite is true for infants with a Cesarean delivery mode where *Bacteroides* are less abundant within its network of taxa connections. Finally, for month 2, *Rikenellaceae* taxa can be observed to be more (less) abundant relative to both *Clostradiacea* and *Proteus* taxa for infants with a vaginal (Cesarean) delivery mode.

2.6.6 Association between abnormal fecal calprotectin levels and gut microbiome

Here we apply SelEnergyPerm to analyze WGS microbiome data from the integrative human microbiome project [Proctor et al., 2019], a longitudinal study designed to uncover interactions between disease and human-associated microbial communities. Specifically, using the inflammatory bowel disease

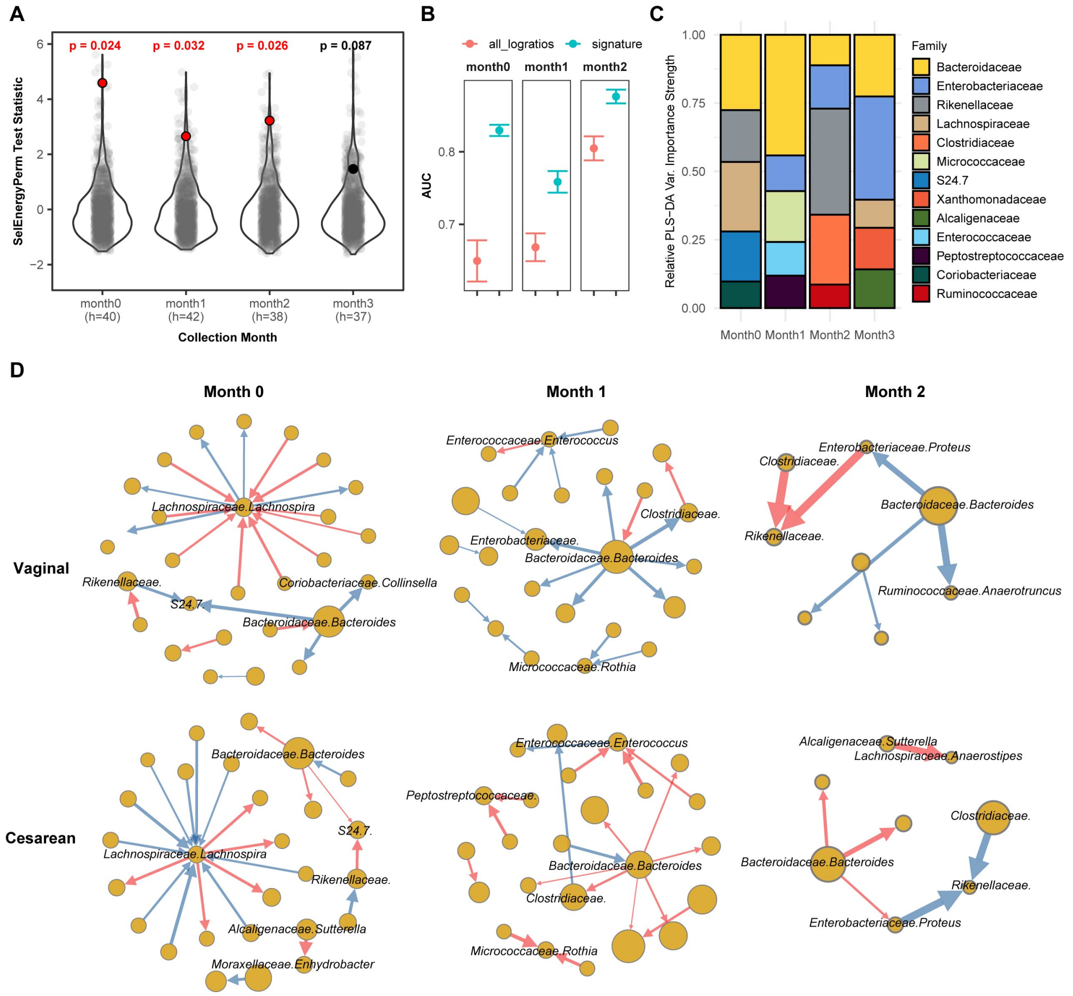

Figure 2.6: SelEnergyPerm case study examining the association between delivery mode and the gut microbiome composition of infants over the first three months of life using 16S data. A. SelEnergyPerm permutation test (permutations = 1000) results displaying the null distribution of the test statistic (violin and grey points) and the empirical test statistic (red if significant, black otherwise) with Benjamini-Hochberg corrected p-values. Test statistics values were z-score scaled (by Collection Month) for ease of visualization. B. AUC comparisons of delivery mode discrimination using PLS-DA. Models were trained with repeated (r = 20) 5-fold stratified (delivery mode and host) cross-validation using either the SelEnergyPerm signature or all logratios. Points represent the mean AUC and error bars indicate the 95% CI. C. Relative taxa strength by family measuring the importance of each taxon for discriminating between delivery modes across each collection time point. Relative strength was computed using the top 5 nodes derived from the PLS-DA variable importance weighted logratio networks across each collection time. D. Directed (edges point from numerator to denominator) network of the SelEnergyPerm-derived signature by month and delivery mode weighted by the absolute logratio means (nodes = taxa, node size = mean strength, edge = logratio, edge width = logratio mean, red edges = negative logratio mean (incoming node more abundant), blue edges = positive logratio mean (outgoing node more abundant)).

(IBD) part of the integrative human microbiome project study, we tested whether there exists an association between the gut microbiome composition and abnormal levels of fecal calprotectin, a protein marker of intestinal inflammation [Proctor et al., 2019]. Processed microbiome data were extracted from the Inflammatory Bowel Disease Multiomics Database [Lloyd-Price et al., 2019] resulting in 399 samples (93 individuals) reporting fecal calprotectin levels that were above 120 (abnormal; n = 190) or below 50 (normal; n = 209). There were 122 species identified (7,381 logratios) as being present in at least 10% of all samples.

Using restricted permutation testing, accounting for the order of visit and diagnosis of Ulcerative Colitis, Crohn's Disease, or non-IBD, SelEnergyPerm identified a significant association (combined-F = 92.507, p = 0.000999, 1000 permutations) between the composition of the gut microbiome and abnormal levels of fecal calprotectin in corresponding stool samples (Figure 2.7A). Notably, both ANOSIM and PERMANOVA with restricted permutation designs using all pairwise logratios (PLR) also detected this association. To assess whether the associated SelEnergyPerm log-ratio signature (25 logratios between 31 species) retained enough information to adequately discriminate between levels of fecal calprotectin, we estimated the discriminatory ability both using the reduced signature and using all PLR. Using repeated cross-validation with PLS-DA we found the SelEnergyPerm signature (AUC = 0.829, 0.803 – 0.854 95%CI) to have comparable performance to PLS-DA models trained using all logratios (AUC = 0.833, 0.803 – 0.862 95%CI) (Figure 2.7B). Examination of the latent space projection of a final PLS-DA model fit using the SelEnergyPerm signature reveals strong separation between individuals with normal vs. abnormal fecal calprotectin levels (Figure 2.7C). A directed log-ratio network of the SelEnergyPerm signature weighted by PLS-DA variable importance shows logratios involving *Dialister invisus*, *Streptococcus salivarius*, *Bacteroides fragilis*, *Escherichia coli*, and *Blautia wexlerae* to be most important for discriminating between levels of fecal calprotectin (Figure 2.7D). Interestingly, stratifying the log-ratio signature by diagnosis reveals both shared (significant between diagnosis differences across all groups) and distinct (significant between diagnosis differences among a single group) gut microbiome differences (Figure 2.7E). Particularly increased abundance of *Dialister invisus* relative to *Bacteroides ovatus*, *Intestinimonas butyriciproducens*, and *Anaerotignum lactatifermentans* was observed to be associated with abnormal fecal calprotectin independent of diagnosis.

2.6.7 Association between the infant gut microbiomes and allergen sensitization

In this case study, we apply SelEnergyPerm to WGS gut microbiome data from the DIABIMMUNE study [Vatanen et al., 2016]. The focus of this longitudinal study was to characterize interactions between the immune system and the gut microbiome in the context of autoimmunity and allergy. Specifically, the gut microbiomes of infants from Finland, Russia, and Estonia were profiled monthly during the first 3 years of life. Here we apply SelEnergyPerm to test if associations exist between allergy status and the composition of the gut microbiome at 6-month intervals during the first 2 years of life. Allergy status was defined as food allergy (FA) if the host reported an allergy to egg, peanuts, and/or milk at year 2 (non-FA otherwise). We extracted 646 samples from 192 infants (Russia = 53, Finland = 70, Estonia = 59) across 170 unique species (14,365 logratios).

Using restricted permutation testing to account for repeated host microbiomes and host country we applied SelEnergyPerm to each timeframe and corrected for multiple comparisons using the BH procedure. We found significant differences in the composition of the gut microbiomes between allergy status during both the first 6 months and the 6–12 month collection periods (Figure 2.8A). PERMANOVA and ANOSIM using all taxa PLR detected differences between the gut microbiome during the first 6 months of life but did not detect differences between the gut microbiomes during the remaining time frames at $\alpha = 0.05$ after correcting for multiple comparisons. This difference is further apparent when comparing the discriminatory ability between the SelEnergyPerm signature and all logratios. Using Partial Least Squares Discriminate Analysis (PLS-DA) with repeated cross-validation stratified by allergy status, host, and month, we observed the AUC of the SelEnergyPerm-derived signatures to be significantly higher across all time points when compared to models trained with all logratios (Figure 2.8B). Using the SelEnergyPerm log-ratio signatures and the corresponding PLS-DA variable important scores we next examine which taxa are important for discriminating between food allergy statuses later in life. Stratifying by month and selecting the top 5 species by strength (weighted degree) from our variable importance log-ratio network, we found *Clostridium ramosum*, *Streptococcus parasanguinis*, and *Bifidobacterium bifidum* to be major contributors to the DCV score between allergy status during the first 6 months of life (Figure 2.8C). However, for the 6–12 month period we found the abundance of *Clostridium hathewayi*, *Bacteroides dorei*, and *Haemophilus haemolyticus* to be major contributors to DCV (Figure 2.8C). A review of the log-ratio mean networks (Figure 2.8D) between allergy status during the first 6 months shows *Clostridium ramosum* is, in general, more abundant relative to species (node strength indicated by size) it is connected to in infants with FA vs. non-FA. Further, during the 6–12 month period we see more distinct differences in the log-ratio mean networks whereby *Bacteroides dorei* can be observed to be more abundant relative to species it is connected to in FA infants. We also observe *Clostridium hathewayi* to be more (less) abundant than the species it is connected to in infants with FA (without FA).

2.7 Discussion

We have here presented SelEnergyPerm, a group association testing framework for high-dimensional metagenomic data with sparse microbiome associations between groups. Our framework directly accounts for the compositional sample space imposed on these data as a result of technical variations in sample-wise library size. This is done by using embedded feature selection on a set of robust all pairwise logratios to improve the detection and interpretation of sparse signals hidden in these data. Each logratio is first ranked using our novel DCV score followed by the application of network methods and feature selection techniques to effectively select subsets of logratios. In tandem, these steps help to identify logratio signatures capable of explaining microbiome-derived phenotypic differences. Further, false discovery is properly controlled for by repeating the entire process with permuted labels using appropriate permutation test design (e.g., restricted design for longitudinal supervised data) for statistical significance [Ernst, 2004].

We assessed our method by conducting an extensive simulation study to rigorously benchmark performance relative to popular alternatives for both feature selection and association testing. Our simulation scenarios included data from both synthetic and empirical distributions where scenarios included settings with small/large location shifts embedded in sparse/dense signals and small/large location shifts with covariance differences embedded in sparse/dense association signals, carried out on both balanced and unbalanced sample designs. When compared to popular alternatives, we show our SelEnergyPerm feature selection approach is, overall, able to select fewer logratios, guarantee log-ratio subsets are independent, and better maximize between-group associations with relatively modest computational time requirements. Additionally, when compared to common association testing methods used in metagenomic studies, we show SelEnergyPerm can consistently detect associations better than or comparable with the alternatives in nearly all simulation settings tested. The better performance of SelEnergyPerm is most notable when sparse association signals are present.

Our demonstration of how SelEnergyPerm can be used to gain robust and unique biological insight was carried out in detail with data from 4 case studies. Particularly, we first apply SelEnergyPerm to test for sparse associations. After using SelEnergyPerm to confirm the statistical significance of associations, we extract SelEnergyPerm-derived log-ratio signatures and use traditional statistical techniques to better interpret and visualize (e.g. PCA, PLS-DA) how the microbiome is associated with the phenotype of interest.

Because multivariate effect sizes are not well studied, here we use AUC as a proxy of effect size between groups (i.e., AUC$= 1$ indicates perfect separation, and AUC $= 0.5$ for no separation). Most importantly, AUC used in this context indicates strength of association rather than out-of-sample predictive accuracy. In the first case study, SelEnergyPerm successfully detected a confirmed association between the composition of the microbiome in CSF and PIH/NPIH disease status in Ugandan infants using a reduced log-ratio signature (13 of 1,596 possible logratios). Further, we show given these data that our log-ratio signature can discriminate between disease statuses and explain differences between infants to a greater degree than with a single feature or all pairwise logratios. In our second case study, SelEnergyPerm detected an association between delivery mode and the gut microbiome composition in infants during the first 2 months of life and at the time 0 collection time. Notably, PERMANOVA and ANOSIM applied over the same time course with all logratios failed to detect this association. In the third case study, SelEnergyPerm detected an association between the composition of the gut microbiome and abnormal fecal calprotectin levels. Here we found our fecal calprotectin associated log-ratio signature (25 logratios) had a comparable discriminatory ability to the uninterpretable-due-to-size set of all 7,381 pairwise logratios, thus enabling easier biological interpretation. In the final case study, SelEnergyPerm detected and characterized associations between the microbiome composition in early life and the development of food allergy later in life.

Overall, our results demonstrate that SelEnergyPerm is a powerful framework for detecting sparse association under various scenarios. However, in the presence of heterogeneity of variance and/or unbalanced group designs — both of which are common enemies of multivariate association testing methods — the power of SelEnergyPerm was reduced, albeit to a lesser degree than the standard methods tested. Therefore, caution should be used when applying SelEnergyPerm in these settings. Additionally, in some scenarios with dense association signals, the performance of SelEnergyPerm was slightly reduced when compared to standard methods. While the power reduction was small, the enhanced interpretation from a smaller log-ratio signature may nevertheless outweigh the loss of power in some such settings.

Notwithstanding these limitations, SelEnergyPerm is the first method to our knowledge to fully utilize the pairwise logratio compositional approach in a group association testing framework for metagenomic data. Importantly, given the compositional sample space imposed on these data, where features are relative, our approach enables the discovery of associations using pairwise logratios which, by design, robustly interpret features relative to one another rather than alone. While the benefits of employing logratios are well documented, implementing and carrying out these analyses can be challenging and time consuming in practice. To this end, we developed an R package, SelEnergyPermR, with functions to perform the method presented in this paper. Additionally, our package enables rapid preprocessing of relative abundance data, calculation of all pairwise logratios, and multiplicative zero imputation. Our package also includes functions to simulate data from all scenarios presented in this work. Lastly, our approach adds to a small list of compositional methods for testing associations [Fernandes et al., 2014, Mandal et al., 2015a, Lin and Peddada, 2020] and is to our knowledge, the first compositional data method developed for sparse multivariate group association testing in metagenomic data. We also add to a small list of compositional approaches for feature selection [Susin et al., 2020]; however, unlike these other methods, our approach directly uses pairwise logratios which enables simple interpretation and may better elucidate taxa-taxa interactions through log-ratio network analysis. While not demonstrated explicitly here, SelEnergyPerm is also compatible with multi-class (>2 groups) group association testing (implemented in R package) . Future directions to usefully expand this methodology could focus on incorporating covariate information and extending the framework to longitudinal data.

In conclusion, we developed SelEnergyPerm to be a versatile group association testing method for detecting and understanding sparse associations in high-dimensional metagenomic data. We showed through rigorous simulation study with synthetic and real data distributions that SelEnergyPerm selects parsimonious subsets of independent logratios that better maximize between-group associations when compared to existing feature selection methods. Our simulation results also demonstrate SelEnergyPerm is significantly better at detecting sparse associations when compared to existing multivariate group association tests. Overall, SelEnergyPerm will enable researchers to robustly detect, characterize, and understand sparse associations in metagenomic data using novel log-ratio signatures. The SelEnergyPerm method is implemented in the R package SelEnergyPermR and is freely available on GitHub (https://github.com/andrew84830813/selEnergyPermR.git).

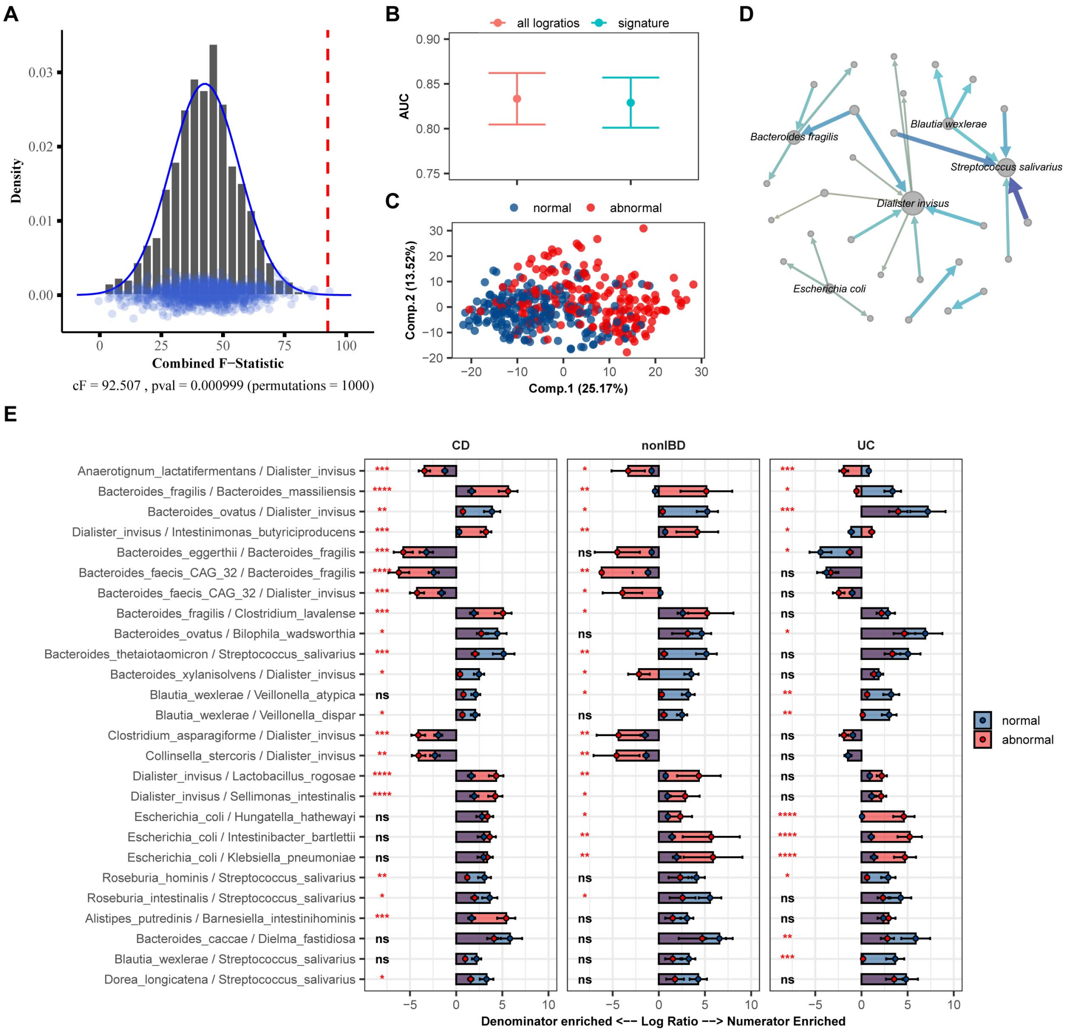

Figure 2.7: SelEnergyPerm case study examining the association between abnormal fecal calprotectin levels and the gut microbiome composition in nonIBD and IBD individuals using WGS data. A. SelEnergyPerm permutation test results displaying the null distribution of the cF statistic (Histogram, Density, and Points) and the empirical cF statistic (dashed red vertical line). B. AUC comparisons of fecal calprotectin level (Abnormal/Normal) discrimination using PLS-DA with 2 components. Models were trained with repeated (r = 20) 10-fold cross-validation using either the SelEnergyPerm signature or all logratios. Points represent the mean AUC and error bars indicate the 95% CI. C. PLS-DA latent space projection plot extracted from final PLS-DA model fit using the full dataset with the SelEnergyPerm signature. Points represent non-IBD or IBD samples. D. Directed network (edges point from numerator to denominator) of the SelEnergyPerm-selected log-ratio signature (nodes = taxa, node size = DCV strength, edges = logratio, edge width/color = PLS-DA Variable Importance). The top 5 taxa names by strength (PLS-DA Variable Importance) are displayed. E. logratio means comparison (normal/abnormal fecal calprotectin level) of each logratio included in the SelEnergyPerm signature stratified by Crohn's Disease (CD), Ulcerative Colitis (UC), and non-IBD individuals. Significance codes (ns, *, **, ***, ****) indicate BH corrected (within diagnosis) p-value < (Not Significant, 0.05, 0.01, 0.001, 1e-4, 0) for normal versus abnormal Wilcoxon Rank Sum Test. Error bars indicate the 95% CI of the mean. Notably, error bars that do not span 0 indicate numerator/denominator is on average more abundant than the opposite.

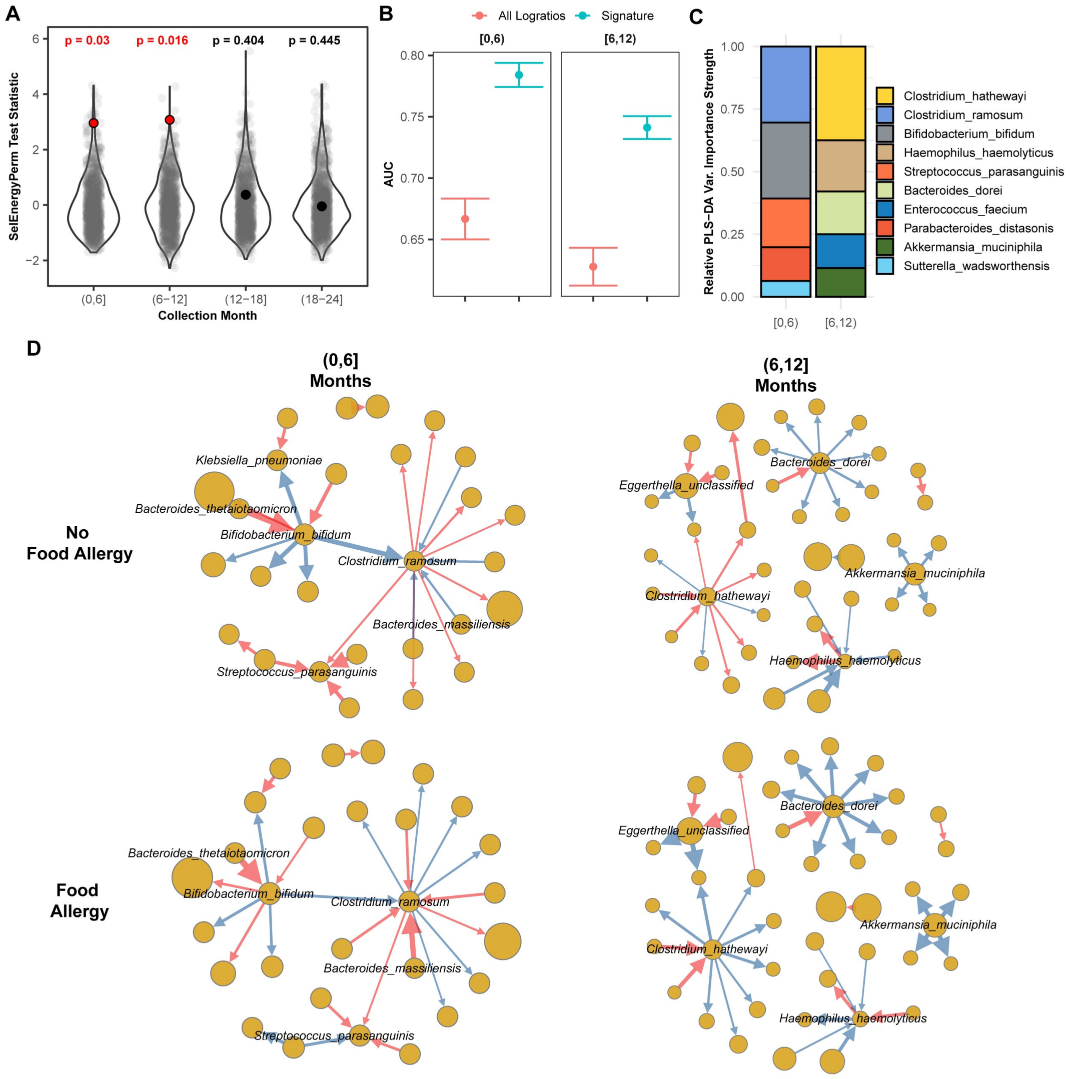

Figure 2.8: SelEnergyPerm case study examining the association between the gut microbiomes of infants in early life and the development of food allergy later in life. A. SelEnergyPerm permutation test (permutations = 1000) results displaying the null distribution of the test statistic (violin and grey points) and the empirical test statistic (red if significant, black otherwise) with Benjamini–Hochberg-corrected p-values. Test statistics values were z-score scaled by collection period to improve visualization. B. AUC comparisons of future food allergy development discrimination using PLS-DA. Models were trained with repeated (r = 20) 10-fold stratified (host and food allergy development) cross-validation using either the SelEnergyPerm signature or all logratios. Points represent the mean AUC and error bars indicate the 95% CI. C. Relative taxa strength by family measuring the importance of each taxon for discriminating between food allergy statuses later in life across each collection month. Relative strength was computed using the top 5 nodes derived from the PLS-DA variable importance weighted logratio networks across each collection month. D. Directed (edges point out from numerator to denominator) networks of the SelEnergyPerm-derived signature by collection period and food allergy development weighted by the absolute logratio means (nodes = taxa, node size = mean strength, edge = logratio, edge width = logratio mean, red edges = negative logratio mean (incoming node more abundant), blue edges = positive logratio mean (outgoing node more abundant)).

CHAPTER 3: DIFFERENTIAL COMPOSITIONAL VARIATION MACHINE LEARNING[2]

3.8 Introduction

Ubiquitous use of rapidly advancing metagenomic sequencing technologies are allowing researchers to uncover profound insights into precisely how alterations of microbial communities that live in and on the human body are associated with human disease. These promising technologies are being rapidly explored for use as non-invasive diagnostic and screening tools [Zackular et al., 2014, Schlaberg, 2020]. The success of such efforts relies on the identification of robust microbial biomarkers that are predictive of disease onset and/or progression. To this end, researchers are exploring novel ways to apply the latest advances in supervised machine learning methodology to unlock key biomarker signals across unique and costly metagenomic data [Marcos-Zambrano et al., 2021]. Fortunately, numerous publicly-available curated metagenomic datasets [Oliveira et al., 2018, Pasolli et al., 2017, Gonzalez et al., 2018] have helped researchers develop and test new methodologies on complex metagenomic datasets across diverse disease groups.

In order to discover generalizable metagenomic biomarkers, new prediction methods will need to address important statistical challenges related to high-dimensional metagenomic data. In particular, whole genome shotgun (WGS) and 16S ribosomal-RNA (16S) sequencing techniques utilized in metagenomic studies produce count data that have arbitrary limits on the total number of reads obtained for each sample by the instrument [Gloor et al., 2017]. Because of this, analysis of these data is limited to relative rather than absolute comparisons. Data with such constraints, formally known as compositional data [Aitchison, 1982], have a simplex sample space and, notably, important dangers arise when ignoring compositional constraints during analysis, including non-linear distances between samples as subsets of parts (e.g. taxa, metabolites) change, spurious correlations, and generalizability of models [Gloor et al., 2017]. Additionally, sparsity, discreteness, and distribution of total library sizes can influence conclusions and further complicate analyses [Lovell et al., 2020]. The importance of appropriately analyzing compositional metagenomic data has been

[2] This chapter was adapted from our manuscript(preprint) available on bioRxiv. The preprint citation is as follows: Hinton, A. L. & Mucha, P. J. Differential Compositional Variation Feature Selection: A Machine Learning Framework with logratios for Compositional Metagenomic Data. bioRxiv 2021.12.08.471758 (2021) doi:10.1101/2021.12.08.471758.

discussed extensively in Refs. [Gloor et al., 2017, Quinn et al., 2019] and the need for suitable normalization approaches to account for these challenges has been highlighted in Ref. [Weiss et al., 2017].

Supervised classification predictive modeling in metagenomics studies aims to classify disease based on learned patterns in microbial compositions. These models can then be deployed in non-invasive screening, diagnostic, or prognostic tests. It is not uncommon for researchers to train standard machine learning algorithms such as random forests, support vector machines, or LASSO regularized logistic models on untransformed relative abundance data and group labels (e.g. disease, phenotype) to identify microbial compositional differences between groups of interest. Indeed, standard approaches such as those in MetaML [Pasolli et al., 2016] ignore compositional data constraints intrinsic in relative abundance data; as a result, model interpretability and beyond-study generalizability may be severely limited. Fortunately, various log-ratio transformations have been proposed to overcome these challenges [Aitchison, 1982, Egozcue et al., 2003]. In particular, given signals from p parts, one seeks to properly describe the corresponding point on the $(p-1)$-dimensional simplex. Only the $(p-1)$-feature basis from the additive logratio (ALR) transformation and the spanning frame of all $\binom{p}{2} = p(p-1)/2$ features from the pairwise logratio (PLR) transformation provide both simple interpretation [Greenacre, 2019] and subcompositional coherence [Aitchison, 1982]. We note in particular that the ALR transformation has been shown to be an effective way to statistically analyze omics data [Greenacre et al., 2021] and that variable selection approaches that address compositional constraints on metagenomic data were important to the recently proposed selbal [Rivera-Pinto et al., 2018] and coda-lasso [Susin et al., 2020] methods.

We here introduce the differential compositional variation machine learning framework (DiCoVarML) for guided feature selection to efficiently train, robustly test, and flexibly interpret supervised classification models using robust additive logratio or pairwise logratio features. To more fully motivate the utility of this framework, we first demonstrate important generalization limitations of machine learning models that are trained directly with either relative abundance or centered logratio (CLR) transformed data: in a simple low-dimensional setting, even with proper cross-validation performance estimation, AUC estimates obtained for models trained with relative abundance or CLR transformed data may be unreliable. We then show through detailed simulation study and analysis of publicly-available metagenomic datasets that DiCoVarML provides significantly better classification performance than existing compositional feature selection approaches. To demonstrate its clinical utility, we then apply DiCoVarML to predict the onset of necrotizing enterocolitis in premature infants using fecal metagenomic data from the NICU NEC study [Olm et al., 2019], and we develop a novel meta-analysis covering 9 studies [Feng et al., 2015, Gupta et al., 2019, Hannigan et al., 2018, Wirbel et al., 2019, Thomas et al., 2019, Vogtmann et al., 2016, Yachida et al., 2019, Yu et al., 2017a, Zeller et al., 2014] to classify microbial differences between the gut microbiome and colorectal cancer.

3.9 Results

3.9.1 The DiCoVarML framework

DiCoVarML attempts to obtain an optimal set of logratios between parts (e.g., taxa, metabolites) for use as features in machine learning classification models. Broadly speaking, feature selection for compositional metagenomic data analysis can be applied at either (a) the parts level (selbal, clr-lasso, coda-lasso) or (b) the log-ratio level (ALR or PLR). In DiCoVarML, we utilize a novel multi-level feature selection approach to simultaneously identify a robust subset of logratios between a targeted number of parts. Our versatile approach importantly supports both targeted (number of target parts selected by expert) and untargeted (number of selected parts optimized empirically) feature selection. A targeted approach is especially useful when weighing trade-offs between predictive performance and cost for diagnostic test development, where limiting the number of parts to include in the final assay may be of particular concern. Moreover, depending on study priorities (high interpretability, high prediction), the DiCoVarML framework naturally allows for either interpretable but generally less accurate ridge-regression models or complex but generally more accurate average probability ensemble models for classification. Additionally, biomarker signatures discovered by DiCoVarML can be found via ALR (lower computational cost but possibly lower insight) or PLR (higher computational cost for possibly higher insight).

To robustly select features and estimate performance while minimizing overfitting, the DiCoVarML framework uses a double-nested (inner and outer) k-fold cross-validation learning schema to discover log-ratio signatures and estimate classification performance (Figure 3.1a & 1b). That is, the relative metagenomic dataset is randomly split in the outer cross-validation loop into k discovery and test partitions with appropriate stratification (by class, sample, study, etc.) given the study design (cross-sectional, longitudinal, repeated measures, etc.) (Figure 3.1a) with each fold held out and used once as a test set and the remaining folds used for discovery and training. On each partition, the targeted multi-level feature selection (tarMFS) method (Figure 3.1b) is applied, yielding user-selected classification performance measures (e.g., AUC, accuracy, AUPRC, etc.). To discover an appropriate number of parts (targeted or untargeted) and classification model (from a set selected *a priori* by the user; ensemble and ridge regression are considered here), an inner cross-validation loop is used (Figure 3.1b): the discovery set of the outer loop is further split into k_{inner} folds (here $k_{\text{inner}} = 2$) for training and validation. Each classification model (for numbers of parts listed in the array T) is evaluated on each partition using the targeted discovery mode (Figure 3.1c, see “Targeted multi-level feature selection” in Methods). The best performing model and number of parts (m_{max} and t_{max} in Figure 3.1b) is selected and used to train the discovery set model and classification performance is then obtained on the corresponding test fold. Overall classification performance is then estimated by averaging the performance across discovery-test folds (Figure 3.1a). Depending on study objectives and computational resources, the entire process can be repeated under different random splits to better estimate out-of-study classification performance.

3.9.2 Beyond-study generalization limitations of untransformed data

Using a simple three-part composition (p, q, r) toy dataset, we demonstrate here that machine learning models trained with relative abundance or CLR-transformed count data can easily fail to generalize beyond the available training samples, but that nevertheless the use of additive (ALR) or pairwise logratios (PLR) succeeds in such settings. In particular, we demonstrate this in scenarios with and without feature selection. We simulate two distinct classes of data across two partitions (train, test) with a common decision boundary within the simplex. While the train and test data partitions share decision boundaries in our simulated data, the deliberate geometric separations between the two simulated partitions are designed to represent differences as might arise from variabilities in study design, sample preparations, different instruments, noise from measurement error, or other cross-study differences. The three-part composition yields a two-dimensional simplex sample space, with the relative proportions of the three parts visualized in the ternary diagrams of Figure 3.2.

We first evaluated a scenario without feature selection where two classes of data are simulated from the additive logistic normal distribution (see Methods) with a single noisy part (Figure 3.2a). In this scenario, the decision boundary of the (p, q, r) components is an isoproportional $p/r = 1$ line in the simplex. Using a random forest model trained directly on the three-part relative abundance (proportions) data, we observed a mean cross-validated AUC = 0.976 in the train set; however, after applying this trained model to the test set, the generalization performance on the test set drops to AUC = 0.505. As observed clearly in the Figure, the random forest model trained on relative abundances failed to learn the correct decision boundary except in the area immediately local to the training set. Similarly, the corresponding model trained on the CLR transformation of the three-part composition yielded a train set AUC = 0.977 but test set AUC = 0.505. In

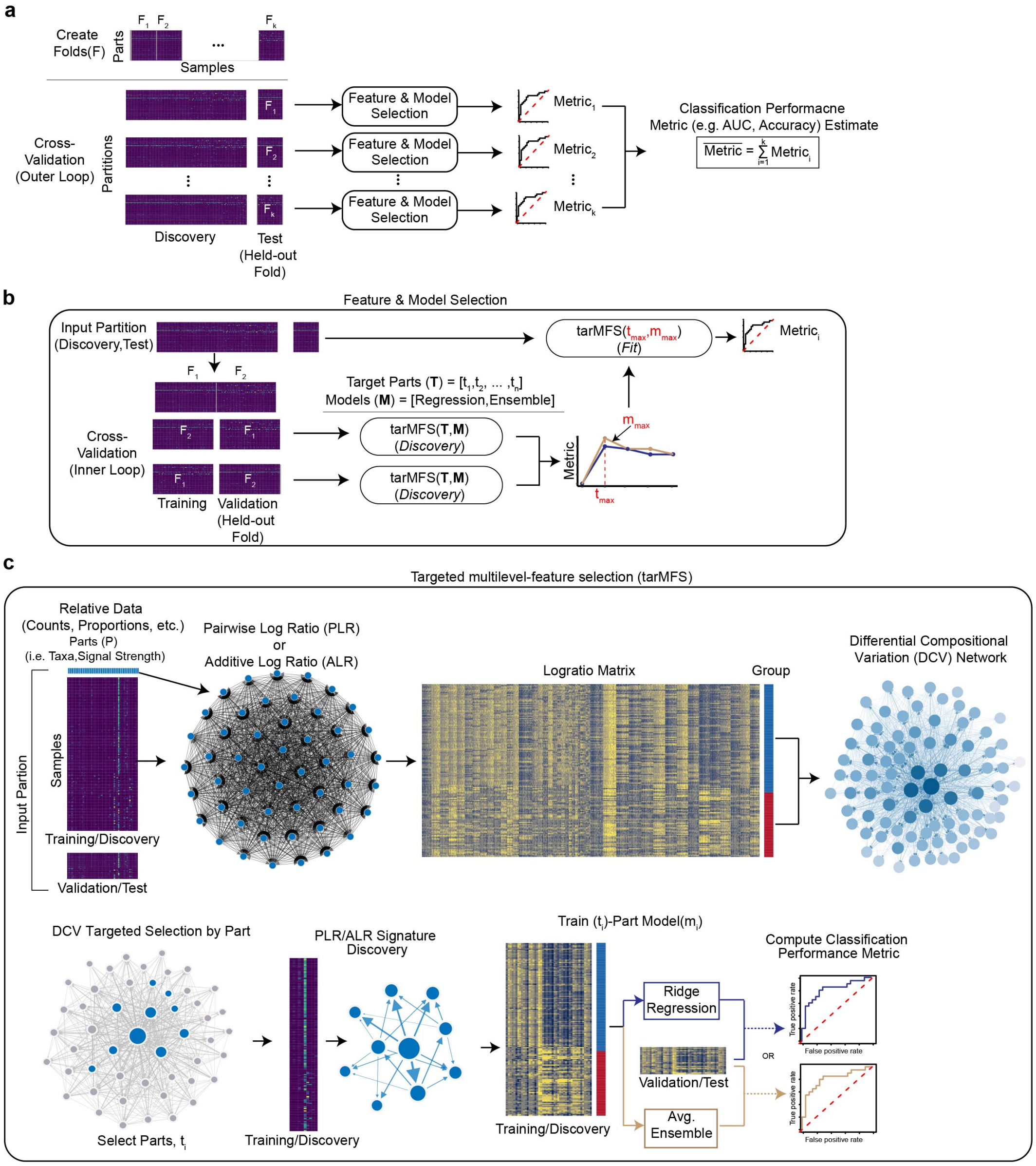
a
Create Folds(F)
Parts
F1 F2 ... Fk
Samples
Cross-Validation (Outer Loop)
Partitions
Discovery
Test (Held-out Fold)
Feature & Model Selection
Metric1
Metric2
Metrick
Classification Performacne Metric (e.g. AUC, Accuracy) Estimate
Metric = Σ Metrici
b
Feature & Model Selection
Input Partition (Discovery,Test)
tarMFS(tmax,mmax) (Fit)
Metrici
Target Parts (T) = [t1,t2, ... ,tn]
Models (M) = [Regression,Ensemble]
Cross-Validation (Inner Loop)
Training
Validation (Held-out Fold)
tarMFS(T,M) (Discovery)
mmax
tmax
Metric
c
Targeted multilevel-feature selection (tarMFS)
Relative Data (Counts, Proportions, etc.)
Parts (P) (i.e. Taxa,Signal Strength)
Input Partion
Samples
Training/Discovery
Validation/Test
Pairwise Log Ratio (PLR) or Additive Log Ratio (ALR)
Logratio Matrix
Group
Differential Compositional Variation (DCV) Network
DCV Targeted Selection by Part
Select Parts, ti
PLR/ALR Signature Discovery
Training/Discovery
Train (ti)-Part Model(mi)
Ridge Regression
Validation/Test
Avg. Ensemble
OR
Compute Classification Performance Metric
True positive rate
False positive rate

Figure 3.1: The differential compositional variation machine learning (DiCoVarML) framework for high dimensional compositional metagenomic data. (a) High level overview of the DiCoVarML framework showing the outer cross-validation procedure for estimating out-of-sample performance including the data partitioning, feature/model selection, and classification performance metric estimation. (b) Overview of the partition-specific feature and model selection process showing how targeted multilevel-feature selection (tarMFS) is used in the inner loop to select the number of parts and the model to then be used in the outer loop to estimate classification performance. (c) Overview of the nested targeted-multilevel feature selection method for identifying key parts, logratio signatures, and estimating classification performance.

contrast, training the random forest model with all pairwise logratios (PLR) leads to a train set AUC = 0.971 with test set AUC = 0.980, demonstrating accurate beyond-sample generalization from using pairwise logratios. In contrast to these stark performance differences, we note that LASSO logistic regression using any of these considered data transformations performs well on this simple, low-dimensional isoproportional boundary scenario.

We next examined a scenario with feature selection on data simulated on opposite sides of a compositional line decision boundary specified by a $[0.6, 0.3, 0.1]$ leading compositional vector through the barycenter (see “Compositional line decision boundary” in Methods). In this scenario, we trained and tested a regularized logistic model on each data transformation to evaluate its generalization properties (Figure 3.2b). For both relative abundances and CLR transformed counts we obtained train set AUC = 0.541 and test set AUC = 0.501, indicating significant under fitting. That is, in each of these cases it appears that the regularization eliminated the degrees of freedom needed to reconstruct the true logratios that combine to form the simulated decision boundary. In contrast, when working with the full frame of all pairwise logratios, the regularized regression achieves (empirical) train set AUC = 1 and test set AUC = 1. We additionally note that all of the considered data transformations generalize poorly when random forest models are trained without feature selection in this scenario, emphasizing the importance of feature selection.

Whereas the use of either pairwise or additive logratios generalizes well beyond the training sample, we note with these simple illustrative scenarios that there are important beyond-study generalization dangers when using either relative abundances or CLR-transformed data to train machine learning models. Importantly, these limitations are independent of the number of dimensions and can be present (or even expected) in high-dimensional metagenomic data. Notably, the ability of a classification model to generalize true patterns beyond the study immediately at hand is essential both to understanding how parts contribute to a classification and to developing robust diagnostic assays.

3.9.3 Classification performance evaluation using synthetic data

To understand how untargeted DiCoVarML with PLR, ALR, or hybrid signatures compares to existing compositional approaches with feature selection (selbal, clr-lasso, coda-lasso) we generated synthetic counts mimicking WGS and 16S binary-class data (see Methods) and compared the classification performance of each method (Figure 3.3). In particular, we studied three scenarios with increasing signal density (number of associated parts) from 2% to 50% of all taxa (WGS = 270, 16S = 124). For each scenario, we increased the percent mean difference (mean shift) between associated taxa from 18% to 30% (see Methods). As seen in Figure 3.3, for the 16S data with 2% signal sparsity, clr-lasso and selbal performed better than DiCoVarML at the 18% mean shift level, comparably at the 24% level, and slightly worse at the 30% level.

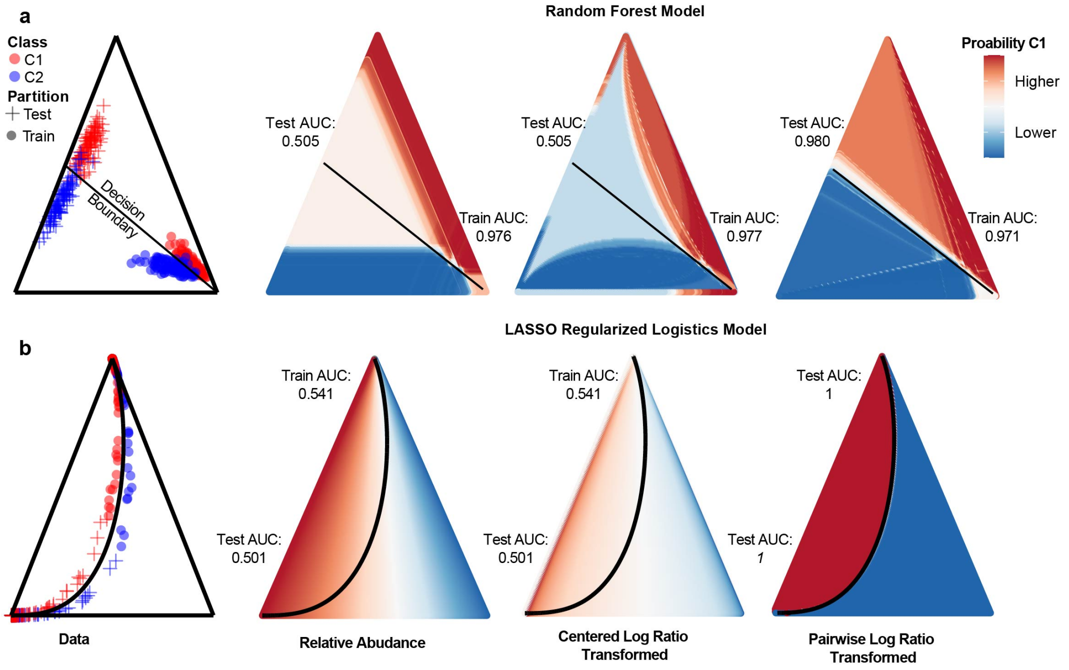

Figure 3.2: Generalizability limitations from training machine learning models with relative abundance or centered logratio data. Ternary diagrams visualize compositional data from three component parts in barycentric coordinates. For each scenario (row), data were simulated from two classes (C1, C2) with distinct data distributions for training (circles) and test (+ symbols) partitions in the first column on the left. The corresponding global decision boundaries are shown as solid lines for each scenario. Training set and test set AUC are shown for models trained from relative abundance (second column), centered logratios (third column), and all pairwise logratios (right column). Prediction probabilities for class C1 for each model fit to the training partition are indicated by shaded coloring throughout the simplices. (a) Random forest models fit to normally-distributed data on the 2-simplex with an isoproportional decision boundary. (b) LASSO regularized logistic model fit to data separated by an $\alpha \odot \{0.6, 0.3, 0.1\}$ compositional line (through barycenter) decision boundary.

As signal density increased, DiCoVarML outperformed the other methods and maintained consistent performance across all signal sparsity settings. Notably, for existing methods on the 16S data, we observed decreasing performance as the signal density increased. These findings are also consistent with simulation results reported in Ref. [Susin et al., 2020]. Notably, coda-lasso performed poorly across all 16S scenarios tested. For the WGS data, similar trends were observed with existing methods (including coda-lasso) performing better or similar to DiCoVarML at the lower signal sparsity levels but worse as the signal density increased. For existing methods on WGS data, we again observed large reductions in performance as signal density increased. Moreover, for DiCoVarML signatures on WGS data, we note a general but smaller reduction performance as signal density increases. Notably, ALR-derived signatures performed worse than PLR and Hybrid signatures in complex WGS data but similarly on less complex 16S data. From this, our simulation results demonstrate DiCoVarML maintains consistent performance over a range of scenarios with 16S or WGS characteristics when compared to existing methods.

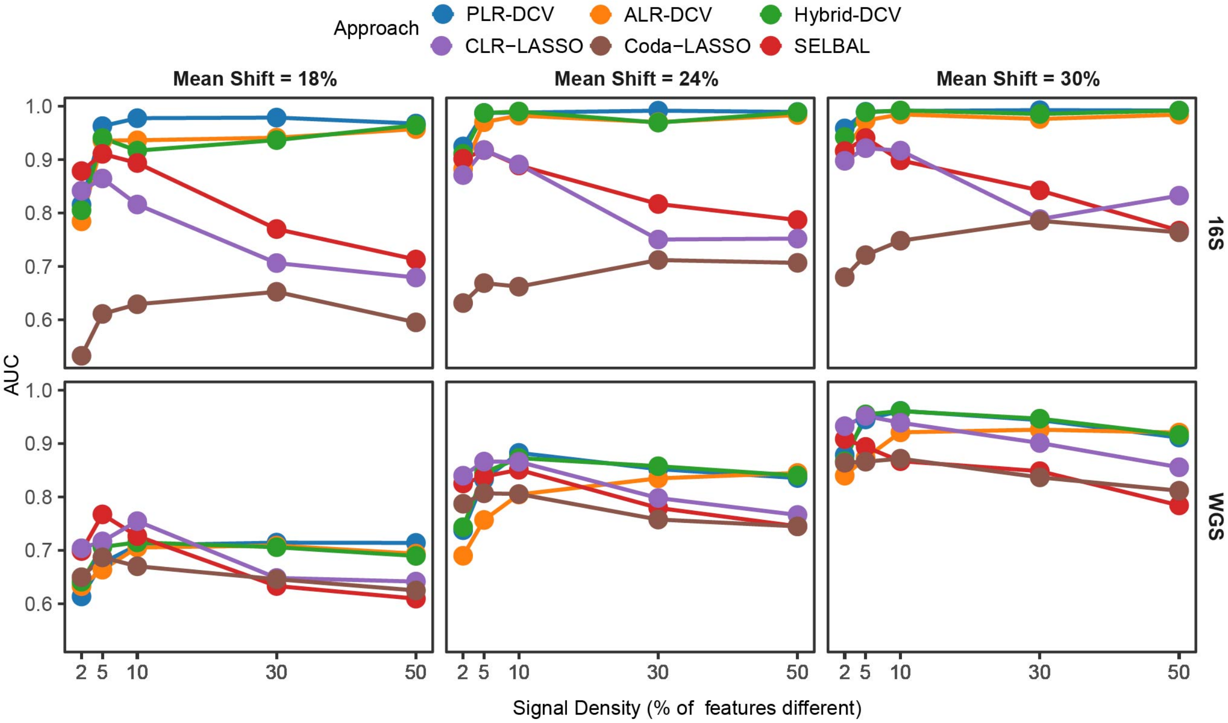

Figure 3.3: Classification performance comparison using synthetic data distributions. Mean binary classification AUC (5 repeats x 2-fold cross validation) for each method applied to simulated two-class data (n = 100 per class) corresponding to 16S (top row; 165 features) or WGS (bottom; 270 features) characteristics. Signal density (x-axis) measures the percentage of total parts that are different between classes. Columns represent effect size (percentage of between-class mean differences) of each signal feature.

3.9.4 Classification performance evaluation using real metagenomic data

We next compared the binary classification performance of untargeted DiCoVarML signatures to existing approaches using real case-control 16S and WGS gut microbiome datasets. To better assess the versatility of each approach, we tested classification performance in eight unique disease settings across eight cohorts from four publicly-available data sources (see Methods). Using a paired design, we trained and tested each approach on the same partitions using 15 repeats of 2-fold cross-validation. For ease of interpretation here, we characterize paired mean AUC differences ($\Delta = \text{AUC}_{DiCoVarML} - \text{AUC}_{existing}$) between approaches as: “modest” $\in (0.00, 0.01]$, “moderate” $\in (0.01, 0.05]$, and “large” > 0.05.

For 16S datasets, DiCoVarML signatures achieved the highest mean AUC in all four datasets tested (Figure 3.4A) when compared to existing methods. Particularly, ALR signatures achieved the highest AUC on three of four 16S datasets with the Hybrid signature doing best on the remaining dataset. To understand the magnitude and significance of differences in AUC between the existing approaches and DiCoVarML, we used Wilcoxon signed rank tests with a significance level $= 0.05$ (Figure 3.4B). Large mean differences with selbal were observed in all datasets tested across all DiCoVarML signature types. We observed moderate differences with coda-lasso for the CDI, NAFLD, and the Crohn's classification tasks. Notably, for the HIV classification task, we only note significant moderate differences in AUC for the signature obtained by DiCoVarML with the Hybrid approach. We observed modest to moderate differences compared to clr-lasso on the CDI/Crohn's classification task and large differences in AUC on the NAFLD and HIV classification tasks for all DiCoVarML signatures.

For WGS datasets, DiCoVarML signatures again achieved the highest mean AUC in all four datasets tested (Figure 3.4B), with PLR (2 of 4) and Hybrid (2 of 4) signatures achieving the highest mean AUC. Examination of mean differences compared to selbal revealed significant large mean differences in AUC across all datasets tested. For both coda-lasso and clr-lasso we observed significant yet modest to moderate differences for the cirrhosis (Cirr) and soil-transmitted helminth (STH) classification tasks and large significant differences for the Colorectal Cancer (CRC) and Schizophrenia (SCZ) tasks. Overall, our findings from real datasets demonstrate DiCoVarML signatures significantly outperform existing compositional methods for classifying disease using metagenomic data. In addition to achieving better classification performance, DiCoVarML signatures are sparse and easily interpretable. Furthermore, cross-validated performance estimates obtained from models trained with ALR or PLR are more robust and better representative of out-of-sample performance.

3.9.5 DiCoVarML predicts onset of NEC in preterm-infants

In this case study, we apply untargeted DiCoVarML with PLR and ridge regression to predict the onset (> 7days) of necrotizing enterocolitis (NEC) in pre-term infants, using publicly-available fecal microbiome data from MicrobiomeDB [Oliveira et al., 2018] collected as part of the longitudinal NICU NEC study [Olm et al., 2019] (Figure 3.5). After data preprocessing (see Methods), the dataset contained 136 genera (384 unique taxa) across 902 (non-NEC = 779, NEC = 123) samples from 144 pre-term infants (non-NEC = 120, NEC = 24). First, we set out to understand if the microbiome compositions were predictive of future NEC. In order to unbiasedly evaluate the performance of classifiers in this imbalanced setting, we focus on AUC to assess performance across thresholds in a manner that is not biased to either the minority (NEC) or majority (nonNEC) group [Fawcett, 2006]. We estimated the classification performance with 20 repeats of 5-fold stratified (by NEC status and sample) cross-validation using DiCoVarML with ridge regression to obtain a

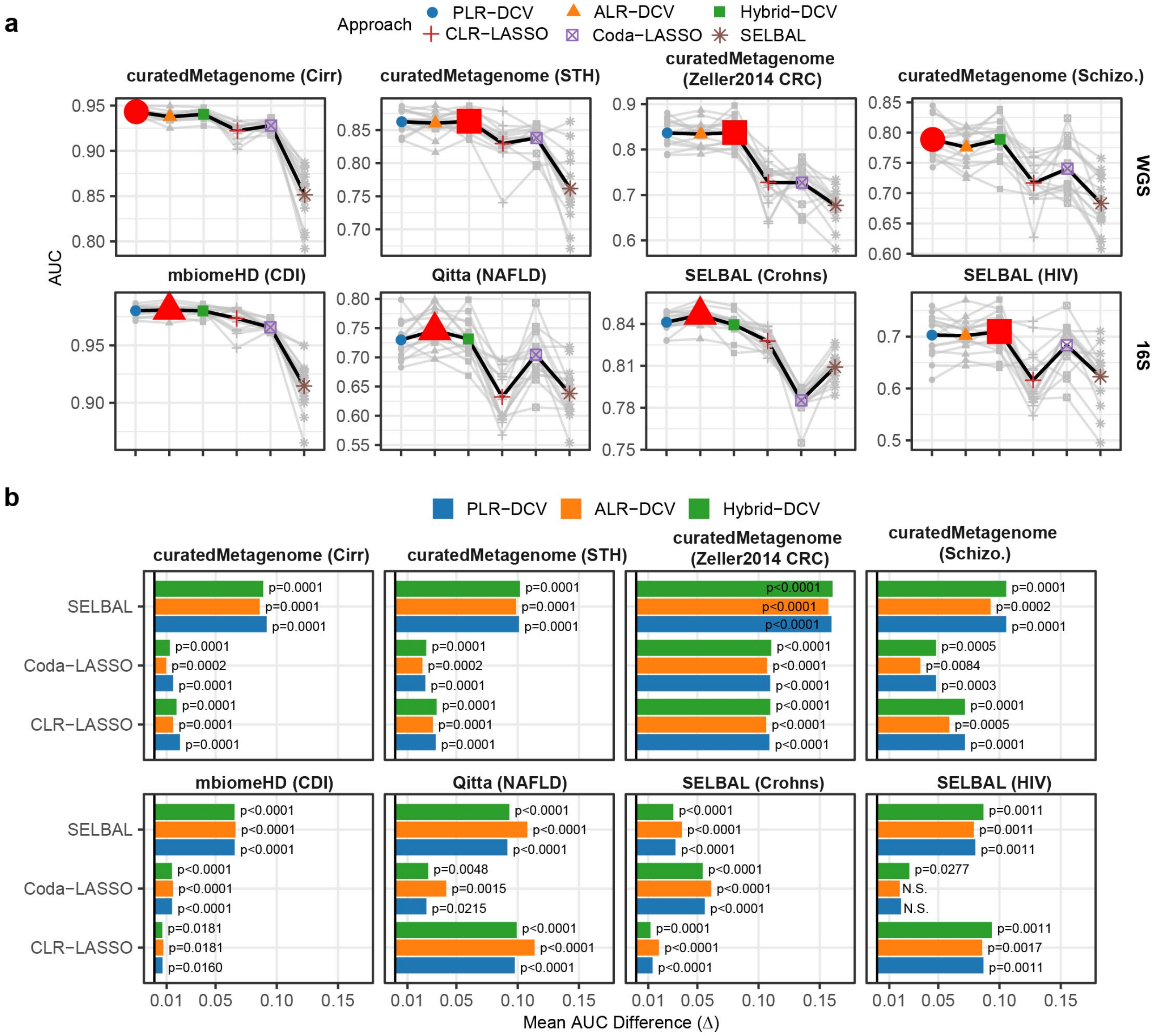

Figure 3.4: Classification performance comparison using real 16S and WGS datasets. Paired 15 repeats of 2-fold cross validation for each approach when applied to publicly-available case (shown) vs. control datasets. (a) AUC results for each method using WGS (top row) or 16S (bottom row) datasets. Grey points represent seed specific results. Grey lines connect paired seed specific AUC scores. Heavyweight points in color indicate overall mean AUC for each approach. Red point indicates approach with highest mean AUC. (b) Comparison of mean AUC differences between existing compositional approaches (y-axis) and DiCoVarML (bars). Results from Wilcoxon signed-rank test comparing AUC scores are shown. Benjamini–Hochberg-corrected p-values rounded to the nearest 0.0001 indicate high levels of significance in most cases ("N.S." indicates not significant at the 0.05 level).

mean AUC = 0.676 (95% CI: 0.655 – 0.696) (Figure 3.5A). To ensure that patterns learned by the classifier were non-random, we computed the AUC under permuted disease labels within the same folds, achieving a mean AUC = 0.616 (95% CI: 0.595 - 0.636). Using the Wilcoxon signed rank test, we confirmed the true classifier performed better than random ($p = 2.465 \cdot 10^{-5}$) indicating the classification model learned non-random disease specific association patterns (Figure 3.5B/5C). To further explore these patterns and identify a candidate set of biomarkers, we trained the final DiCoVarML guided ridge regression on the full dataset, uncovering a microbial network connecting 8 genera by 14 ratios that is predictive of future NEC (Figure 3.5B), revealing increased abundance of *Staphylococcus*, *Klebsiella*, *Cutibacterium*, and *Gemella* relative to this microbial network were associated with future NEC onset. Using the final trained model, we next examined the regression scores for each sample relative to the number of days until NEC onset (Figure 3.5D), showing classification performance was strongest 11–17 days before onset with a notable decrease in accuracy 18 or more days before onset. Finally, analysis of survival agnostic regression scores from the NEC positive samples revealed a significant association ($p = 4.566 \cdot 10^{-5}$) between the regression score and survival with higher scores associated with samples from infants that did not survive (Figure 3.5E).

3.9.6 DiCoVarML reveals association between CRC and gut microbiome composition

Lastly, we demonstrate the targeted DiCoVarML with ensemble modeling approach in a novel 11-cohort meta-analysis to classify fecal microbiome samples as colorectal cancer (CRC) vs. control. Specifically, we demonstrate how targeted (T = 50) DiCoVarML can be used to identify a predictive PLR signature (between species). Using the curatedMetagemic R package [Pasolli et al., 2017] we compiled and processed case-control WGS data from 11 studies (see Methods): keeping only samples with at least 10^6 total reads, our final dataset included 1,305 samples (CRC = 653, control = 652). To estimate CRC vs. control classification performance we used 252 repeats of stratified (by dataset) cross-validation where, for each partition, both splits (5 and 6 of the 11 total datasets) were used once for training and testing. From this, we observed a mean AUC = 0.795 (95% CI: 0.791 – 0.798) which was significantly ($p < 2.22 \cdot 10^{-16}$) higher than the mean AUC = 0.498 (95% CI: 0.493 – 0.503) obtained with permuted labels (Figure 2.6A), confirming non-random predictive patterns. We next used DiCoVarML to infer, from the full dataset, a targeted 50 species PLR signature predictive of CRC. In doing this, we identified 259 logratios between the targeted 50 species as being important for classification (Figure 2.6B). To understand the model predictions, we computed ensemble model scores (see Methods), revealing higher scores among CRC samples with control samples generally having lower scores (Figure 2.6C). After simplifying the log-ratio network, we next examined the multivariate association of individual taxa to each group (CRC/control) (Figure 2.6D), revealing *Peptostreptococcus stomatis*, *Gemella morbillorum*, and *Fusobacterium nucleatum* as the top three taxa associated with CRC, with increased abundance of these species within the microbial network (Figure 2.6B) associated with higher ensemble model scores (Figure 2.6C). Likewise, we found *Alistepes*

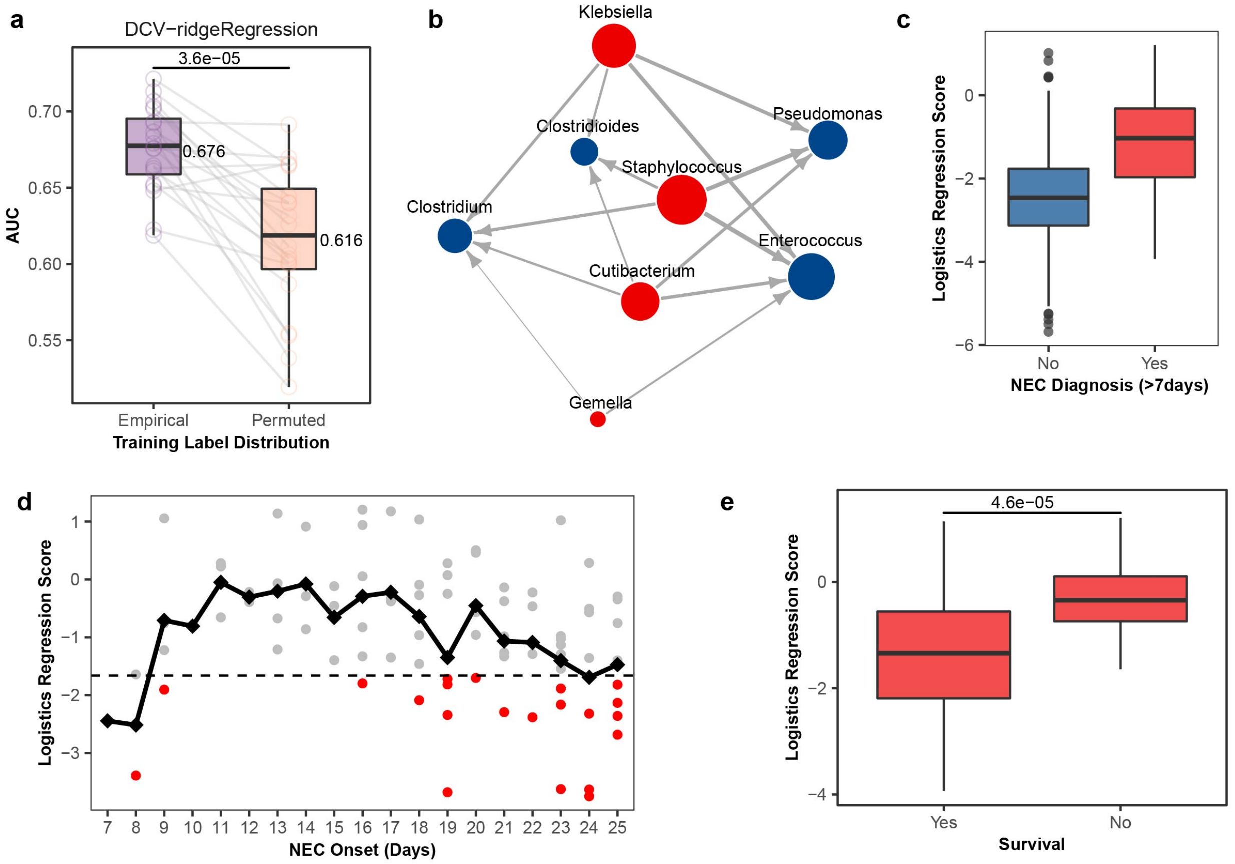

Figure 3.5: Predicting NEC onset (> 7 days) in premature infants using gut microbiome composition in the first few months of life using untargeted DiCoVarML with ridge regression applied to metagenomic profiled fecal samples (n=1,100) from the NICU NEC study. (a) AUC (paired by seed) from stratified (by sample and class) cross-validation (20 repeats x 5-fold) with empirical and permuted labels with p-value from a Wilcoxon signed-rank test and overall mean AUC shown. (b) Predictive microbial logratio network after applying DiCoVarML to the full data. Nodes (v = 8) represent selected genera with sizes indicating the sum of absolute coefficients (β_i) associated with each node. Each directed edge (e = 14) indicates a ratio (outgoing part over receiving part) with thickness proportional to the absolute coefficient β_i (larger absolute coefficient = thicker edges). Blue/Red node colors indicate increased abundance (relative to network) is associated with nonNEC/NEC samples. (c) Distribution of logistic regression scores ($\text{Score} = \Sigma_{i=1}^{e}\beta_i \cdot Logratio_i$) from model fit to full data. (d) Black line represents overall mean score by day. Dashed line indicates the non-NEC (lower) vs. NEC (above) decision threshold. Red points indicate misclassification (nonNEC vs. NEC). (e) Logistic regression scores of NEC positive samples only stratified by survival, with p-value from Mann-Whitney U test.

inops, *Solobacterium moorei* and *Eubacterium eligens* to be the top three species associated with control samples (Figure 2.6D) where increased abundances of these species within the microbial network are associated with lower ensemble model scores. Finally, we tested if there was an association between the ensemble scores and cancer stage via AJCC ($n = 258$) or TNM ($n = 168$) staging. Indeed, using the stage-agnostic ensemble scores, we found strong associations between the ensemble score and stage for both AJCC ($p < 3.696 \cdot 10^{-5}$; Figure 2.6E) and TNM ($p < 0.043$; Figure 2.6F) labeled samples. Combined, our results indicate higher ensemble model scores are associated with advanced cancer stages, highlighting potential clinical utility of DiCoVarML.

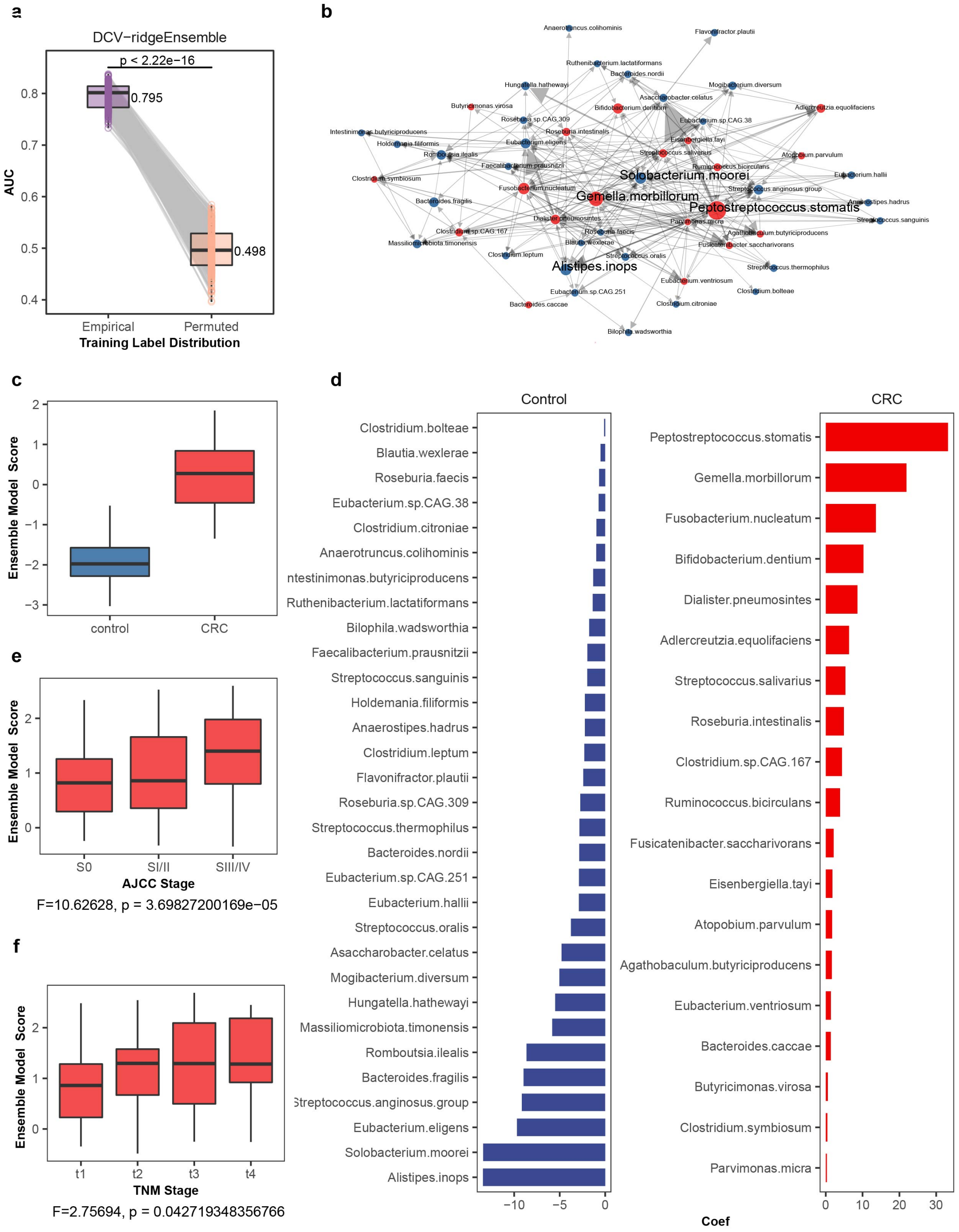

Figure 3.6: Caption on next page

Figure 3.6: Meta-analysis classifying CRC vs. control from gut microbiome composition using targeted (T=50) DiCoVarML with ensemble model using samples (n=1,305) from 11 publicly-available cohorts. (a) AUC (paired by seed) with stratified (by dataset) cross-validation (252 splits x 2-fold) with empirical and permuted labels with p-value from a Wilcoxon signed-rank test and overall mean AUC shown. (b) Predictive microbial log ratio network after applying DiCoVarML to the full data. Nodes (v = 50) represent selected genera with sizes indicating the sum of absolute coefficients (β_i) associated with each node. Each directed edge (e = 259) indicates a ratio (outgoing part over receiving part) with thickness proportional to the absolute coefficient β_i (larger absolute coefficient = thicker edges). Blue/Red node colors indicate increased abundance (relative to network) is associated with nonCRC/CRC samples. (c) Distribution of ensemble model scores from model fit to full data. (d) Species level contributions to ensemble model scores. Coefficient size (x-axis) indicates overall contribution to score where absence of nonCRC (blue bars) species are associated with CRC and the absence of CRC (red bars) genera are associated with nonCRC samples. (e) Boxplot of ensemble model scores stratified by AJCC stage (n = 258). (f) Boxplot of the ensemble model scores stratified by TNM stage (n = 168). F-statistics and p-values from Score $\sim$ Stage linear model.

3.10 Discussion

In this paper we introduced and benchmarked the DiCoVarML framework for robust feature selection and classification of compositional datasets using log-ratio-transformed metagenomic features. Through our simulated example data scenarios and real-world data case studies, we have demonstrated the appropriateness and utility of this framework for supervised classification, as well as its particular relevance for metagenomic data. Importantly, our framework flexibly supports both targeted and untargeted feature selection in a multi-level manner to identify selected parts (e.g., taxa, metabolites) and a subset of logratios between those parts. All of the code used to generate our results are included as part of the DiCoVarML R package developed to implement this framework.

3.11 Methods

3.11.1 DiCoVarML framework

We here describe the details of the various components of the DiCoVarML framework.

3.11.1.1 Data processing: sparse features and treatment of zeroes

The DiCoVarML framework takes as input taxonomic count tables from 16S or WGS sequencing, with or without zeroes. For taxonomic tables, reads should be assigned a taxonomy using a suitable database (e.g. SILVA, Greengenes, RDP, NCBI) and aggregated to a taxonomic level of interest (e.g. Genus, Family). DiCoVarML can handle any relative data where logratio based predictors are of interest. In building models in terms of logratios, one must necessarily select a strategy for handling zeroes and sparsely-counted taxonomic parts.

Sparse features are removed with a default 10% threshold, retaining taxa (parts) if present ($\text{counts} \geq 1$) in at least 10% of samples. After sparse features have been removed, any samples with zero total reads on the remaining parts are also removed. While flexible, by default the DiCoVarML framework handles zeroes via the multiplicative replacement strategy described in Ref. [Martín-Fernández et al., 2015] as implemented in Ref. [Hinton and Mucha, 2021] for metagenomic data. Importantly, because this strategy reinterprets a zero to mean that taxa are present but below a detection limit (uniform across parts), the logratios between non-zero parts remain preserved. Moreover, in this strategy the logratio between two zero-replaced parts becomes zero and thus contributes a zero value to a regression formula. Notably, while we find this strategy to be sufficient for robust predictions in our study here, zeroes can instead be imputed using others techniques such as those found in the zCompositions R package if desired.

In the application of our framework here, we aim to most conservatively minimize the possibility of any information leaking from test/validation sets leaking into discovery/train sets. To this end, we perform the removal of sparse parts/samples and the multiplicative zero replacement on both the training and discovery sets (described below).

3.11.1.2 Outer-loop cross-validation: classification performance

Because feature selection is an important step in the DiCoVarML framework, our cross-validation schema must account for this to minimize over-fitting. DiCoVarML uses nested feature selection to minimize the chance of reporting artificially high AUC estimates that might be obtained when feature selection is performed on the full dataset before cross-validation. Specifically, in frameworks where feature selection is not nested, inflation of AUC or other classification performance metrics can be directly attributable to information leakage from the test set during model training. The nested schema used in the DiCoVarML framework directly prevents such information leakage.

We estimate classification performance through r repeats of k-fold cross validation, stratified by group labels. Importantly, selection of the model type and number of parts within the DiCoVarML framework is nested and treated as a hyper-parameter that is tuned, with part and ratio feature selection performed as part of the training. Each fold is left out once for testing with the remaining folds used for discovery in the inner loop (described below). This process creates k unique discovery-test partitions for each repeat. For concreteness here, we denote the ith testing fold of the jth repeat as $\mathbf{\Upsilon}_{ij}$ and the union of the remaining folds in the jth repeat as the discovery set $\mathbf{\Psi}_{ij} = \bigcup_{l \neq i} \mathbf{\Upsilon}_{lj}$.

Each discovery set is further partitioned in the inner loop (as described below) to perform model and part selection. After fitting this model and selected parts to $\boldsymbol{\Psi}_{ij}$ (and re-selecting a subset of logratios from the available parts as part of the fitting procedure), we then assess performance on the corresponding test set, $\boldsymbol{\Upsilon}_{ij}$. The overall classification performance estimate is obtained by averaging over all ij test fold indices. We focus here on AUC as the primary metric of interest, but other measures could also be used (e.g., AUPRC, accuracy, sensitivity, etc.).

We note that the computational costs involved in the DiCoVarML framework are highly dependent on the number of parts analyzed; reasonable computational time and memory requirements can be expected for < 500 parts after preprocessing.

3.11.1.3 Inner-loop cross-validation: model and part selection

Model selection, including selection of either a targeted or untargeted number of parts selection, occurs within an inner cross-validation loop. In this inner loop, we further partition the (outer) discovery set $\boldsymbol{\Psi}_{ij}$ into 2 folds, $\boldsymbol{\Psi}^1_{ij}$ and $\boldsymbol{\Psi}^2_{ij}$, to be used as training-validation pairs. While this process can be repeated, the default DiCoVarML setting is to use a single repeat in this step to reduce computational time. Training on 1 of these 2 folds at a time, we identify a subset of T parts (e.g., taxa, metabolites, etc.) which then restrict the set of available logratio features for model building. In our *targeted* feature selection mode, the parts subset size T is user defined. In contrast, T is selected automatically in the *untargeted* mode.

3.11.1.4 Targeted multi-level feature selection

The targeted multi-level feature selection method takes as input raw count data from a training set $\boldsymbol{\Psi}^f_{ij}$, discovery set $\boldsymbol{\Psi}_{ij}$, or the full data for final model development (after out-of-sample performance has been estimated from the outer-loop cross-validation). Letting p be the number of parts after preprocessing, either $m = \binom{p}{2} = p(p-1)/2$ pairwise logratios are computed for PLR signatures or the $m = p - 1$ additive logratios are computed for ALR signatures forming the base log-ratio matrix $\mathbf{R}$. For ALR/Hybrid signatures the reference part is selected such that given this reference, the Procrustes correlation between the ALR and the full PLR geometry is maximized [Fawcett, 2006]. Part-level feature selection starts by computing the differential compositonal variation (DCV) scores [Hinton and Mucha, 2021] for each logratio in $\mathbf{R}$. The DCV scores are given by $DCV(\mathbf{R}) = \check{V}$. Specifically, DCV scores are computed using the dcvScores function from the selEnergyPermR R package [Hinton and Mucha, 2021]. We then construct a weighted DCV network defined as $G = (V, E, \check{V})$ where V is the set of p part vertices, E is the set of m edges or pairwise logratios between parts, and $\check{V} \in \mathbb{R}^{m \times 1}$ are the weights given by logratio DCV scores. Let the (unweighted) adjacency matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$ and weight matrix $\mathbf{W} \in \mathbb{R}^{p \times p}$ be derived from G. We compute the vertex strength for the uth vertex (representing the uth part) as $s_u = \sum_{v=1}^{p} a_{uv} w_{uv}$. We select the top t_i parts from these strengths and compute all $\binom{t_i}{2}$ pairwise logratios between them for PLR/Hybrid signatures or all $t_i - 1$ additive logratios (with a Procrustes correlation maximizing reference) for ALR signatures to form the targeted logratio matrix $\mathbf{L}$. Log-ratio-level feature selection is then carried out by first computing the DCV scores for each logratio, retaining only those with positive DCV scores. With this, the subset of logratios is selected by sequential addition from the list of logratios sorted by DCV score (high DCV scores first) until the subset includes at least t_i parts.

3.11.1.5 Model and target part selection

Using inner-loop cross-validation the best model and number of target parts can be selected by averaging the AUC estimates across the {train,validate} sets. From these results we select the model (m_{max}) and number of target parts (t_{max}) with the highest average AUC to to estimate classification performance (across discovery sets $\mathbf{\Psi}_{ij}$) or build a final model (using full data).

3.11.1.6 Machine learning models

While the choices of which machine learning model paradigms to use within DiCoVarML is flexible and can be determined by the user, we have here focused on the ridge regression and ensemble model modeling paradigms as the defaults to be utilized in DiCoVarML. Ridge regularized logistic regression in DiCOVarML uses the glmnet R package with alpha = 0 and type set to either binomial for binary classification or multinomial for multi-classification. Cross-validated AUC values are computed and stored. For final model development, regression scores and coefficients can be directly interpreted from the model using either the predict() or coef() functions from the R STATS package.

For the average probability ensemble model that we use by default in DiCoVarML, we apply the following machine learning algorithms: random forest (ranger R package), support vector machine with radial basis function (kernlab R package), regularized regression (glmnet R package), random forest with extra trees (ranger R package), and partial least squares discriminate analysis (pls R package). Each model is fit (to the appropriate fold) and the predicted class probability $\bar{\phi}_i$ for the ith sample (e.g., in application to the corresponding validation or test fold) is computed as the average probability across all models used. Cross-validated AUC values are computed and stored. For all machine learning modeling the Caret R package is used. All models are trained with default parameters except random forest models are trained with 500 trees and mtry tuning grid using either 1 or the square-root of the number of logratio features.

3.11.1.7 Final model scores and logratio coefficients

In DiCoVarML, final model scores are computed after cross-validation performance has been estimated. For ridge-regression models, the final model score for each sample is obtained after fitting a ridge regularized logistic model to all data (after processing to remove sparse features/samples and zeros) using the cv.glmnet() from the glmnet R package with alpha = 0 and default settings otherwise. For ridge regression models the model score after fitting becomes:

$$y_i = \beta_0 + \sum_{j=1}^{|\mathbf{P}^{\text{final}}|} \beta_j p_{ij}$$

The fitted β coefficients ($j \geq 1$) correspond to logratio features selected for the final logratio signature.

Ensemble model scores for binary classification are obtained after fitting the ensemble model (see above) to all data (after processing). While direct interpretations of ensemble models are difficult, in DiCoVarML we transform each average prediction probability $\bar{\phi}_i$ to an unconstrained ensemble model score $\varepsilon_i = \log(\bar{\phi}_i/(1-\bar{\phi}_i))$, imputing non-zero values in the numerators or denominators of this transformation by the same procedure described previously for zeros in the data. We then fit a secondary ridge penalized linear regression model (adjusting for class/group) to understand how each logratio in the data contributes to ε, using the glmnet R package with default λ grid. The resulting model coefficients may be used to interpret how each logratio contributes to the final ensemble model score. Specifically, we solve the following penalized linear model problem with adjusting for group/class (κ):

$$\min_{\beta_0,\beta_1,\beta} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} \left(\varepsilon_i - \left[\beta_0 + \beta_1\kappa + \beta^T p_i\right]\right)^2 + \lambda \frac{||\beta||_2^2}{2}$$

From this, for ensemble models, β_i (for $i \geq 2$) corresponds to the coefficients for each logratio in $\mathbf{P}^{\text{final}}$ and can be used interpret how each logratio contributes to the final ensemble model score.

3.11.2 Synthetic compositional data generation

In this section we describe in detail how we generate synthetic data and decision boundaries for the examples in Figure 2 and the synthetic data distribution benchmarks in Figure 3.3. Notably, decision boundaries can be described as hypersurfaces that separate a space into different classes/groups. We generate binary classification problems in our synthetic examples and benchmarks. We also describe the machine learning analysis used to assess classification performance.

3.11.2.1 Isoproportional decision boundary

Given a three part composition (p, q, r), an isoproportional line representing a fixed ratio between 2 parts is a straight line on the 2-dimensional simplex. That is, the isoproportional line between 2 parts (p, r) is defined by the relationship $p/r = C$ for constant C (independent of q). We simulate two-class compositional data from a simple additive logistic normal distribution [Aitchison, 1982] with $D = 3$ parts (p, q, r) as follows. Working in the $d = D - 1 = 2$ dimensional space of logratios $(\log \frac{p}{r}, \log \frac{q}{r})$, we define the mean vectors $(\mathbf{m}_1, \mathbf{m}_2) \in \mathbb{R}^d$ and covariance structures $(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) \in \mathbb{R}^{d \times d}$ for each class. The simulated training data in the top row of Figure 2 was generated from multivariate Gaussian distributions with means $\mathbf{m}_1 = (-0.5, 2.0)$ and $\mathbf{m}_2 = (0.5, 2.0)$ and covariances $\mathrm{diag}(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) = 0.2$ and off-diagonal covariance elements drawn uniformly at random on $[0.0, 0.1]$, then ensuring the covariance matrices are semi-positive definite by application of the nearPD() function from the R Matrix package. We note in particular that these simulated points correspond to a ground-truth isoproportional boundary $p/r = 1$ (i.e., $\log \frac{p}{r} = 0$). Finally, we apply the one-to-one additive logistic transformation to map each simulated logratio point $\boldsymbol{r} \in \mathbb{R}^d$ back to $\boldsymbol{x} \in \mathbb{R}^D$ in the space of compositions by $x_i = \exp(r_i)/S$ for $i = 1, \ldots, d$ and $x_D = 1/S$, with normalization constant $S = \exp(r_1) + \cdots + \exp(r_d) + 1$.

The test data in the top row of Figure 2 was generated by the same process as the training data but with means $\mathbf{v}_1 = (-0.5, -2.0)$ and $\mathbf{v}_2 = (0.5, -2.0)$. Importantly, the positions of the test means correspond to the same isoproportional boundary as the training data but centered in a different region of the simplex due to different-scaled contributions from the q component, representing, e.g., possible out-of-study differences across populations, methodologies, or noise levels.

3.11.2.2 Compositional line decision boundary

Using the so-called "Aitchison's Geometry," compositional boundaries on the simplex can be defined using geometric power transformations and perturbation operations [Pawlowsky-Glahn et al., 2007]. Given a compositional vector $\mathbf{z} = [z_1, \ldots, z_D]$ on a $(D-1)$-dimensional simplex (that is, constrained to $\sum_i z_i = 1$), the power operation is defined by $\alpha \odot \mathbf{z} = [z_1^\alpha, \ldots, z_D^\alpha]/\sum_i z_i^\alpha$. Similarly, the perturbation operation between two compositional vectors $\mathbf{y}$ and $\mathbf{z}$ is defined by $\mathbf{y} \oplus \mathbf{z} = [y_1 z_1, \ldots, y_D z_D]/\sum_i y_i z_i$. Armed with these algebraic operations, we can define a compositional-line decision boundary $\mathbf{b}(\alpha) = \mathbf{y} \oplus (\alpha \odot \mathbf{z})$ by specifying $\mathbf{y}$ and $\mathbf{z}$, with $\alpha \in [-\infty, \infty]$. The boundary in the bottom row of Figure 2 was defined by setting $\mathbf{y} = (1/3, 1/3, 1/3)$ and $\mathbf{z} = (0.6, 0.3, 0.1)$.

For interpretation, we emphasize that a decision boundary constructed in this way is equivalent to the zero level set of particular linear combinations of logratios, noting that different linear combinations of logratios may be algebraically equivalent to one another. Indeed, such linear combinations of logratios are then equivalent to a linear combination of log part values which are automatically constrained to be independent of a constant multiplicative factor applied uniformly to all values. In particular, in terms of the composition (p, q, r), the decision boundary in the bottom row of Figure 2 using the $\mathbf{y}$ and $\mathbf{z}$ listed above is equivalent to the zero level set of $f(p, q, r) = \frac{1}{\log 2}\log(\frac{p}{q}) - \frac{1}{\log 3}\log(\frac{q}{r}) \doteq 1.44 \log p - 2.35 \log q + 0.91 \log r$. That is, $f > 0$ is on one side of the boundary with $f < 0$ is on the other side. We again note there are other linear combinations of logratios giving the same zero level set boundary. We call attention to the fact that the resulting coefficients on the logged parts (i.e., on $\log p$, $\log q$, $\log r$) sum to zero by construction, since any linear combination of logratios is independent of an overall multiplicative scale applied uniformly to each component part. We also note that the decision boundary in the bottom row of Figure 2 has zero offset because of the symmetry in the selected $\mathbf{y}$ used here, but non-zero offsets would result from other $\mathbf{y}$ selections. Finally, we note that the zero level set of a linear combination of logratios (with possible non-zero offset coefficient) is left unchanged by a scalar multiple applied across all coefficients; that is, in a sense aligned with logistic regression, the $f = 0$ boundary of equal probability would be left unchanged but the certainty one might assign to a class label prediction is affected by how quickly f changes away from the zero level set.

To generate simulated data (perfectly) separated by the compositional line boundary $\mathbf{b}(\alpha)$, we generate $n = 100$ random α values uniformly on $[-10, 10]$ and evaluate $\mathbf{b}(\alpha)$ for each point. We then randomly shift each of these points to the same side of the boundary by sampling the shift size Δ uniformly on $[0.1, 0.2]$ and moving the point to $\mathbf{x} = \mathbf{b}(\alpha) \oplus (\Delta \odot \boldsymbol{\rho})$, where the constant shift direction here is set to $\boldsymbol{\rho} = (1, 1, 100)/102$. We similarly generate points on the other side of the boundary by the same procedure, starting from $n = 100$ new α values and a new Δ for each point, but with the shift given by $\mathbf{x} = \mathbf{b}(\alpha) \oplus ([-\Delta] \odot \boldsymbol{\rho})$. Finally, we partition this data into train and test sets according to the value of the first component ($x_1 = p$ in (p, q, r)) of each data point, with $x_1 > 1/3$ points assigned to the train set and those with $x_1 \leq 1/3$ assigned to the test set.

3.11.2.3 Machine learning analysis

For the isoportional decision boundary simulation 100 samples were generated for both classes of data for both training and testing sets. Using the R caret and Random Forest packages, random forest models were fit to the training data with mtry = 2 with ntrees = 500. Training set AUC was estimated using 10 repeats of 10-fold cross-validation. Final models for all data transformations were fit to the full training dataset and used for prediction on the testing dataset. Classification performance was estimated using the AUC metric from the R pROC package.

For the compositional line decision boundary, 200 samples for each class were simulated for both training and testing sets. Using the training set, feature discovery and model fitting was carried out with the glmnet R package with alpha = 1 (LASSO penalty). Only the selected features from these procedures were included in the final training and test data. Notably, to simulate true feature selection, such as in the case of biomarker discovery, for relative abundance normalized and CLR-transformed data, the normalization/transformation operation was applied to the selected features. For PLR-transformed data, only the selected ratios were computed on the training and test data. A final logistic regression model was fit to the training data and tested on the test data for all data transformations. Classification performance was assessed using the AUC performance metric.

3.11.3 Synthetic 16S and WGS data simulation

To simulate data with real 16S and WGS characteristics, we used the selEnergyPerm R package. Specifically, we used the simFromExpData.largeMeanShft() function from the selEnergyPermR package to simulate count data with differences in the mean vectors of additive logratio transformed data. To do this, we used the process described in Ref. [Hinton and Mucha, 2021] where a zero inflated negative binomial model (ZINB) is fit to healthy WGS sequenced samples from the Observational Study of Blood Glucose Levels and Gut Microbiota in Healthy Individual trial [Zeevi et al., 2015] using the curatedMetagenomicData [Pasolli et al., 2017] and 16S sequenced fecal samples from the TwinsUK population [Goodrich et al., 2014] using the zinbwave R package. We note parametric simulations and goodness of fit analysis in Ref. [Calgaro et al., 2020] demonstrate ZINB models are reasonable for modeling metagenomic count data. Once the ZINB models were fit we synthesized binary class count data for $n = 200$ samples (100 in each class), with true differences between the classes. In addition to the raw count matrix, the primary experimental parameters of the simFromExpData.largeMeanShft() function are the featureShiftPercent parameter (effect size) which controls the magnitude of difference between mean vectors of each class and the perFixedFeatures parameter (Signal Density) which controls the number of parts that will have simulated between class mean differences where higher values are associated with fewer parts being shifted (sparse signal density). From a machine learning classification perspective, these parameters directly control how difficult or easy the learning task will be where smaller effect sizes among fewer parts are more difficult to learn vs. larger effect sizes among more parts which are easier to detect. Here we set featureShiftPercent = $\{18\%, 24\%, 30\%\}$ and perFixedFeatures = $\{50\%, 70\%, 90\%, 95\%, 98\%\}$.

3.11.4 Publicly-available datasets used for benchmarking

We utilized 16S and WGS data from several publicly-available data sources as benchmark data, extracting processed OTU/taxa tables with matched clinical data. The following data sources were used:

3.11.4.1 curatedMetagenomicData

Metagenomic datasets were extracted from version 3.0.0 of the curatedMetagenomicData [Pasolli et al., 2017] R package. The curatedMetagenomicData R package provides a common interface to access standardized metagenomic data (e.g. relative abundance, pathway abundance) for new analyses. Using the curatedMetagenomicData() function in this package, the following studies were extracted by study ID: ZhuF_2020 [Zhu et al., 2020], RubelMA_2020 [Rubel et al., 2020], WirbelJ_2018 [Wirbel et al., 2019], and QinN_2014 [Quinn et al., 2019]. The SummarizedExperiment R package rowData() and assay() functions were used to extract the taxa information and count tables respectively. Sample clinical data were extracted directly from the curatedMetagenomicData object using the SummarizedExperiment colData() function. Each study was checked for repeated measures on subjects to ensure appropriate cross-validation schemes were applied. Only samples with at least 10^6 total reads were included in our analysis. Additionally, taxa were aggregated to the species taxonomic level.

3.11.4.2 Qitta

Metagenomic data for the gut microbial metabolism and nonalcoholic fatty liver disease (NFALD) study [Sharpton et al., 2019] were downloaded from Qitta [Gonzalez et al., 2018]. Qitta is an open source multi-omics platform that provides database resources for storing and processing publicly-available omics studies. Using this platform, the sample information and two related BIOM files were downloaded directly from the web interface using Qitta Study ID 11635. The 16S BIOM files were loaded in R using the biomformat R package. For this study, the total samples ($n = 290$) were filtered to only include samples with a NFALD status collected with at least 1,000 total reads, yielding 185 samples (NFALD+Cirrhosis = 44, nonNFALD+Cirrhosis=141). Taxa were aggregated to the genus taxonomic level.

3.11.4.3 MicrobiomeHD

16S data from the clostridium difficile infection (CDI) microbiome study [Schubert et al., 2014] were downloaded from the microbiomeHD database [Duvallet et al., 2017], a standardized database of uniformly processed publicly-available case-control studies. Specifically, the

cdi_schubert_results.tar.gz file was downloaded from the microbiomeHD web interface. Within this directory, we extracted the OTU data from *cdi_schubert.otu_table.100.denovo.rdp_assigned* and sample clinical data from *cdi_schubert.metadata.txt*. Taxa data were aggregated up to the genus level.

3.11.4.4 Selbal

The 16S Crohn's and HIV benchmark datassets originally included in the presentation of the selbal method [Rivera-Pinto et al., 2018] were extracted from the cloned selbal R package available on gitHub [Rivera-Pinto et al., 2018].

3.11.5 Publicly-available case study datasets

3.11.5.1 Necrotizing enterocolitis (NEC)

Fecal metagenomic data were extracted from microbiomeDB version 23 [Oliveira et al., 2018], a web based discovery tool and database that uniformly processes and stores 16S and WGS datasets along with related clinical metadata, using the NICU NEC (WGS) study ID. The longitudinal NICU NEC study [Olm et al., 2019] contains 1,100 microbiome-profiled fecal samples from 150 pre-term infants collected during the first months of life where some infants go on to develop NEC. For our case study, we downloaded processed WGS taxon abundance (*NICUNEC.WGS.taxon_abundance.tsv*) and sample detail (*NICUNEC.WGS.sample_details.tsv*) tables. The NEC positive samples were filtered to include only microbiomes related to an onset of greater than 7 days (Days of period NEC diagnosed > 7) and samples after the first day (Age at sample collection days > 0) of life (since some samples developed NEC within the first 7 days of life). The filtered NEC positive samples were then merged with the NEC negative samples. The taxa tables were aggregated at the genus level. Further, because these data are longitudinal and involve repeated measures, cross-validation folds were evenly stratified by infant and NEC.

3.11.5.2 Colorectal cancer (CRC) meta-analysis

Fecal metagenomic datasets for our 11-cohort meta-analysis were processed and collected from the curatedMetagemic R package (also used for some of the benchmark datasets, as described above). In particular, we aggregated sample clinical data and WGS fecal microbiome data ($n = 1,395$) from the following study IDs: FengQ2015 ($n = 107$) [Feng et al., 2015], GuptaA2019 ($n = 60$) [Gupta et al., 2019], HanniganGD2017 ($n = 55$) [Hannigan et al., 2018], WirbelJ2018 ($n = 125$) [Wirbel et al., 2019], ThomasAM2018b ($n = 60$) [Thomas et al., 2019], ThomasAM2018c ($n = 80$) [Thomas et al., 2019], VogtmannE2016 ($n = 104$) [Vogtmann et al., 2016], ThomasAM2018a ($n = 53$) [Thomas et al., 2019], YachidaS2019 ($n = 509$) [Yachida et al., 2019], YuJ2015 ($n = 128$) [Yu et al., 2017a], ZellerG2014 ($n = 114$) [Zeller et al., 2014]. Only samples with at least 10^6 total reads were retained ($n = 1,305$). To most conservatively estimate out-of-sample generalizability, all (outer loop) cross-validation partitions in this meta-analysis were strictly stratified by dataset only. Because only 9 samples from the HanniganGD2017 were retained, we always included this dataset in the partition with 6 data sets (the other partition containing 5 data sets). Our analysis then considered all $\binom{11-1}{5} = 252$ such partitions into 5/6 datasets, using each split once for training and testing.

3.11.6 Use of existing methods in benchmarks

3.11.6.1 Selbal

To estimate the classification performance of selbal we used the selbal R pacakge available on github [Rivera-Pinto et al., 2018]. Because the primary selbal function in the selbal package performs cross validation internally, the following process was used to comparably nest the selbal feature selection method and test on common folds. Using the count data from the ijth discovery set, we identify taxa to include in the balance using the selbal.aux() function with the zero.rep = “one”. If the number of taxa returned after this step is ≤ 1 then all taxa are included. After taxa have been identified we extract the balance values for both the discovery and testing folds using the bal.value() function. A final logistic regression model is fit to the balances with corresponding class labels for the discovery set and used to make predictions on the test. Performance metrics are computed as described above.

3.11.6.2 CLR-LASSO

To implement the clr-lasso method we use standard functions in a manner consistent with [Susin et al., 2020]. To comparably apply the clr-lasso method in a nested way to common folds for benchmarking, we first compute the CLR transformation on the discovery/test sets after adding a pseudo count of 1. Feature scaling is first estimated from the discovery set and then applied to scale features in the discovery and test sets. The glmnet R package is then used to fit a LASSO model (alpha = 1) to the discovery data for feature discovery. If the number of features selected after this step is ≤ 1 (indicates LASSO failed to converge) then all features are used. After subsetting the discovery and test sets for the ijth parititon to include only the features selected (nested feature discovery), the CLR transformation is recomputed. Finally, after scaling features as described above, a LASSO penalized logistic regression model is fit to the discovery data (after feature selection) and used to make predictions (i.e. using the predict() function with ”lambda.min”) on the test data (after feature selection). Overall performance metrics are computed as described above.

3.11.6.3 Coda-LASSO

To implement the coda-lasso method for feature selection we cloned functions from the referenced github repository in [Susin et al., 2020] and carried out the analysis similar to the github documentation. Specifically, to comparably estimate classification performance in a nested manner on common folds, the following process was used. We first added a 1 pseudo count to the discovery and testing data sets (default for coda-lasso). Here, to minimize test data leaking into the discovery set, the z-transform (coda-lasso specified) was estimated from the discovery data only and applied to both the discovery and test data for the ijth partition. To identify a suitable lambda value for input into the coda_logistic_lasso() function, we used the glmnet R package to estimate a grid of lambda values and then selected the lambda value that both included > 0 features and maximized the “proportion of explained deviance” estimated from the coda_logistic_lasso() function. Once lambda was estimated, the coda_logistic_lasso() was used to select a subset of features for further model development. After subsetting the discovery and test sets for the ijth partition to include only the features discovered in the previous step, a pseudo count of 1 was added and the z-transformation was again performed as described above. If the number of features selected after this step was ≤ 1 (indicates LASSO failed to converge) then all features were used. We again optimize lambda and construct a final LASSO regularized logistic regression model using the coda_logistic_lasso(). The returned “beta” coefficients from the coda_logistic_lasso() function were used to compute the final scores for both the discovery and test data. From this, a logistic regression model was fit to the discovery data (after feature selection) and used to make predictions on the test data. Overall performance metrics are computed as described above.

3.12 Data availability

All data used to benchmark the DiCoVarML framework is publicly available and can be accessed as described in the Methods section.

3.13 Code availability

All functions required to implement the DiCoVarML framework have been made publicly available in the DiCoVarML R package available on gitHub at https://github.com/andrew84830813/DiCoVarML.git. All code used here to benchmark the DiCoVarML framework, conduct case studies, and produce the figures is accessible on github at

https://github.com/andrew84830813/DiCoVarML.ProjectRepo.git.

CHAPTER 4: SELENERGYPERM NASAL MICROBIOME ANALYSIS[3]

4.14 Introduction

Approximately 7 million adults and more than 3.5 million youth are current electronic cigarette (e-cigarette) users [Wang et al., 2017, Leventhal et al., 2019, Wang et al., 2020]. E-cigarettes heat and aerosolize e-liquids containing nicotine and flavorings dissolved in humectants propylene glycol and glycerin. E-cigarette use has been steadily increasing over the past decade, especially among teenagers and young adults, reversing the previous decline in youth tobacco use [Wang et al., 2017, Cullen et al., 2019]. Public health crises, such as the outbreak of e-cigarette and vaping-associated lung injury in 2019-2020 and the ongoing SARS-CoV-2 global pandemic, highlight the importance of research examining the effects of e-cigarettes on respiratory immune function [Kiernan et al., 2021, McAlinden et al., 2020].

There is emerging evidence that e-cigarettes disrupt respiratory innate immunity. Previous work has demonstrated the potential for e-cigarette toxicity and impairment of respiratory immune defense using in vitro and in vivo models as well as in samples from human subjects [Martin et al., 2016, Reidel et al., 2018, Clapp et al., 2017, Madison et al., 2019, Ghosh et al., 2019]. For example, e-cigarette users have altered markers of innate immune responses in induced sputum and bronchoalveolar lavage fluid in comparison with smokers and nonsmokers [Reidel et al., 2018, Ghosh et al., 2019], and chronic e-cigarette exposure in mice can dysregulate endogenous lung lipid homeostasis and innate immunity [Madison et al., 2019, Sussan et al., 2015]. In vitro studies have demonstrated that e-liquids, e-cigarette aerosols, and their components can impair the function of ciliated airway cells and respiratory immune cells [Clapp et al., 2017, Gerloff et al., 2017, Muthumalage et al., 2017, Behar et al., 2018, Clapp et al., 2019, Hickman et al., 2019]. Furthermore, e-cigarette exposure has been shown to enhance bacterial virulence and adhesion to airway cells [Hwang et al., 2016, Miyashita et al., 2018], suggesting that e-cigarette exposure may impact the

[3] This chapter was adapted from our manuscript(preprint) available on Research Square. For context, the full details from the manuscript/preprint are included here: follows: Elise Hickman*, Andrew Hinton*, Bryan Zorn et al. E-Cigarette Use, Cigarette Use, and Sex Modify the Nasal Microbiome and Nasal Host-Microbiota Interactions, 03 August 2021, PREPRINT (Version 2) available at Research Square [https://doi.org/10.21203/rs.3.rs-725763/v2]. *indicates that these authors contributed equally to the work; My contributions to this work include the multivariate statistical and machine learning methods and analysis as described in the Methods and Results of the manuscript.

respiratory microbiome. However, the effects of e-cigarette use on the respiratory microbiome in humans have not been evaluated.

The respiratory microbiome includes distinct communities of microbiota along the length of the respiratory tract [Man et al., 2017]. Similar to microbial communities at other body sites, respiratory microbiota interface with the host immune system, and dysbiosis of the respiratory tract microbiome has been associated with diseases, including cystic fibrosis, chronic obstructive pulmonary disease, asthma, and chronic rhinosinusitis, as well as with disease exacerbations and smoking cigarettes [Man et al., 2017, Ubags and Marsland, 2017, Charlson et al., 2010, Ramakrishnan et al., 2016]. Sampling the nasal microbiome is straightforward in contrast to the lower airway microbiome, which is easily contaminated with oral microbiota during specimen collection [Grønseth et al., 2017]. In addition, the nose is an important gatekeeper in the respiratory tract, as potential pathogens must often colonize this region before progressing to the lower respiratory tract [Man et al., 2017]. This role has become even more clear and relevant with the emergence of SARS-CoV-2, with recent studies showing associations between the nasal microbiome and SARS-CoV-2 infection [Di Stadio et al., 2020, Rosas-Salazar et al., 2021]. Of note is that dysbiosis of the nasal microbiome specifically has been associated with smoking cigarettes [Charlson et al., 2010], and gene expression and histopathological changes due to smoking are similar in the nasal and lower airway epithelium [Martin et al., 2016], supporting the use of the nasal microbiome for studying the effects of environmental exposures on the respiratory microbiome.

Mechanistic study of the human microbiota is an important focus when studying the human microbiome, where identifying microbes associated with disease is paramount [Gilbert et al., 2018]. To uncover complex interactions in microbiome association studies changes to classical statistical methods are required [Li, 2019]. In addition, computational methods that robustly integrate disparate data types with 16S microbiome data for association testing have been limited [Jiang et al., 2019]. In particular, microbiome datasets have interspecies interactions, small sample sizes, high dimensionality (where the number of features greatly exceed the number of samples), are sparse (where the data matrix contains many zeroes), and when converted to relative abundance are compositional, meaning the total number of reads is not informative [Gloor et al., 2017]. Combined, these challenges significantly confound the multivariate integrative analysis required to improve our understanding of host-microbiome interactions. Thus, novel analytical tools are necessary to uncover true signals hidden within small sample size microbiome data.

In this study, we sampled the nasal microbiomes of smokers, nonsmokers, and e-cigarette users using a non-invasive absorptive strip to collect nasal epithelial lining fluid. We then used high-throughput sequencing of the bacterial 16S rRNA gene from the strips to identify bacteria present and analyze the bacterial composition of the nasal microbiome in our subjects. Because these microbial communities are composed of highly interdependent taxa that have complex interaction patterns, multivariate data analysis is critical to extract biologically relevant information.

Here, we leverage Selection Energy Permutation [Hinton and Mucha, 2021], a novel multivariate association test that simultaneously tests associations while identifying robust subsets of pairwise logratios in the setting of high-dimensional, low sample size data. These reduced subsets are then used to integratively analyze nasal microbiome and matched cell-free nasal lavage fluid mediator data to determine: 1) whether there were significant compositional differences in the nasal microbiomes of E-cigarette users, smokers, and nonsmokers, 2) whether levels of nasal lavage fluid (NLF) mediators are significantly different in e-cigarette users and smokers in comparison with nonsmokers, and 3) whether changes in levels of these mediators correlate with nasal microbiome dysbiosis. Our data demonstrate nasal microbiome dysbiosis and unique networks of host-microbiota mediators in e-cigarette users and smokers in comparison with nonsmokers. This is indicative of disrupted respiratory mucosal immune responses in these groups and potentially increased susceptibility to infection by specific bacterial taxa. We also observed significant sex differences in the nasal microbiome, highlighting the importance of including sex as a biological variable in nasal microbiome studies.

4.15 Methods

4.15.1 Subject recruitment

Nasal epithelial lining fluid (NELF) strips, nasal lavage fluid (NLF), and venous blood were obtained from healthy adult human e-cigarette users, smokers, and nonsmokers as described previously (Table 1) [Rebuli et al., 2017], forming our exposure groups. Inclusion criteria were healthy adults age 18-50 years who are either nonsmokers not routinely exposed to environmental tobacco smoke, active regular cigarette smokers, or active e-cigarette users. Active cigarette smoking and e-cigarette use were determined as described previously [Martin et al., 2016]. Exclusion criteria were current symptoms of allergic rhinitis (deferred until symptoms resolve), asthma, FEV1 less than 75% predicted at screen, bleeding disorders, recent nasal surgery, immunodeficiency, current pregnancy, chronic obstructive pulmonary disease, cardiac disease, or any chronic cardiorespiratory condition. After the consent process was completed, a medical history and substance use questionnaire was obtained, and subjects were issued a diary to document smoking/vaping for up to 4 weeks, after which they returned for sample collection. E-cigarette users averaged less than 1.5 cigarettes/day in their smoking/vaping diaries, while cigarette users ranged from 4.93-20 cigarettes per day in their diaries. To compare demographic characteristics between subjects in the different exposure groups, age, BMI, and serum cotinine levels were tested for normality using the Shapiro-Wilk test, and groups were compared using the Kruskal-Wallis test followed by the Steel-Dwass method for non-parametric multiple comparisons (analogous to a one-way ANOVA with Tukey's HSD for parametric data).

4.15.2 Serum Cotinine Measurement

Venous blood was collected in BD Vacutainer serum-separating tubes (Fisher Scientific, Waltham, MA) and allowed to clot for a minimum of 15 minutes at room temperature. The blood was then centrifuged at 1200 x g for 10 minutes, and the serum layer was transferred to a fresh tube and stored at $-80°$C until samples were collected from all subjects. Serum was assayed for cotinine, a metabolite of nicotine that can be measured as a biomarker of nicotine consumption, using a commercially available ELISA kit (Calbiotech, Mannheim, Germany) per manufacturer's instructions. Absorbance was read on a CLARIOstar plate reader (BMG Labtech, Ortenberg, Germany). The limit of quantification for serum cotinine was 5 ng/mL. For samples below the limit of detection, a value of zero was assigned. Serum was not available for one subject in the cohort.

4.15.3 NELF Strip 16S Sequencing

DNA was extracted from whole NELF strips using Powersoil DNA Isolation Kit (MoBio Laboratories). Sequencing libraries were prepared as previously described [Muhlebach et al., 2018]. Samples were sequenced on an Illumina MiSeq kit version V3 2x300 paired end over the V3-V4 bacterial 16s gene. Raw sequencing data were demultiplexed and processed to generate a table of operational taxonomic units (OTUs). Specific primer schema, qPCR data, and the OTU table (having at least 10 sequences per OTU across all samples) are provided in the supplement. Raw sequence data have been uploaded under the BioProject accession number PRJNA746950 within the Sequence Read Archive.

4.15.4 NLF Processing and Soluble Mediator Measurement

Cell-free nasal lavage fluid was obtained via processing of raw nasal lavage fluid as described previously [Horvath et al., 2011]. Briefly, raw nasal lavage fluid from each nostril was pooled and centrifuged at 500x g through a 40 mum strainer for 10 minutes. Supernatant (cell-free NLF) was collected and stored at $-80°$C until samples were collected from all subjects. Cell-free NLF was assayed for mediators of host-microbiota interaction (neutrophil elastase, immunoglobulin A (IgA), lactoferrin, lysozyme, interleukin 8 (IL-8), alpha-defensin 1, beta-defensin 1, beta-defensin 2, cathelicidin (LL-37)) using commercially available ELISA kits per manufacturer's instructions as described in (Appendix B: Table A.1). Absorbance was read on a CLARIOstar plate reader. For samples below the limit of detection, a value of $^1\!/_2$ the lowest standard was assigned. Cell-free nasal lavage fluid was not available for one subject in the cohort (Appendix B: Figure A.5).

4.15.5 Sequencing Data Processing and Filtering

Five samples were removed from the dataset due to a low number of reads (Appendix B: Figure A.5). A spiked pseudomonas positive control was identified correctly as pseudomonas. To control for potential contamination on the NELF strips, the decontam R package was used to remove contaminants [Davis et al., 2018]. This package uses an algorithm that takes into account the relative abundance of OTUs in samples and controls to remove the most likely contaminants and has been shown useful for respiratory samples [Drengenes et al., 2019]. This reduced the number of OTUs from 5346 to 4677. Alpha diversity measures (Observed, Chao1, ACE, Shannon, Simpson, Fisher) were calculated using the phyloseq R library before trimming OTU counts less than 5 for downstream analysis. This brought the number of OTUs to 3059 for downstream analysis.

4.15.6 Alpha diversity

Shannon and Simpson diversity indices were computed for each sample. Diversity indices were tested for normality using the Shapiro-Wilk test and further statistical tests to compare groups were carried out using the appropriate parametric (two-tailed t-test, ANOVA) or non-parametric (Kruskal-Wallis, Steel Dwass) tests. These analyses were performed using JMP Pro 14 and GraphPad Prism 8.

4.15.7 Nasal Microbiome Compositional Data Analysis

To limit spurious findings and because absolute sequencing counts are uninformative [Gloor et al., 2017, Quinn et al., 2019, Tsilimigras and Fodor, 2016], compositional data analysis (CoDA)[Aitchison, 1982] was carried out on the OTU count table after aggregating OTUs (O = 3059) by family (min. level assigned) and genera (max level assigned) and removing taxa not present in at least 20% of samples. The 20% sparsity threshold was selected to maximize class-specific information (Sex, Exposure group) while ensuring the microbial signatures were robust and contained minimal noise due to excessive sparsity. After aggregating OTUs, we define the taxa count matrix, $\mathbf{X} \in \mathbb{R}^{\mathbf{n}\times\mathbf{p}}$, with n = 62 samples and p = 143 taxa. The closure operator, $C[\cdot]$, was then used to map the count data of each element x_{ij} of $\mathbf{X}$ onto its corresponding coordinate on the unit-sum simplex, defining $\mathbf{X}' = C[\mathbf{X}]$ in terms of matrix elements as

$$x'_{ij} = (C[\mathbf{X}])_{ij} = \frac{x_{ij}}{\sum_{k=1}^{p} x_{ik}}. \tag{4.12}$$

Because the presence of zeros is a major limitation of the logratio transformation essential to CoDA, all zeroes must be robustly imputed to non-zero values. To overcome this we use the ratio-preserving multiplicative replacement strategy which has been shown to have several theoretical advantages over simple additive replacement:[Martın-Fernandez et al., 2003] we set the δ imputed values to a single constant equal to the smallest nonzero value encountered in $\mathbf{X}'$. Here the constant δ gives provides a consistent interpretation of zeroes across samples. From this, we impute zeros and replace $\mathbf{X}'$ with $\mathbf{Z}$ defined in matrix elements as

$$z_{ij} = \begin{cases} \delta\,, & x'_{ij} = 0 \\ \left(1 - \sum_{k|x_{ik}=0} \delta\right) x'_{ij}\,, & x'_{ij} > 0 \end{cases} . \tag{4.13}$$

4.15.8 Partial redundancy analysis to remove variation due to Sex

To remove the significant effect of Sex (which otherwise obscures the exposure group effect) on $\mathbf{Z}$, partial Redundancy Analysis (pRDA)[Legendre and Legendre, 2012] was used. Here we encode the Sex variable into the design matrix S. Additionally, to ensure multiple regression computations used in pRDA are performed on symmetric vectors in real space that preserves the inter-sample Euclidean distances, a center logratio (clr) transformation was applied [Aitchison, 1982] to $\mathbf{Z}$, defining the clr values in $\mathbf{C} \in \mathbb{R}^{n \times p}$ where for the clr of the ith sample is:

$$\mathbf{C} = \mathrm{clr}(\boldsymbol{z}_{i*}) = \boldsymbol{c}_{i*} = \left[\log \frac{z_{i1}}{\left(\prod \boldsymbol{z}_{i*}\right)^{1/p}}, \ldots, \log \frac{z_{ip}}{\left(\prod \boldsymbol{z}_{i*}\right)^{1/p}}\right] \text{ for } i = 1, \ldots, n. \tag{4.14}$$

With $\mathbf{C}$ defined, pRDA was carried out in the vegan R package [Oksanen et al., 2015]. Multivariate linear regression of $\mathbf{C}$ on $\mathbf{S}$ (i.e. computed as a series of multiple linear regression on individual features) was used to produce the fitted values $\hat{\mathbf{C}}$. To remove the Sex effect as in analyzing the residuals after pRDA, the adjusted values of $\mathbf{C}$ were computed by $\mathbf{P} = \mathbf{C} - \hat{\mathbf{C}}$ where $\hat{\mathbf{C}}$ contains the variation attributable to Sex. With the residuals matrix $\mathbf{P}$ defined in clr-coordinates which are not suitable for downstream pairwise logratio transformations, an inverse clr transformation was applied to map the adjusted coordinates back to the unit-sum simplex. The Sex adjusted relative abundance matrix $\mathbf{M}$ by applying the inverse clr transformation as

$$\mathbf{M} = \text{clr}^{-1}(\boldsymbol{p}_{i*}) = C\left[\exp(\boldsymbol{p}_{i*})\right] \text{ for } i = 1, \ldots, n. \tag{4.15}$$

4.15.9 Nasal Microbial Signature Identification using Selection Energy Permutation

To identify microbial logratio signatures in the setting of high-dimensional microbiome data we utilized the recently developed Selection Energy Permutation (SelEnergyPerm) method [Hinton and Mucha, 2021]. The SelEnergyPerm, which is inspired by the direction-projection-permutation (DiProPerm) [Wei et al., 2016] method, detects sparse associations in high-dimensional compositonal microbiome data in conditions (low sample size) where traditional non-parametric association test (e.g. PERMANOVA, ANOSIM) have limited statistical power. More specifically, treating the data as compositional, SelEnergyPerm: (1) selects, from the full frame of PLRs, a set of logratio signature that maximizes the the association between groups and (2) test the statistical significance of the association using distribution free permutation testing. Key to SelEnergyPerm is the statitsic used to measure the overall between group association. In this work, we use SelEnergyPerm with the energy statistic (E-statistic)[Rizzo and Székely, 2016]. To describe the energy statistic, let the logratio signature matrices be defined as $\mathbf{X}$ and $\mathbf{Y}$ where n, m are the number of samples in each group and f is the number of PLR logratios in the selected. From this the E-statistic[Rizzo and Székely, 2016] is defined by

$$\mathcal{E}^{\alpha}_{n,m}(\mathbf{X}, \mathbf{Y}) = 2A - B - C, \tag{4.16}$$

where $\alpha = 2$ and A, B, and C are specified by

$$A = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\|x_i - y_j\|_\alpha, B = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|x_i - x_j\|_\alpha, C = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{m}\|y_i - y_j\|_\alpha.$$

From this, extending the E-statistic to two or more groups the multi-group E-Statistics is defined as

$$S = \sum\left(\frac{n_j + n_k}{2N}\right)\left[\frac{n_j n_k}{n_j + n_k}\mathcal{E}^{\alpha}_{n_j,n_k}(\mathbf{A}_j, \mathbf{A}_k)\right]. \tag{4.17}$$

where $0 < \alpha < 2$ is the α-norm, A is the pooled samples for the jth and kth group, K is the number of groups, and n_j, n_k are the number of samples for the jth and kth groups.

In selEnergyPerm, the multi-group E-statistic is then maximized using greedy forward selection on a subset selected from the set of pairwise logratios. As in the approach in Greenacre et al.,45 selEnergyPerm requires the reduced subset of logratio to be non-redundant. Computationally, it is infeasible in the case of high dimensional metagenomic data to test all possible non-redundant subsets of pairwise logratio which can be expresed as the Cayley number($p^{(}p - 2)$) [Greenacre, 2019]. To overcome this, SelEnergyPerm first ranks the strength of the between group variation for each logratio using the differential compositional variation (DCV) scoring algorithm [Hinton and Mucha, 2021]. Specifically, the DCV score, which is comprised of five diverse statistics (Welch's t-statistic, F-statistic, Brown-Forsythe F-Statistic, Kolmogorov–Smirnov statistic, and Information), measures the variation (location, scale, distributional, etc.) between groups for each logratio. The single DCV score, representing the relative strength of the between group variation for each logratio, is computed as the sum of the z-score scaled (by score) statistics. Network methods are then used to compute a maximum spanning tree (MST), which is intrinsically non-redundant, from a DCV weighted logratio network. Greedy forward selection is then carried out on the MST to identify the final subset of logratios. Specifically, given a logratio signature discovered with true labels, Overall, SelEnergyPerm tests if the observed mutli-sample E-statistic (S^*) is more extreme than random using permutation testing [Ernst, 2004]. That is, multi sample E-statistic are sampled from the permutation distribution of logratio signatures selected under random labels (S_i, indexing different random-label samples). With γ such multi-samaple E-statistics randomly sampled from the permutation distribution the one-sided p-value becomes

$$\hat{p} = \frac{1 + \sum_{i=1}^{\gamma} I(S_i > S^*)}{\gamma + 1}. \tag{4.18}$$

4.15.10 Network Visualization of Microbial signature

As shown in [Greenacre, 2019], logratio can be conveniently visualized using networks. To visualize the microbial logratio signatures, we constructed undirected networks connecting the key taxa (vertices/nodes) by edges representing the formation of a ratio between two taxa with edge weight corresponding to the between-group Kruskal-Wallis H-statistic. While the full logratio structure is directed in distinguishing numerators from denominators, directedness in the visualizations used here does not fundamentally change our interpretation. Networks were visualized using Gephi[Bastian et al., 2009] and R-igraph [Csardi and Nepusz, 2005].

4.15.11 Multivariate statistical test for microbial signals

To specifically test the associations between microbial logratio signatures and Sex/Exposure group identified by SelEnergyPerm, traditional multivariate hypothesis testing was done using permutational multivariate analysis of variation[Anderson, 2017] and implemented using the R vegan package [Oksanen et al., 2015]. Unsupervised lower-dimensional projections of samples and group centroids were done using principal coordinate analysis (PCoA) and were implemented using the R stats package.

4.15.12 Partial Least Squares Discriminate Analysis

Because effect sizes and direction are different to interpret and not well explored in the context of multivariate statistics, we utilized partial least squares discriminate analysis (PLS-DA) [Barker and Rayens, 2003, Brereton and Lloyd, 2014]. PLS-DA is a versatile multivariate statistical regression technique where here we used it to model and understand the relationship between Sex/Exposure group to the SelEnergyPerm selected microbial logratio signatures. We specified a priori the number of PLS-DA components (ncomp) as follows: for the between Sex nasal microbial signature, ncomp = 1; for the between Exposure group nasal microbial signature, ncomp = 2. Model fitting was done using the R caret [Kuhn, 2008] plsda function, with latent space projections and loadings extracted from the final models fit using all samples using R caret [Kuhn, 2008]. PLS-DA biplots were created by scaling and superimposing the loading vectors onto the score coordinates extracted from the final fitted model. PLS-DA biplots were visualized using the R ggplot2 package [Wickham, 2016].

4.15.13 Receiver operating characteristic curve analysis and PLS-DA performance metric

To understand how well the binary PLS-DA models discriminate between Sex using the nasal microbiome signature, we utilized the area under the receiver operating characteristic metric (AUC), as a measure of effect size. We note, in this context (effect size), AUC is not used to estimate out of sample predictive performance. Specifically, AUC represents the probability that a randomly selected instance of class 1 will be ranked higher than a randomly selected instance of class 2 where a value of 0.5 indicates no discrimination between groups and a value of one indication perfect discrimination between groups [Fawcett, 2006]. Additionally, to understand the discriminatory potential of the PLS-DA exposure group models, the multi-class AUC metric was used. Multi-class AUC generalizes binary AUC through pairwise class AUC averaging and has the useful property of being independent of cost and priors as in AUC while having a similar interpretation to misclassification rate [Hand and Till, 2001]. AUC metrics were estimated using repeated k-fold cross-validation [Stone, 1974]. The R pROC package[Robin et al., 2011] was used to compute all AUC metrics. ROC curves, which graph the false positive and true positive rate of a classifier over a range of thresholds, were computed using the R pROC package[Robin et al., 2011] and visualized using the R ggplot2 package [Wickham, 2016].

4.15.14 NLF mediator and microbiome data integration

In this section we detail the exposure group logratio signature identification and data integration for both the NLF mediator and nasal microbiome data. For NLF mediators, we define the NLF data matrix, $\mathbf{L} \in \mathbb{R}^{n \times \tau}$, where n = 66 samples and $\tau = 7$ mediators. Here we treated the NLF mediator data as multivariate relative data such that the sample-wise total concentrations are relative (Appendix B: Figure A.7A). Zeroes in $\mathbf{L}$ were imputed after applying the closure operator as described in our compositional data analysis methods. From this, using a compositional approach, we define all pairwise logratios of $\mathbf{L}$ to be $\mathbf{L}' \in \mathbb{R}^(n \times k)$, with $k = \tau(\tau - 1)/2 = 21$. With a focus on multivariate associations we sought to remove uninformative NLF mediators using feature selection. To do this we computed the differential compositional variation (DCV) score [Hinton and Mucha, 2021] and assigned each NLF mediator logratio a score by averaging the within-fold DCV score using 20 repeats of 10-fold cross-validation. NLF logratios with a DCV score < 0 were considered uninformative and were removed (Appendix B: Figure A.7B). From this $\mathbf{L}'$ was reduced to $\hat{\mathbf{L}} \in \mathbb{R}^{n \times k}$ where k=4 (Appendix B: Figure A.7C) logratios. With key logratio features selected univariate associations testing between NLF mediator logratios in $\hat{\mathbf{L}}$ and Exposure group the Kruskal-Wallis test was applied followed by pairwise Wilcoxon rank-sum testing if $\alpha < 0.05$.

For the nasal microbiome signal, we obtained the exposure group logratio signature by applying the SelEnergyPerm method to $\mathbf{M}$ to get $\hat{M} \in \mathbb{R}^{n \times r}$ where $n = 62$ samples and $r = 9$ SelEnergyPerm selected logratios. Concatenating these data, we define the integrated NLF mediator and nasal microbiome matrix as $\mathbf{D} \in \mathbb{R}^{q \times f}$ where $q = 61$ (6 samples were removed due to either missing nasal microbiome or NLF data) and $f = 13$ (4-nasal lavage plus 9 microbiome log-ratio features). Exposure group discrimination was estimated separately for $\hat{\mathbf{L}}$, $\hat{\mathbf{M}}$, and $\mathbf{D}$ using multi-class AUC from 50 repeats of 10-fold cross-validation using 2-component PLS-DA models. Multi-class AUC estimates using $\hat{\mathbf{L}}$, $\hat{\mathbf{M}}$, and $\mathbf{D}$ were compared between groups using the non-parametric Wilcoxon rank-sum test.

4.15.15 Nasal NLF mediator and microbiome association analysis

A final 2-component PLS-DA model to discriminate between exposure groups was fit to $\hat{\mathbf{M}}$. Using dimensionality reduction inherent to PLS-DA, the first PLS-DA component (explaining the most variation) was extracted as a latent variable for further analysis. Pearson's correlation coefficients (PCC) and subsequent p-values were computed between the first PLS-DA component and $\mathbf{L}'$ defined in "NLF mediator and microbiome data integration" method section above. The p-values obtained for PCC were adjusted for multiple comparisons (q-value) using the Benjamini-Hochberg (BH) correction[Benjamini and Hochberg, 1995] and were considered significant if $q \leq 0.10$. These analyses were carried out using the R stats and caret packages.

4.15.16 Between Exposure group Correlation analysis

Partitioning the samples of $\mathbf{D}$ into 3 matrices based on exposure group (nonsmokers, e-cig users, or smokers), we calculated all pairwise PCC and p-values between features for each group. We corrected p-values for multiple testing using the BH procedure and report q-values. Correlations were considered significant if $q \leq 0.10$. Significant PCC within each subject were then aggregated across all exposure groups and visualized as a graph using the R igraph package [Csardi and Nepusz, 2005].

4.15.17 Confidence Intervals and univariate statistical test for logratios

Logratio 95% confidence intervals were calculated for each logratio i as

$$\mathrm{CI}_i = \bar{x_i} \pm 1.96\frac{s_i}{\sqrt{n}}. \tag{4.19}$$

where for the ith logratio, x_i =sample mean, s_i =sample standard deviation and n=number samples. logratios with confidence intervals bounds that do not include 0 are interpreted as enriched on average for the numerator if $x \geq 0$ or denominator if $x < 0$. The Kruskal-Wallis and Wilcoxon rank-sum test were used for univariate comparisons of logratios between Sex or Exposure groups. Moreover, p-values were adjusted for multiple comparisons using the BH correction using the R stats library and are reported as q-values. For more conservative control of false discovery, false coverage rate adjusted confidence intervals can be considered.

4.16 Results

4.16.1 Subject Demographics

Demographic, questionnaire, and smoking/vaping diary data are summarized in Table 1. The study cohort was comprised of 30% nonsmokers (n = 20), 42% e-cigarette users (n = 28), and 28% smokers (n = 19) with at least n = 8 per sex within each exposure group. E-cigarette users were significantly younger (26.39 ± 1.44) than nonsmokers (30.75 ± 1.32) and smokers (31.89 ± 1.91) ($p < 0.05$). BMI did not differ significantly between the exposure groups. Questionnaires and smoking/vaping diaries were completed for 95% (19/20) of nonsmokers and 100% of e-cigarette users and smokers. However, there was variability in the completeness of diaries filled out by e-cigarette users, particularly for the e-cigarette use parameters (mL/day, puffs/day, nicotine concentration, flavor, device). Cigarette users smoked an average of 12.68 ± 0.96 cigarettes per day, whereas 25% (7/28) of e-cigarette users smoked a cigarette during the diary period with an average of 0.14 ± 0.07 cigarettes per day, while 13 e-cigarette users reported puffs per day and 16 reported mL e-liquid/day and e-liquid nicotine concentration in mg/mL. These e-cigarette users averaged 53.90 ± 16.54 puffs/day, 3.60 ± 0.70 mL of e-liquid, and 19.43 ± 4.92 mg/mL nicotine in e-liquids. One smoker reported vaping on one day of the diary, which is the reason for the non-zero values for e-cigarette use parameters in the smoker category. Nonsmokers did not report previous cigarette smoking or marijuana use, whereas 79% (22/28) of e-cigarette users were former cigarette smokers, while 14% (4/28) of e-cigarette users and 21% (4/19) of smokers reported marijuana use in their diaries. Cotinine, a metabolite of nicotine, was not detectable in the serum of nonsmokers and was significantly elevated in the serum of e-cigarette users (127.99 ± 15.42) and smokers (170.16 ± 21.41) in comparison with nonsmokers ($p < 0.0001$), as expected.

	Nonsmokers	**E-Cigarette Users**	**Smokers**
n	20	28	19
Sex (Male/Female)	8/12	19/9	10/9
Race (White/AA/Asian/Other)	16/1/2/1	18/4/5/1	10/8/0/1
Age	30.75±1.32	26.39±1.44#	31.89±1.91
BMI	27.11±1.31	30.01±1.51	27.65±1.43
Cigarettes/Day	0±0	0.14±0.07	12.68±0.96
mL E-Liquid/Day	0±0	3.60±0.70	0.015±0.015
E-Cigarette Puffs/Day	0±0	53.90±16.54	0.466±0.414
E-Liquid Nicotine (mg/mL)	0±0	19.43±4.92	0.158±0.158
Former Cigarette Smoker (Yes/No)	0/20	22/6	19/0
Marijuana Use (Yes/No)	0/20	4/24	4/15
Serum Cotinine (ng/mL)	0±0	127.99± 15.42****	170.16±21.41****

Table 4.1: Subject demographics. Reported values are mean ± standard error. Groups were compared using the Steel Dwass method for non-parametric multiple comparisons. AA = African American. # p<0.05 in comparison with nonsmokers and smokers. **** p<0.0001 in comparison with nonsmokers.

4.16.2 Nasal Microbiome Characteristics

The 4677 OTUs included in the dataset represented OTUs from 19 unique phyla and 225 unique genera. The top four most abundant phyla by average relative abundance across all samples were *Actinobacteria*

(50.2%), *Firmicutes* (36%), *Proteobacteria* (12.0%), and *Bacteroidetes* (1.6%). The top six most abundant genera by average relative abundance across all samples were *Corynebacterium* (40.7%), *Staphylococcus* (19.9%), *Propionibacterium*(11.8%), *Alliococcus* (8.5%), *Moraxella* (5.3%), and *Streptococcus* (4.2%). This microbial composition is similar to previously reported studies of the nasal microbiome [Kumpitsch et al., 2019, De Boeck et al., 2017]. These data are summarized in Figure 4.1, where relative abundances by exposure group and sex are plotted for the most highly abundant phyla and genera.

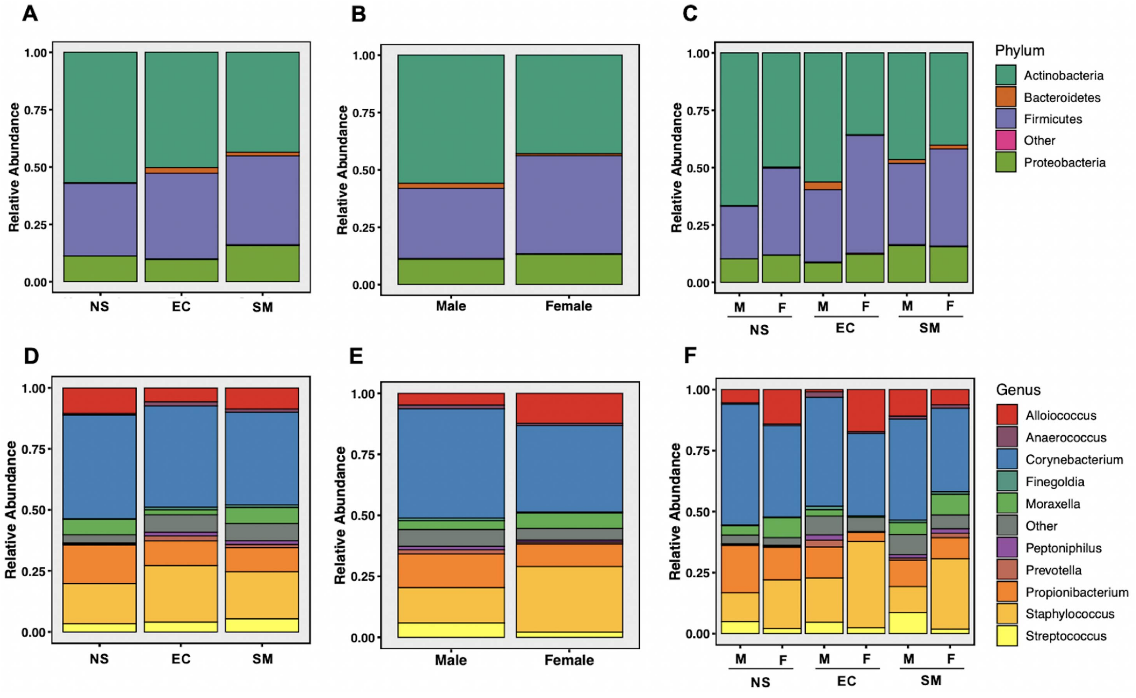

Figure 4.1: Average relative abundances of the top 4 phyla (A-C) and top 10 genera (D-E) plotted by exposure group (A, D), sex (B, E), and sex within exposure groups (C, F). NS = nonsmoker, EC = e-cigarette user, SM = smoker, M = male, F = female.

4.16.3 Alpha Diversity

To determine whether there are differences in alpha diversity between the nasal microbiomes of smokers, nonsmokers, and e-cigarette users, we calculated alpha diversity indices (Observed, Chao1, ACE, Shannon, Simpson, Fisher) using phyloseq [McMurdie and Holmes, 2013]. We did not find any statistically significant differences between the exposure groups for any measure of alpha diversity; however, we did observe a non-significant trend of increased alpha diversity in smokers (Figures 4.2A and 4.2B). Because our group and others have previously observed sex differences in respiratory mucosal immune responses[Rebuli et al., 2018, Cho et al., 2019] we also tested whether alpha diversity was significantly different between male and female subjects. We found that both the Shannon and Simpson indices were significantly higher in males than females (p = 0.021 and p = 0.0078, respectively) (Figures 4.2C and 4.2D). We then tested for the interaction between sex and exposure group and found that sex was a significant source of observed variation (p = 0.0286 for Shannon; p = 0.0102 for Simpson), while exposure group was not. When the data were stratified by exposure group, the only male-female comparison that remained significant was in the e-cigarette user group (p = 0.0361 for Shannon; p = 0.0124 for Simpson) (Figures 4.2E and 4.2F). These results suggest that sex is an important biological variable to consider in studies of the nasal microbiome.

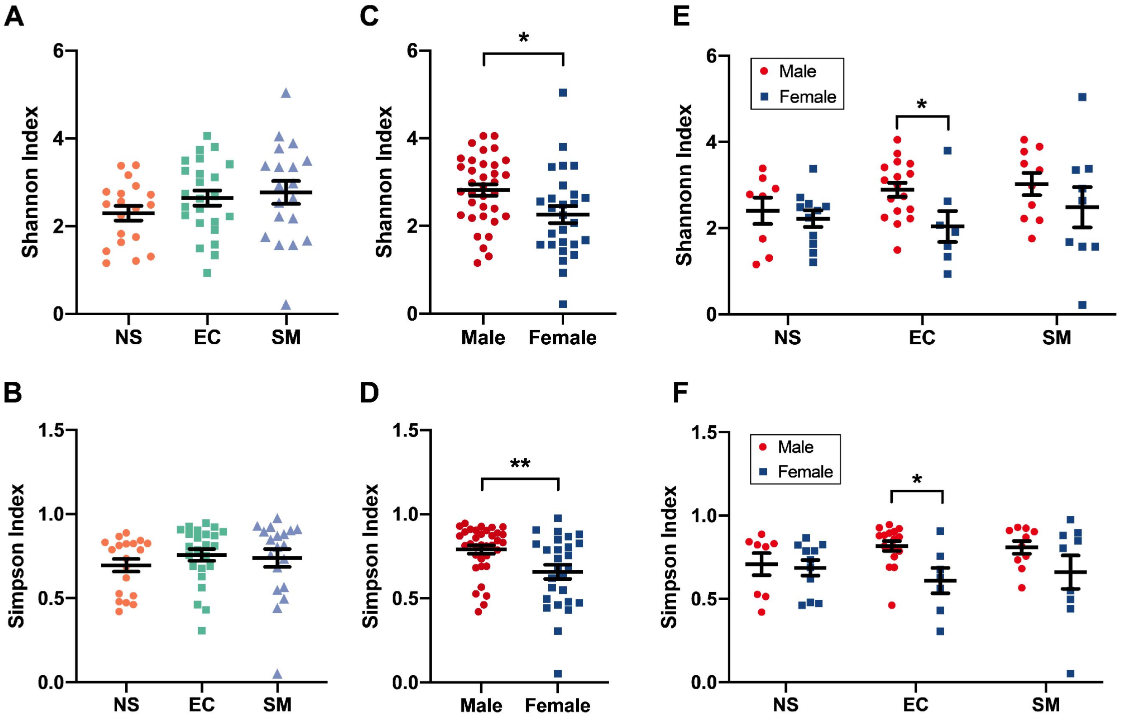

Figure 4.2: Shannon and Simpson indices of alpha diversity are significantly different between sexes, and this difference is most pronounced in e-cigarette users. The Shannon and Simpson indices for alpha diversity were calculated and plotted by exposure group (A, B), sex (C, D), and sex within exposure groups (E, F). NS = nonsmoker, EC = e-cigarette user, SM = smoker. Data are presented as mean ± standard error. * $p < 0.05$, ** $p < 0.01$ by t-test (C), Kruskal-Wallis test (D), or two-way ANOVA with Fisher's LSD (E, F).

4.16.4 Compositional Difference of the Nasal Microbiome by Sex

Because we observed distinctions in alpha diversity between sexes, we next tested whether there were significant compositional differences between the sexes and to identify specific genera capable of explaining these dissimilarities. Given challenges presented by sparse, compositional 16S rRNA sequencing data combined with high-dimensionality (genera = 255) and small sample size (n=62), we leveraged the

SelEnergyPerm [Hinton and Mucha, 2021] method to identify a robust signature of nasal microbiome taxa (among sparse noisy data) capable of explaining compositional differences between sexes.

By applying this method, we discovered (beyond random noise) a subset of genera (g = 6) capable of maximizing the energy distance between male and female samples (p = 0.0123, Appendix B: Figure A.6A). This microbial signature was comprised of four logratios between *Rhodococcus*, *Finegoldia*, *Sneathia*, *Abiotrophia*, *Tannerella*, and *Yaniella* genera (Figure 4.3A). Using the identified logratio signature, PERMANOVA analysis (pseudo-F = 16.586, p =0.0002, Figure 4.3B) also confirmed the existence of differences in the nasal microbiome composition between sex. Analysis of individual taxa logratios between sexes demonstrated important nasal microbiome compositional differences (Figure 4.3C). In female samples, *Yaniella* was more abundant on average than *Rhodoccous* and *Tannerella*, while the reverse was true for males. In male samples, *Abiotrophia* was more abundant on average than *Sneathia*, while the opposite was true for females. Finally, in both males and females, *Finegoldia* was observed to be more abundant than *Yaniella*, however, *Finegoldia* was significantly more enriched relative to *Yaniella* in males compared to females.

Next, we analyzed the microbial signature as a whole using Partial Least Squares Discriminate Analysis (PLS-DA) with a single component to predict sex. Using 20 repeats of 10-fold cross-validation, the average area under the receiver operating characteristic curve (AUC) for predicting sex given the reduced microbial signature was 0.862 (95% CI 0.842 – 0.883, Figure 4.3D). With strong cross-validated predictive performance, a final PLS-DA model was trained on all samples (n=62). Scores from the single PLS-DA component indicated strong separation between sexes (Figure 4.3E). The PLS-DA loading plot (Figure 4.3F), which shows how each logratio contributes to the final score, demonstrates key relationships between taxa logratios. Increased abundance of *Abiotrophia* and *Finegoldia* (in logratios where they appear) were characteristic of males, and increased abundance of *Yaniella* was associated with females. Overall, these findings indicate there exists a compositionally distinct taxa subset that differs strongly in the nasal microbiomes of males and females. Therefore, controlling for sex differences present in the nasal microbiome is important in further analysis.

4.16.5 Compositional Difference of the Nasal Microbiome by Exposure group

We next examined whether there were distinct nasal microbiome compositions between exposure groups (e-cigarette users: n = 24; smokers: n=19; nonsmokers: n=19; See Methods and Table 1). Taking into account nasal microbiome sex differences and applying SelEnergyPerm, we identified a subset of

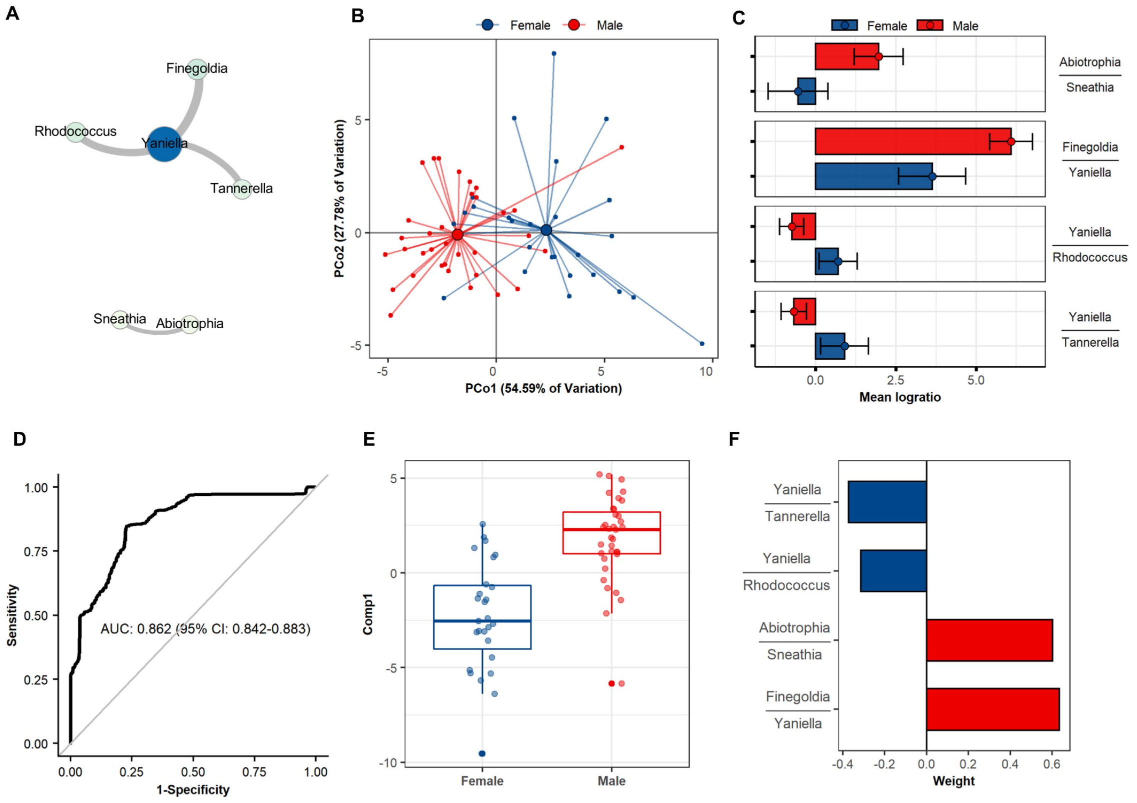

Figure 4.3: Nasal microbiome differences between sexes (Males: n=35; Females: n=27). (A) Network representation of SelEnergyPerm (p=0.0123) derived genus aggregated taxa logratio signature of nasal microbiome differences between sexes (Node = genera; edge = logratio between taxa, Edge-weight = Kruskal Wallis H-statistic between sexes, Size/Color = node strength). (B) Principal coordinate analysis plot of nasal microbiome logratio signature between sex explaining 82.37% of the total variation. (C) Univariate analysis of logratio signature showing average depletion or enrichment of specific taxa logratios between sexes. Error bars reflect 95% confidence intervals of the mean log-ratio value for males and females. (D) Receiver operating characteristics (ROC) curve displaying the area under the curve (AUC) predictive performance (20x10-fold cross-validation) of 1-component partial least squares discriminant analysis (PLS-DA) models trained on nasal microbiome signature between sexes. (E) PLS-DA scores plot of single discriminating component between sexes. Final PLS-DA model fit using all samples (n=62). (F) PLS-DA loadings plot showing contributions of each logratio to final scores.

genera (g = 12) important for explaining key nasal microbiome alterations between exposure groups (p = 0.032, Appendix B: Figure A.6B). This microbial signature comprised nine logratios (edges) between 12 key genera (nodes) (Figure 4.4A). PERMANOVA analysis (pseudo-F = 8.4889, p =0.0002, Figure 4.4B) confirmed differences in nasal microbiome composition between exposure groups given the microbial signature of 9 logratios.

Individual analyses of logratios elucidated specific compositional differences between exposure groups (Figure 4.4C). In e-cigarette users, *Lactobacillus* taxa were significantly more abundant relative to *Bacillus* taxa, while in smokers and nonsmokers, these taxa presented in similar proportions, suggesting an enrichment of *Lactobacillus* among e-cigarette users. E-cigarette users' nasal microbiomes also contained significantly more *Staphylococcus* relative to *Bacillus* than what was observed in nasal microbiomes of both smokers ($q = 0.0097$) and nonsmokers ($q = 0.0031$). In smokers, *Maccrocus* genera were significantly more abundant on average relative to *Hymenobacter*, *Mycobacterium*, *Varibaculum*, and *Rhodococcus*, suggesting that smoking may enrich *Macrococcus* taxa populations in the nasal passage. Additionally, smoker nasal microbiomes contained more *Hymenobacter* relative to *Moryella*, whereas the opposite was true for nonsmokers, both in contrast to e-cigarette users, which maintained on average equal amounts of both genera. In nonsmokers, *Lautropia* taxa were significantly more abundant relative to *Bulleidia*, but this was not observed in smokers and e-cigarette users.

To understand how taxa logratios work together to discriminate between exposure groups, PLS-DA was used with 20 repeats of 10-fold cross-validation (Figure 4D). The estimated multi-classification AUC was 0.851 (95% CI 0.835 – 0.866) suggesting excellent exposure group discrimination. Pairwise examination of exposure group classifications shows strong differences between the nasal microbiomes of nonsmokers/e-cigarette users (AUC = 0.895: 95% CI 0.874 – 0.915) and smokers/e-cigarette users (AUC = 0.893: 95% CI 0.873 – 0.913), with weaker yet distinct differences between smokers/nonsmokers (AUC = 0.803: 95% CI 0.773 – 0.833) (Figure 4.4D). The relative importance of taxa logratios for discriminating between exposure groups was computed using a final PLS-DA model fit using all samples (n=62). The logratio between *Macrococcus* relative to *Hymenobacter* was found to be most important for classifying samples as smoker (least important for e-cigarette user classification), and the logratio between Bacillus taxa relative to taxa from the Micrococcaceae family was most important for samples to be classified as e-cigarette users (least important to be classified as smokers). (Figure 4.4E). Interestingly, inspection of relative logratio importance data failed to uncover logratios disproportionately important for nonsmokers. This observation suggests smoking and e-cigarette use recognizably alter the nasal microbiome in otherwise healthy adults. Overall, analysis of the taxa logratios signature suggests alterations in *Macrococcus* and *Bacillus* genera are important for distinguishing between these exposure groups.

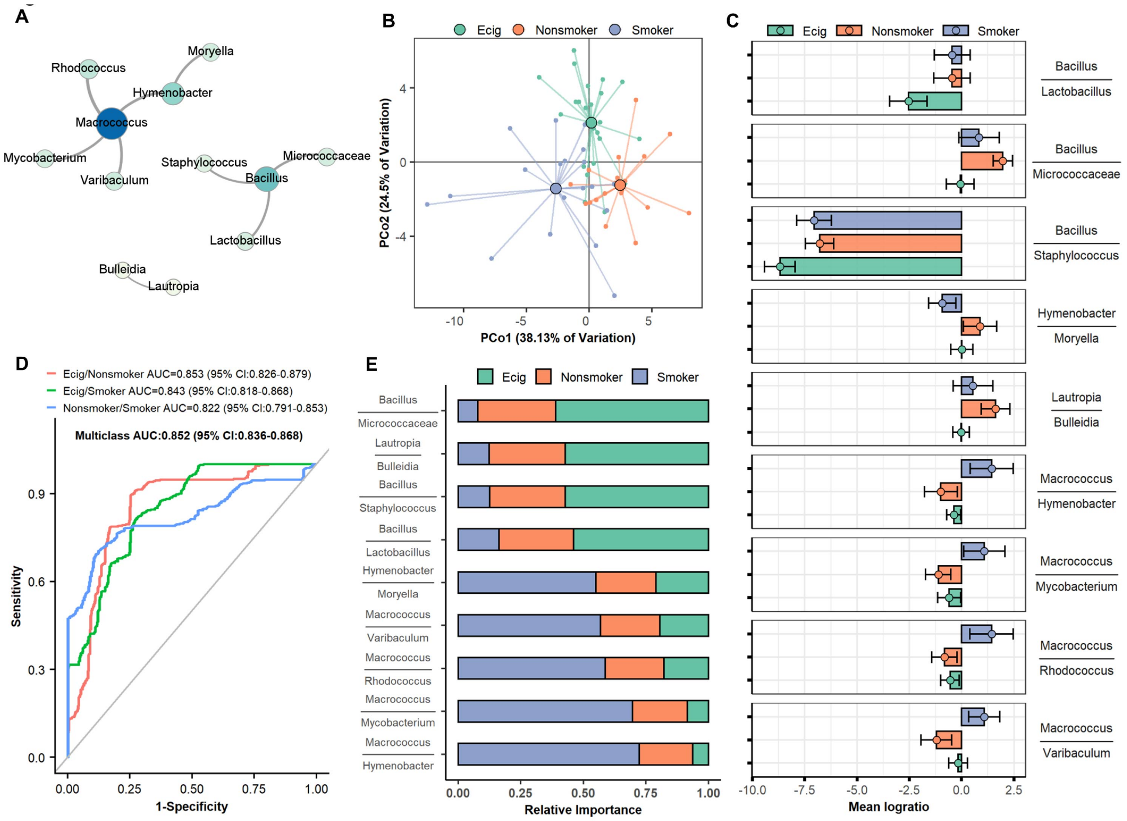

Figure 4.4: Nasal microbiome differences between exposure groups (Ecig-users: n=24; Nonsmokers: n=19; and Smokers: n=19) adjusted for sex. (A) Network representation of SelEnergyPerm (p=0.032) derived genus aggregated taxa logratio signature of nasal microbiome differences between exposure groups (Node = genera; edge = logratio between taxa, Edge-weight = Kruskal Wallis H-statistic between sex, Size/Color = node strength). (B) Principal coordinate analysis plot of nasal microbiome logratio signature between exposure groups explaining 62.63% of the total variation. (C) Univariate analysis of logratio signature showing average depletion or enrichment of specific taxa logratios between exposure groups. Error bars reflect 95% confidence intervals of the mean log-ratio value for each exposure group. (D) ROC curve displaying the multi-classification AUC for predicting exposure group (20x10-fold cross-validation) of 2-component PLS-DA models trained on nasal microbiome signature between exposure groups. (E) Relative importance of logratios for distinguishing between exposure groups in PLS-DA model trained on all samples (n=62).

4.16.6 Differences in NLF mediator Expression Patterns Between Exposure groups

Because smoking and e-cigarette use were associated with distinct changes in the nasal microbiome, we next explored if there was altered expression of innate immune response mediators in the exposure groups. Accounting for differences in absolute concentration (Appendix B: Figure A.7A) and subsequently applying differential compositional variation scoring 32 (See Methods, Appendix B: Figure A.7B), we identified four logratios among NLF mediators that showed strong intergroup variability (Appendix B: Figure A.7C).

These ratios comprised the following NLF mediators: IL-8, DEFB4A-2, neutrophil elastase, IgA, and lactoferrin. Kruskal-Wallis one-way testing (Appendix B: Figure A.7D) of each logratio suggest there exist intergroup differences in NLF mediator expression formed between the concentrations of neutrophil elastase relative to IL-8 (H = 6.4417; p = 0.0399; q = 0.0798) and lactoferrin relative to IL-8 (H = 8.2080; p = 0.0165; FDR = 0.0660). There were no significant differences between exposure groups among logratios formed by IgA relative to IL-8 or DEFB4A-2 relative to neutrophil elastase. However, multivariate analysis with PERMANOVA (pseudo-F = 3.7678, p =0.0030) using the four key logratios confirmed there were differences in NLF mediator expression patterns between exposure groups when considered together. To better understand which groups were different, we applied PLS-DA. Training a PLS-DA model with the NLF mediator expression patterns revealed the strongest between-subject-group discrimination to be among Smokers and Nonsmokers (AUROC = 0.8230, 95%CI 0.7920-0.8530, Appendix B: Figure A.7E). Notably, e-cigarette users' NLF mediators were weakly distinguishable from nonsmokers (AUROC = 0.6720, 95%CI 0.6350-0.7100, Appendix B: Figure A.7E) but more discernible from smokers (AUROC = 0.7480, 95%CI 0.7130-0.7820, Appendix B: Figure A.7E). Together, these results suggest that the expression of NLF mediators in smokers was distinct from that of e-cigarette users and healthy adults.

4.16.7 Integration of NLF mediators and nasal microbiome composition

Finally, we aimed to understand if alterations in NLF mediator expression are associated with nasal microbiome dysbiosis resulting from smoking or e-cigarette use. To this end, we first estimated the discriminatory AUROC of a 2-component PLS-DA model fit on logratios from NLF mediators (Appendix B: Figure A.7C), nasal microbiome (Figures 4.4A), or both nasal microbiome and NLF mediators (Appendix B: Figure A.8A). When compared to individual signatures, improved discriminatory AUROC (Figure 4.5B) was observed when PLS-DA models were fit using the combined nasal microbiome and NLF mediator signatures. Therefore, with established synergy between mediator expression and nasal microbiome composition in discriminating between exposure groups, we next examined if correlations were present between the two.

4.16.8 Association between altered NLF mediator expression and nasal microbiome dysbiosis

Using the first PLS-DA component of the nasal microbiome signature, we found significant correlations with NLF mediator expression, showing an association between the nasal microbiome composition and NLF mediator expression (Figure 4.5C). Examination of the location of samples by exposure group projected along the first PLS-DA component show important projective distinctions between smokers (on average negative projections) and both e-cigarette users and nonsmokers (on average positive projections) (Appendix B: Figure A.8B). Given this, these correlations suggest nasal microbiome dysbiosis caused by cigarette smoke exposure is associated with increased expression of IL-8 relative to neutrophil elastase, Total IgA, and lactoferrin (Figure 4.5C). Moreover, the loadings along the first PLS-DA component (Appendix B: Figure A.8C) show logratios with higher abundance of Maccroccous as being the most important contributor to negative projections. Combined, these data propound an important link between dysbiosis in *Macrococcus* communities within the nasal microbiome and NLF IL-8 expression.

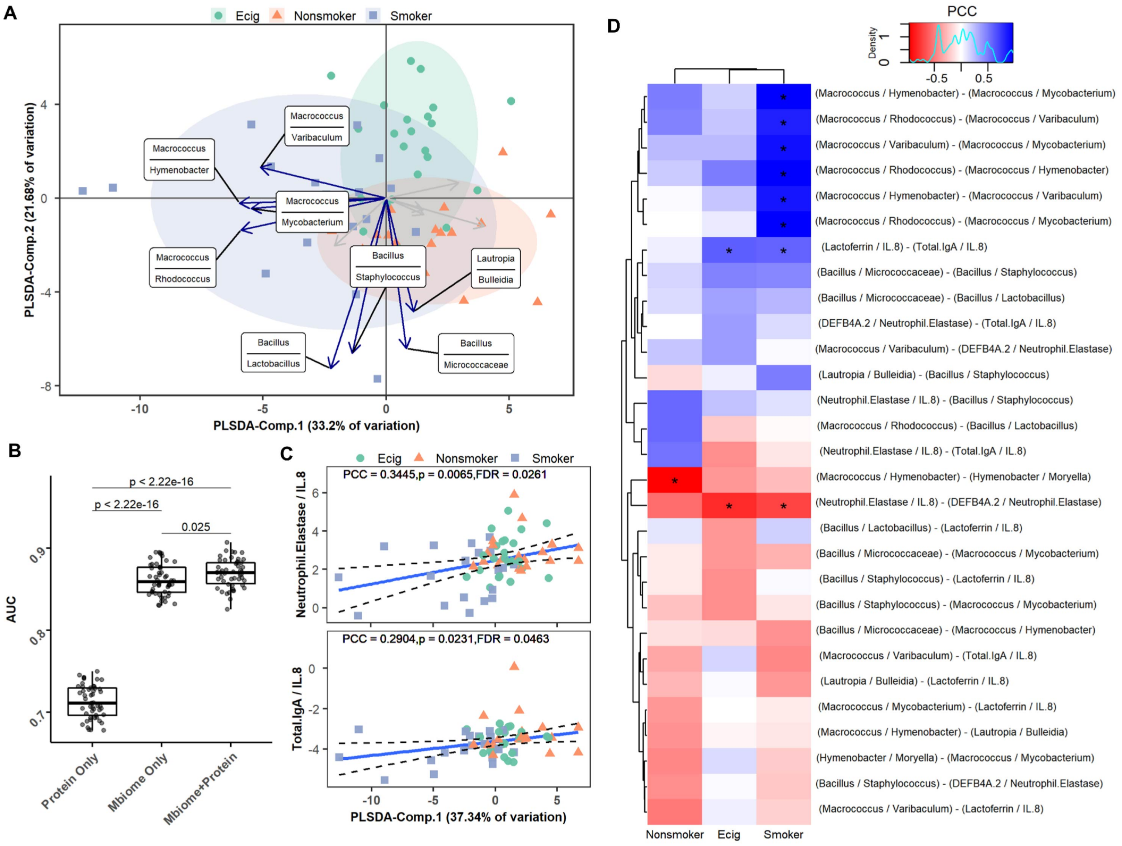

Figure 4.5: Integrating data uncovers association between NLF mediators and nasal microbiome along with identifying distinct correlation patterns between exposure groups (Ecig-users: n=23; Nonsmokers: n=19; and Smokers: n=19). (A) PLS-DA biplot of integrated NLF mediators and nasal microbiome (B) Box and whisker's plot comparing area under the receiver operating characteristic curve performance of 2-component PLS-DA model (50x10-fold cross-validation) using each data type alone or integrated. (C) Scatter plot showing correlations between logratios formed between concentrations(μg/mL) of Lactoferrin, Neutrophil Elastase relative to IL-8 and the first PLS-DA component of the nasal microbiome. (D) Correlation heatmap showing Pearson's correlation coefficients (PCC) between and within the microbiome and protein logratio signatures. (*indicates within group $q \leq 0.10$)

4.16.9 Microbial functional and mediator expression differences between exposure groups

Correlation analysis of the combined NLF mediator expression and nasal microbiome signature reveal distinct correlation patterns within exposure groups suggesting distinct functional differences (Figures 4.5D). Most notably, a significant negative correlation between logratios formed by *Hymenobacter*/*Moryella* and *Macrococcus*/*Hymenobacter* was observed only in the nonsmoker group. This negative correlation highlights a possible role of Hymenobacter, in that it appears to be important for maintaining a healthy balance of *Maccroocous* and *Moryella*. In the e-cigarette and smoking groups, we observed a significant positive correlation between the logratios formed by IgA/IL-8 and Lactoferrin/IL-8. Analysis of this correlation pattern reveals that increased expression of IL-8 in these groups may come at the expense of decreased expression of IgA and lactoferrin or vice versa. We also observed a significant negative correlation between the logratios formed by Neutrophil Elastase/IL-8 and DEFB4A-2/Neutrophil Elastase in the e-cigarette and smoking groups. These strong negative correlations show that increased expression of IL-8 and DEFB4A-2 subsequently results in decreased expression of neutrophil elastase. The final significant correlation pattern observed was in smokers only and consisted of four positively correlated logratios formed by *Macrococcus* relative to *Hymenobacter*, *Mycobacterium*, *Varibaculum*, and *Rhodococcus* (Figure 4.5D). Relatively interpreting these correlations between logratios suggests that as *Macrococcus* becomes more abundant (among these ratios) the abundance of Hymenobacter, Mycobacterium, Varibaculum, and Rhodococcus decreases. This suggests an assocaiotn between cigarette smoke exposure favorable colonization conditions for *Maccroccous* genera which subsequently reduce the abundance of *Hymenobacter*, *Mycobacterium*, *Varibaculum*, and *Rhodococcus*.

From these analyses, our results demonstrate there exists a strong association between altered NLF mediator expression and nasal microbiome dysbiosis. Our findings indicate nasal microbiome dysbiosis from smoking results in the simultaneous increase in IL-8 expression and *Maccroccous* abundance. Additionally, variations in the correlation networks among e-cigarette users and smokers, while similar, were distinct from nonsmokers, suggesting functional differences at the microbial and mediator levels between exposure groups.

4.17 Discussion

Despite the growing body of research showing that e-cigarette use can disrupt the respiratory immune system, no studies to date have assessed the effects of e-cigarettes on the respiratory microbiome and host-microbiota interactions. In this study, after adjusting for sex differences, we found that e-cigarette users, smokers, and nonsmokers have unique nasal microbiomes, with differences driven by the relationships between a subset of key taxa. We also found a subset of immune mediators that had distinct relationships between each other in the different exposure groups. Importantly, we found a link between nasal microbiome dysbiosis and soluble immune mediator networks.

A fundamental feature of our study is that we detected microbial signatures from the nasal microbiome that explained differences between sex and exposure groups using the novel SelEnergyPerm computational method. This method directly accounts for the sparse, high-dimensional and compositional nature of the 16S relative abundance data. Additionally, SelEnergyPerm identifies subsets of robust logratios between taxa, as opposed to analyzing taxa relative abundance alone, yielding higher statistical power in the sparse association setting with low-sample-size compositional data [Hinton and Mucha, 2021]. Most importantly, traditional statistical techniques such as PERMANOVA, ANOSIM, and ANCOM alone were unable to detect these sparse associations within the high-dimensional nasal microbiome feature space. Further, our parsimonious yet statistically significant signatures were then integrated with NLF mediators where we were then able to uncover novel interactions between a taxa subset within the nasal microbiome and the NLF mediators in response to exposure to cigarette or e-cigarette aerosol.

We observed that there were relationships between a subset of taxa that were important in separating the microbial communities of smokers, nonsmokers, and e-cigarette users (Figure 4.4). Only a few studies have previously compared the nasal microbiome of smokers and nonsmokers [Charlson et al., 2010, Yu et al., 2017b]. Charlson et al. found specific bacteria genera that were differentially abundant in smokers and that some genera belonging to the phylum Firmicutes were important in distinguishing smokers from nonsmokers [Charlson et al., 2010]. Other studies did not find any significant differences in diversity measures or relative taxa abundance between smokers and nonsmokers [Charlson et al., 2010]. In our study, which focused on the composition of the nasal microbiome and ratios between taxa rather than relative abundance of individual taxa, we found that alterations in Macrococcus and Bacillus genera are important for distinguishing between exposure groups. Our data also suggest an enrichment of Lactobacillus and Staphylococcus relative to Bacillus in e-cigarette users and enrichment of Macrococcus relative to Hymenobacter, Mycobacterium, Varibaculum, and Rhodococcus in smokers. A shift from Lactobacillus to Bacillus in the lung microbiome has been previously demonstrated in response to influenza A infection and increases in anaerobic bacteria, such as Lactobacillus, have been associated with chronic rhinosinusitis

[Kumpitsch et al., 2019]. Furthermore, Bacillus have been shown to produce antimicrobials against S. aureus [Piewngam et al., 2018], indicating that the patterns we have observed may be directly linked to specific interactions between taxa. An increase in Staphylococcus relative to Bacillus in e-cigarette users is also notable due to the role of species such as Staphylococcus aureus, which is carried normally by about 30% of people and is also considered to be a potential pathogen of the skin and mucosal surfaces [Liu et al., 2015, Sakr et al., 2018]. Our data provide evidence that e-cigarette and smoker nasal microbiomes are distinctly shifted from nonsmokers. Importantly, we also observed that different subsets of taxa were important in separating e-cigarette users and smokers, rather than effects on a continuum from nonsmokers to e-cigarette users to smokers, highlighting the concept that the effects of e-cigarettes are likely unique from those of smokers, even though they are commonly directly compared.

We also measured concentrations of mediators of host-microbiota interactions in nasal lavage fluid to determine whether the changes in the nasal microbiome in different exposure groups are potentially caused by direct effects on the microbiome, mediated by changes in the host immune system, or both. Our data indicate that the expression of immune mediators in nasal lavage fluid samples differed among exposure groups and was driven by shifts in neutrophil elastase and lactoferrin relative to IL-8. Neutrophil elastase and IL-8 are associated with inflammation and neutrophil recruitment, while lactoferrin is an antimicrobial protein primarily produced by epithelial cells and has a wide array of functions, including antioxidant and immune-modulating properties [Actor et al., 2009]. Our results suggest that e-cigarette users and smokers may have altered immune mediator milieu, indicating a shift away from immune homeostasis and towards increased inflammation and neutrophil recruitment. This shift could be partially driving observed differences in the nasal microbiome. Our data indicate that both e-cigarette users and smokers have altered nasal microbial communities and relationships between markers of innate immune response, which could imply that they are at increased susceptibility to respiratory infections and/or that they exist in a state of inflammation and altered immune response. We also uncovered interactions of key immune mediators with the host and microbiota, such as IL-8, neutrophil elastase, and lactoferrin, that are also disrupted by e-cigarette and cigarette use. The microbial shifts we observed in association with e-cigarette and cigarette use could be driven by changes in the microenvironment, such as temperature, pH, free radical formation, and availability of metabolic substrates (e.g. sugars) that could then alter the fitness of different bacteria in the nasal microbial community. The shifts we observed could also be mediated through direct effects on respiratory host defense function, inflammation, and/or specific microbes. Multiple processes are likely at play, but our novel findings on the effects of e-cigarettes on the nasal microbiome add to the growing body of literature demonstrating that e-cigarettes are not without health effects and that they should be more thoroughly investigated for inhalational toxicity.

Because sex differences in the human immune system and its response to respiratory disease and toxicant exposure have been observed previously [Rebuli et al., 2018, Casimir et al., 2013], we also investigated whether there were sex differences in the nasal microbiomes of our subjects. We observed that the relationships between six genera were important in separating the nasal microbiomes of males and females (Figure 4.4A). Increased abundance of Abiotrophia and Finegoldia (in logratios where they appear) were characteristic of males, and increased abundance of Yaniella was associated with Females. Many of these genera have been detected in previous studies of skin, oral, and/or respiratory microbiomes [Charlson et al., 2010, Kumpitsch et al., 2019, Man et al., 2019, Neumann et al., 2020, Hoggard et al., 2017, Chiu et al., 2020, Bacci et al., 2016], but detailed information on the functions of these bacteria as part of the microbial community, as well as their impact on host health, are not available for all taxa. Although some of these genera, such as Abiotrophia and Finegoldia have been associated with disease- and exposure-driven alterations in the respiratory microbiome [Charlson et al., 2010, Kumpitsch et al., 2019, Man et al., 2019, Neumann et al., 2020], we hypothesize that the observed sex difference is neither good nor bad; rather, it is reflective of a different baseline composition in males and females or altered microenvironments in males and females due to differences in toxicant metabolism rates or mechanisms of immune regulation [Zanger and Schwab, 2013, Vemuri et al., 2019]. In other body sites, such as the gut, sex differences have been detected and have been attributed to a variety of factors, including sex hormone levels, pharmaceutical use, and diet [Kim et al., 2020, Shin et al., 2019]. In mice, sex-related differences in gut microbiota were shown to impact pulmonary responses to ozone [Cho et al., 2019]. However, few studies have explored sex differences in the respiratory microbiome [Han et al., 2018]. In the studies that have analyzed data by sex, detection of sex differences is not consistent between studies and is typically not explored in-depth [De Boeck et al., 2017, Liu et al., 2015, De Boeck et al., 2019]. Importantly for the data presented here, compositional differences in the nasal microbiomes of e-cigarette users, smokers, and nonsmokers were not apparent until sex was properly adjusted for, further underscoring the importance of considering sex as a biological variable which significantly modifies exposure effects and can substantially affect data interpretation.

Though our study reveals important community shifts in nasal microbiota and immune mediators associated with e-cigarette and cigarette use as well as with sex, there are limitations to our study. Our novel analysis approach, while properly accounting for the compositional nature of the data, limits us in comparing our work to previous studies, which have been more focused on specific taxa rather than ratios across the microbial community as a whole. As with any study of human subjects, there is also inherent inter-subject variability that can interfere with detection of differences between groups. In our e-cigarette user group, there was considerable variability in factors that could impact the exposure subjects are receiving, including e-liquid flavor, device, nicotine content, and frequency of use. The e-cigarette user group also includes previous smokers and some marijuana use was reported in both smoker and e-cigarette user questionnaires. These factors were included in our analysis and did not show a significant impact on our overall findings due to the nature of the computational models we used. In future studies, larger cohort sizes coupled with more extensive questionnaires could improve the ability to detect which, if any, of these factors may be driving changes in microbiota composition and would also increase power to detect overall changes and shifts in the nasal microbiomes of such subjects given the compositional and sparse nature of 16S sequencing data.

As a whole, our results support and expand on the previously published notion that exposure to inhaled toxicants, including tobacco products, can influence the respiratory microbiome [Charlson et al., 2010, Chen et al., 2020, Mariani et al., 2018]. The novel, robust computational approach in terms of pairwise logratios that we applied allowed us to uncover both exposure- and sex-dependent effects on nasal mucosal host defense responses using straightforward, non-invasive sampling of the upper respiratory tract of human subjects. Importantly, we were able to integrate 16S sequencing data with expression of soluble immune mediators to understand interactions between the nasal microbiome and host milieu by appropriately handling the sparse, compositional data generated by 16S sequencing, accounting for inter-individual variability between subjects' mediator levels, and selecting for features that were most important for separating classes, resulting in interpretable, biologically meaningful results. Conventional analysis pipelines would have limited our ability to integrate these two types of data and detect the exposure and sex-dependent effects we observed, highlighting the importance of applying innovative computational methods to address specific research questions and integrating multiple factors in understanding biological outcomes of exposure and disease.

APPENDIX A: CHAPTER 2 SUPPLEMENTAL DATA

Algorithm 1 Association maximization with greedy forward stepwise selection

```
procedure SELECTIONENERGY($\mathbf{Z}'$, $\mathbf{y}$, $\alpha$, patience, $\varepsilon$)
    if PERMDISP2($\mathbf{Z}'\mathbf{y}$) $< \alpha$ then                ▷ Determine test statistic
        testStatisticFunction = $cF()$
        Metric = ’combinedF’
    else
        testStatisticFunction = $F_{n,\alpha}()$
        Metric = ’discoF’
    end if
    $\mathbf{X} = \mathbf{Z}' \in \mathbb{R}^{n\times 3}$                ▷ Select first 3 columns
    maxF = testStatisticFunction($\mathbf{X}, \mathbf{y}$)
    improvementTime = 0
    for $i \in [4, \ldots, |\mathbf{Z}'|]$ do
        $\mathbf{X}_{new} = \mathbf{X} \cup z'_{*,i}$                ▷ Append $i$th column
        newF = testStatisticFunction($\mathbf{X}_{new}, \mathbf{y}$)
        diff = newF - maxF
        if diff $\geq \varepsilon$ then
            $\mathbf{X} = \mathbf{X}_{new}$
            maxF = newF
            improvementTime = 0
        else
            improvementTime = improvementTime +1
        end if
        if improvementTime $>$ patience then
            Break
        end if
    end for
    return($\mathbf{X}$ , testStat )
end procedure
```

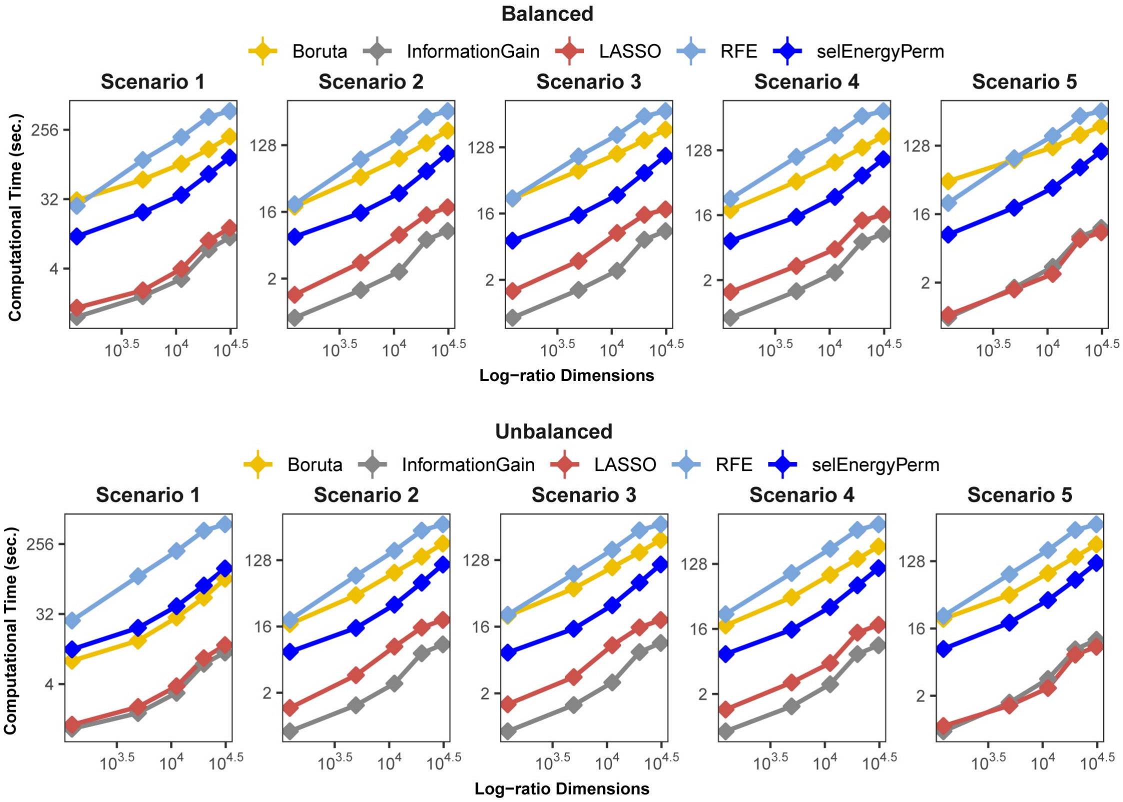

Figure A.1: Feature selection computational time comparisons for balanced and unbalanced sampling designs between SelEnergyPerm, LASSO, RFE, RF, Information Gain, and Boruta across each scenario and dimension. Points are the mean for each experimental condition.

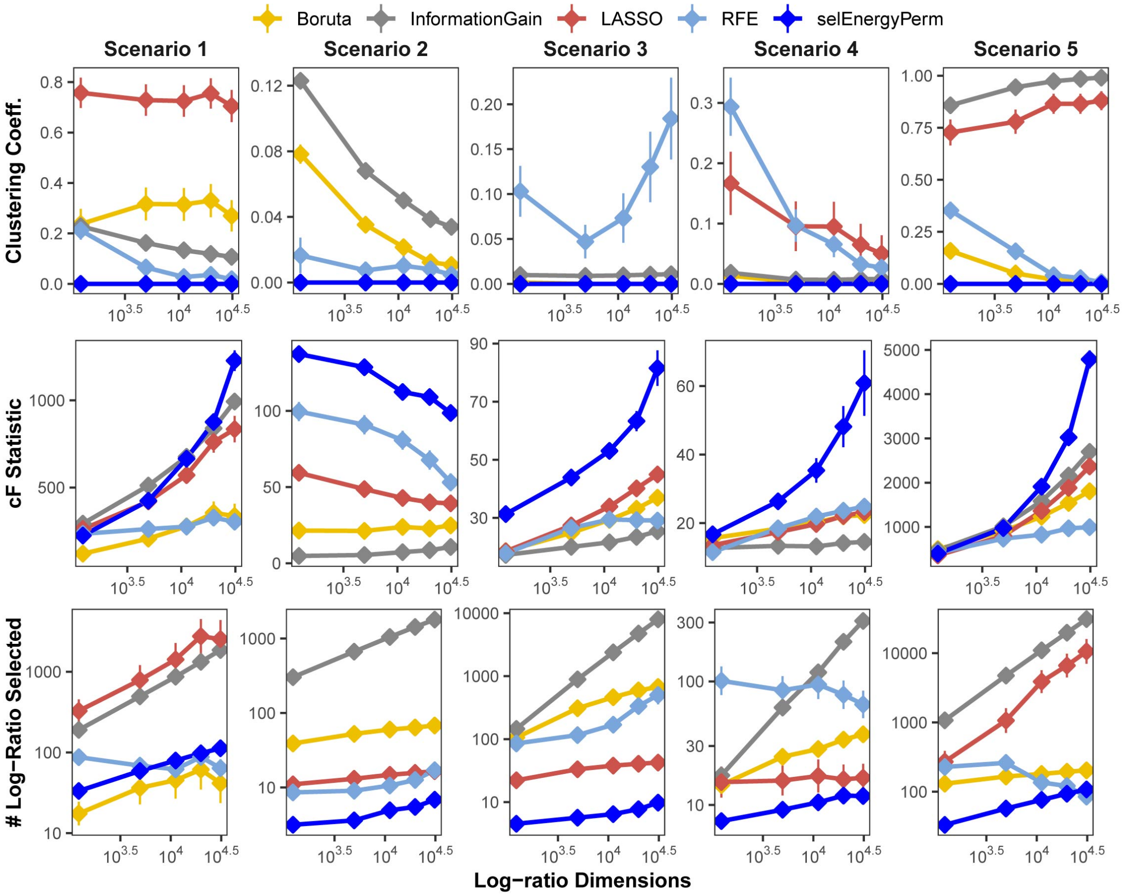

Figure A.2: Comparison of SelEnergyPerm-selected log-ratio subset characteristics with Boruta, Information Gain Filtering, LASSO, and RFE across five simulation scenarios for the unbalanced sampling design. Using 200 simulations for each scenario-dimension by method we assessed: (Top Row) the clustering coefficient of logratio networks formed by selected subsets returned from each method, (Middle Row) the magnitude of the association as measured by the cF-statistic on selected subsets returned from each method, and (Bottom Row) the number of logratios returned by each method. Points are the mean for each experimental condition and error bars indicate 95% confidence interval.

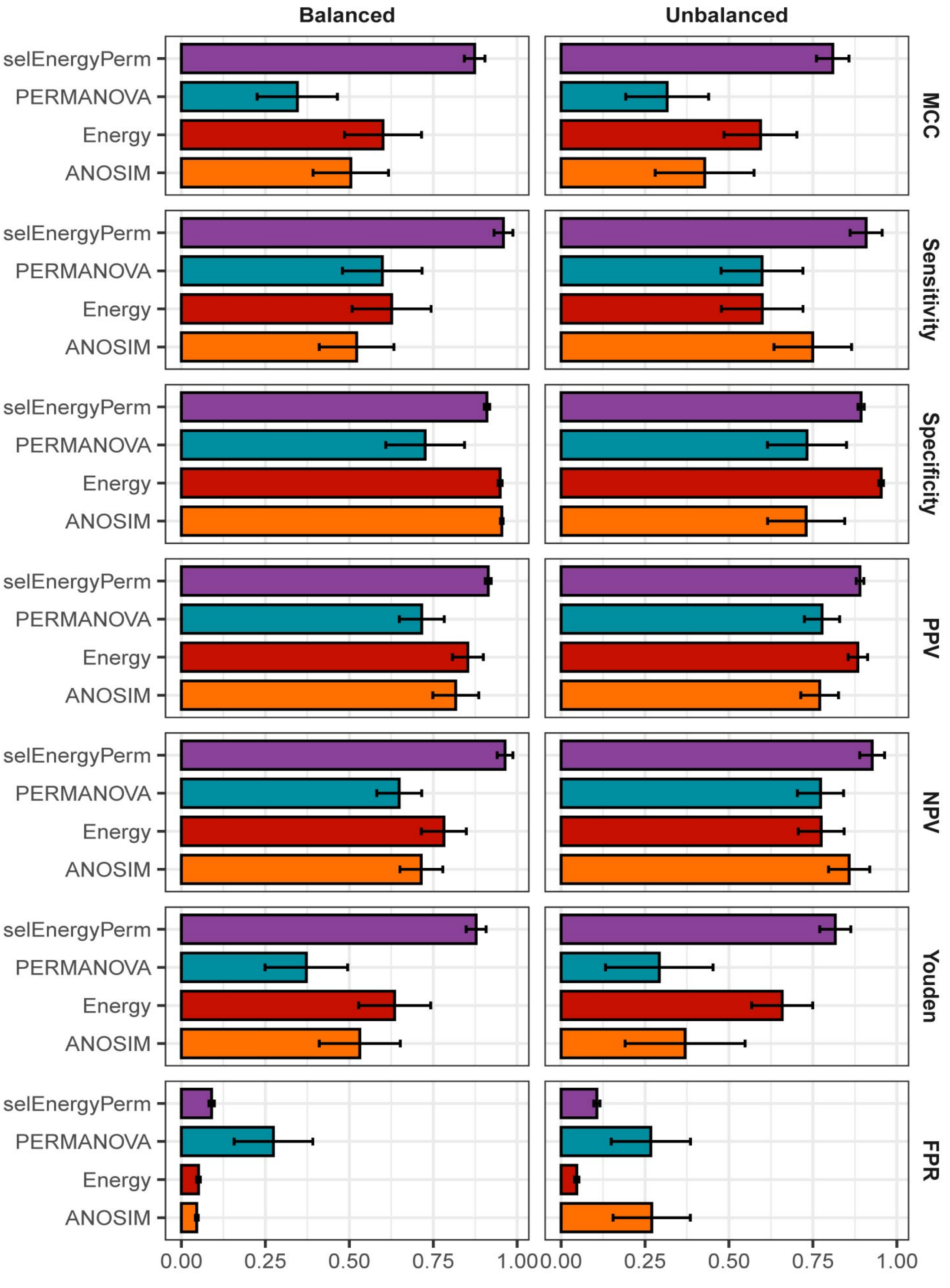

Figure A.3: Overall mean performance comparison for data generated from synthetic distributions aggregated across all scenarios and dimensions using MCC, Sensitivity, Specify, Positive predictive value (PPV), Negative predictive value (NPV), Youden Index, and False Positive Rate (FPR) metric. Error bars indicate standard error.

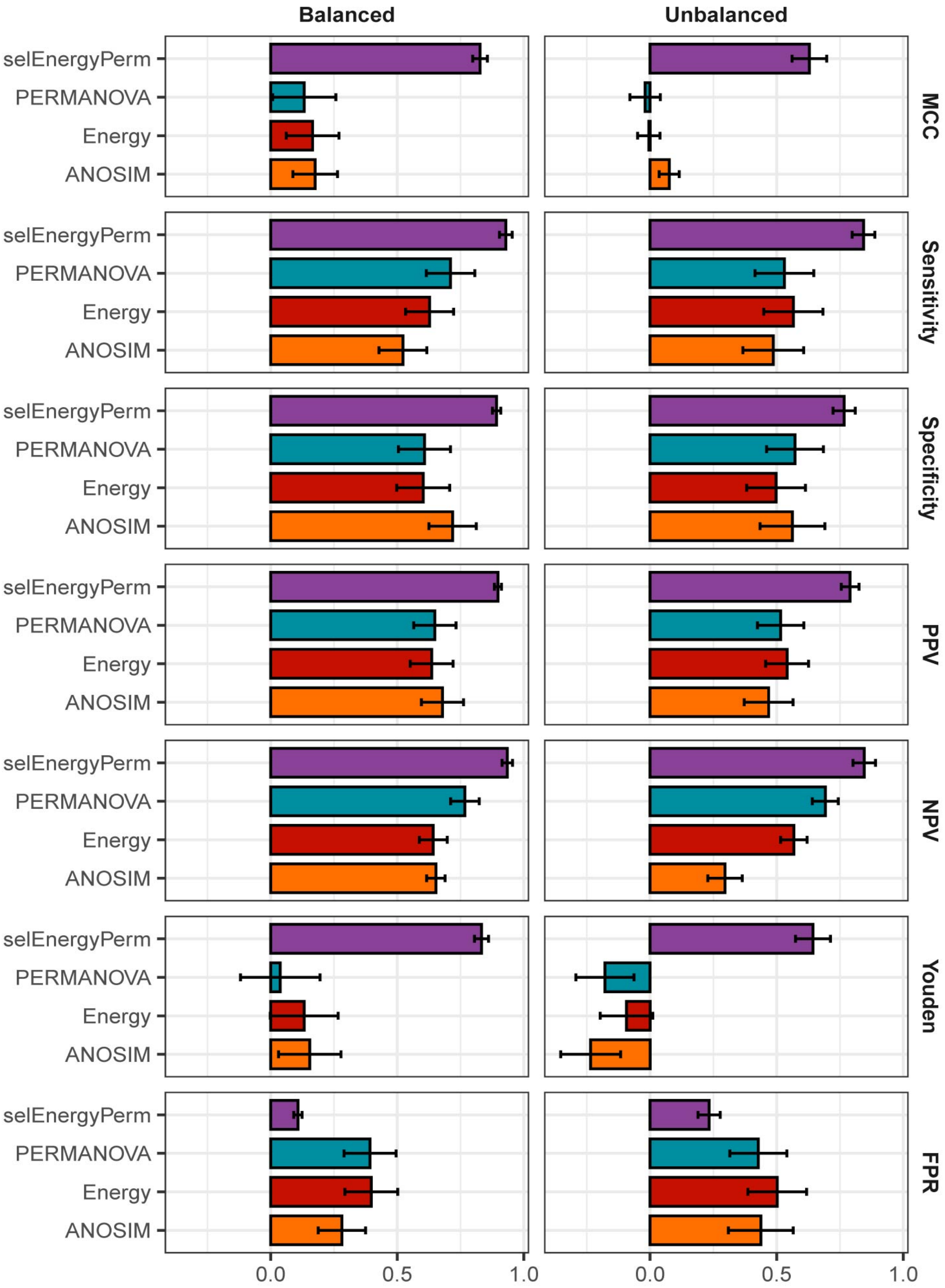

Figure A.4: Overall mean performance comparison for data generated from 16S and WGS synthetic data aggregated across all scenarios and effect levels using MCC, Sensitivity, Specify, Positive predictive value (PPV), Negative predictive value (NPV), Youden Index, and False Positive Rate (FPR) metric. Error bars indicate standard error.

APPENDIX B: CHAPTER 4 SUPPLEMENTAL DATA

Mediator	Limit of Detection	Company	Company Location
Neutrophil Elastase	0.8 ng/mL	Thermo Fisher Scientific (Invitrogen)	Waltham, MA
Total IgA	1.6 ng/mL	Thermo Fisher Scientific (Invitrogen)	Waltham, MA
Lactoferrin	156.3 pg/mL	Abcam	Cambridge, UK
Lysozyme	31.25 pg/mL	Abcam	Cambridge, UK
IL-8	3.1 pg/mL	BD Biosciences	San Diego, CA
Beta-Defensin 1	7.8125 pg/mL	LifeSpan Biosciences	Seattle, WA
Beta-Defensin 2	7.8125 pg/mL	LifeSpan Biosciences	Seattle, WA

Table A.1: Commercially available ELISA kits used to measure mediators of host-microbiota interaction.

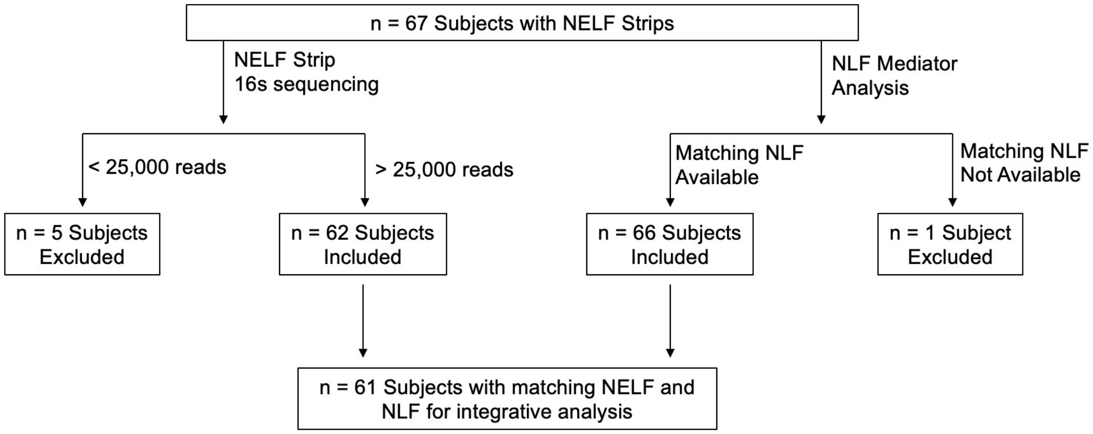

Figure A.5: Flow chart showing inclusion and exclusion criteria for NELF microbiome component, NLF component, and integrative analysis.

Supplementary 1.

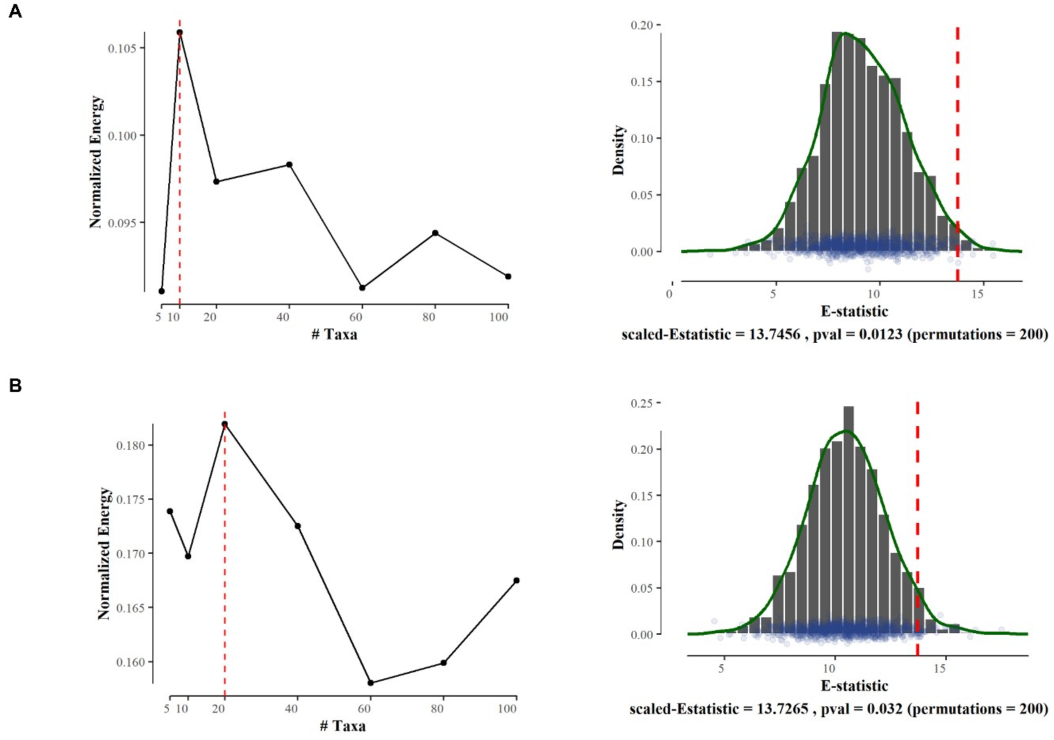

Figure A.6: SelEnergyPerm taxa subset selection and significance results. (A) Left - Selection of number of taxa (t = 10) by normalized energy maximization to test in the final by Sex microbial logratio signature. Right – SelEnergyPerm by Sex microbial signature significance via permutation testing. (B) Left - Selection of number of taxa (t = 20) by normalized energy maximization to test in the final by Subject microbial logratio signature. Right – SelEnergyPerm by Exposure group microbial signature significance results via permutation testing.

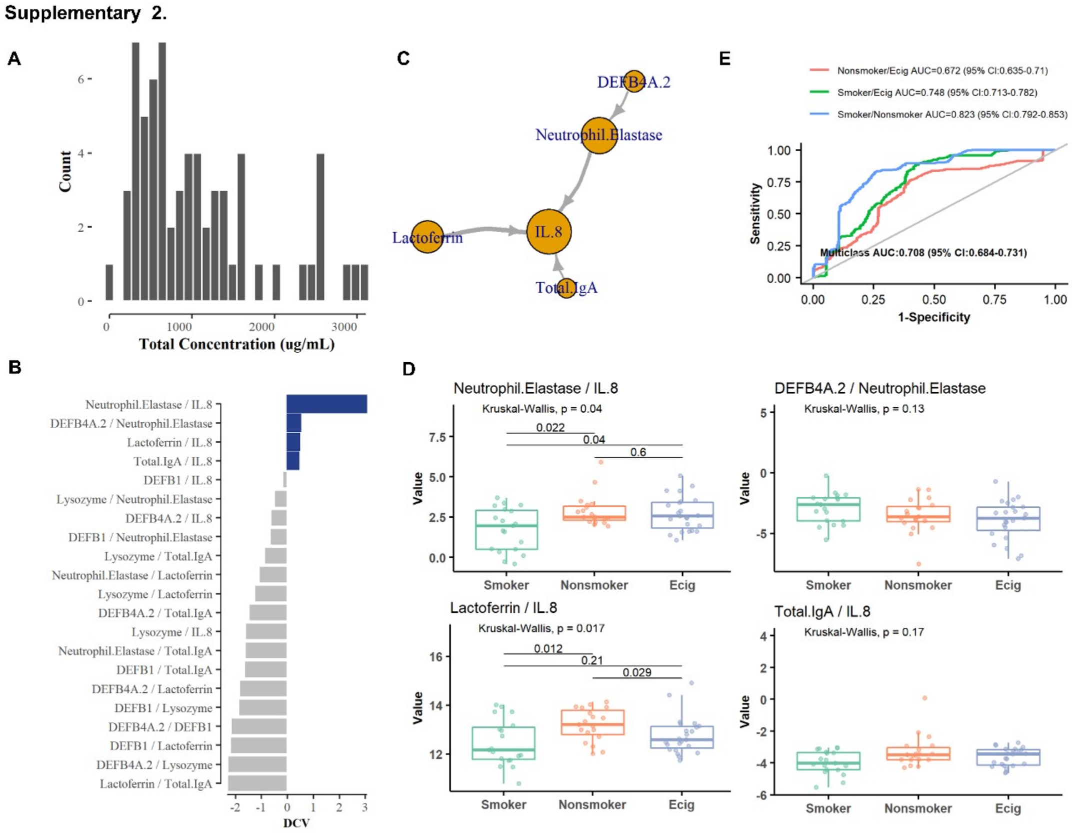

Figure A.7: NLF Mediator Analysis (A) Histogram of total NLF mediator concentrations by sample. (B) DCV scoring of NLF mediator logratios. Grey: DCV < 0; Blue: DCV ≥ 0 (C) Graph representation of DCV derived key NLF mediators. (D) logratio values of key NLF mediator logratios by exposure group with subsequent Wilcoxon Rank-Sum test pairwise comparison displayed. (E) NLF mediator exposure group discrimination via ROC curve displaying the multi-class AUC results of 50 repeats of 10-fold cross-validation using a 2-component PLS-DA model.

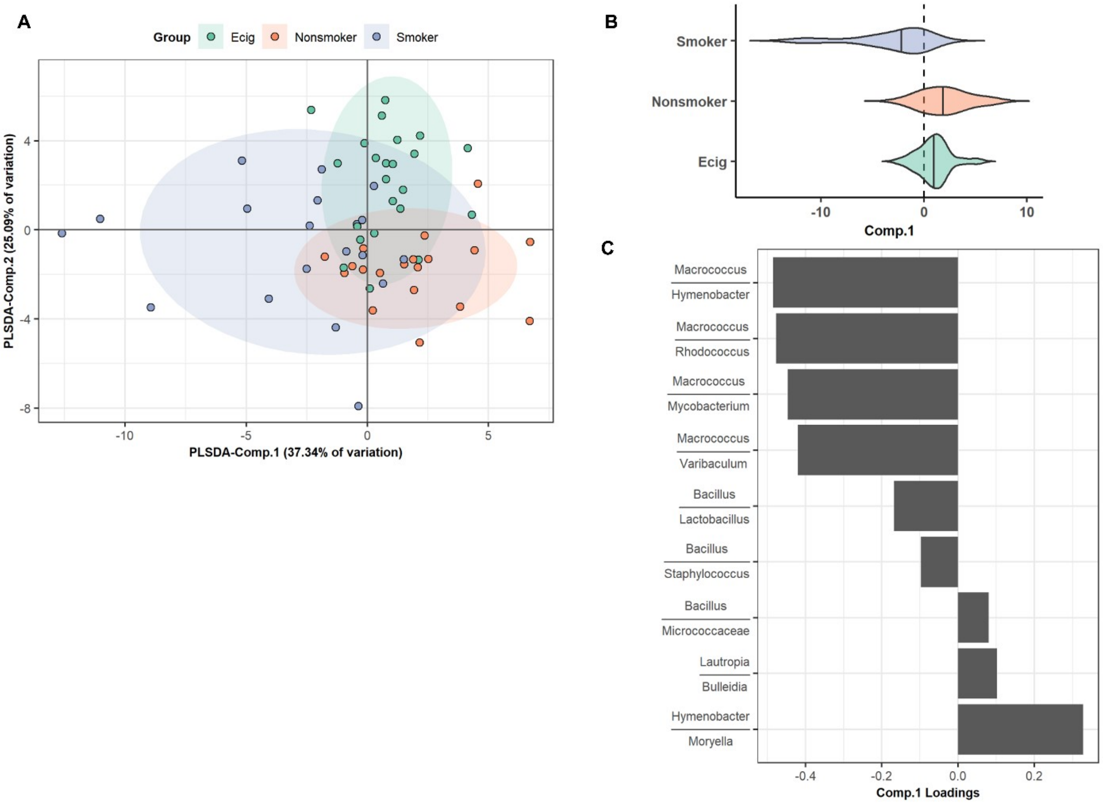

Figure A.8: By Exposure group Nasal Microbial Signature Latent Space Analysis (A) 2-Component PLS-DA Biplot (B) Violin plot with means showing the distribution of first PLS-DA component scores by exposure group (C) PLS-DA loadings on the first component.

BIBLIOGRAPHY

K. Nakayama and T. Nakamura. 15.10 - x-Ray Fluorescence Spectroscopy for Geochemistry. In Heinrich D. Holland and Karl K. Turekian, editors, *Treatise on Geochemistry (Second Edition)*, pages 181–194. Elsevier, Oxford, second edition edition, 2014. ISBN 978-0-08-098300-4. doi: https://doi.org/10.1016/B978-0-08-095975-7.01413-3. URL https://www.sciencedirect.com/science/article/pii/B9780080959757014133.

Karl Pearson. Mathematical contributions to the theory of evolution.—On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, 60(359-367):489–498, January 1897. doi: 10.1098/rspl.1896.0076. URL http://royalsocietypublishing.org/doi/abs/10.1098/rspl.1896.0076. Publisher: Royal Society.

J. Aitchison. The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):139–177, 1982. ISSN 0035-9246. URL https://www.jstor.org/stable/2345821. Publisher: [Royal Statistical Society, Wiley].

J Aitchison. *The Statistical Analysis of Compositional Data*. Chapman & Hall, Ltd., GBR, 1986. ISBN 0412280604.

John Aitchison. A Concise Guide to Compositional Data Analysis. page 134.

J. L. Scealy and A. H. Welsh. Colours and Cocktails: Compositional Data Analysis 2013 Lancaster Lecture. *Australian & New Zealand Journal of Statistics*, 56(2):145–169, 2014. ISSN 1467-842X. doi: 10.1111/anzs.12073. URL http://onlinelibrary.wiley.com/doi/abs/10.1111/anzs.12073. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/anzs.12073.

Vera Pawlowsky-Glahn and Antonella Buccianti. *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons, September 2011. ISBN 978-0-470-71135-4. Google-Books-ID: Ggpj3QeDoKQC.

Gregory B. Gloor, Jean M. Macklaim, Vera Pawlowsky-Glahn, and Juan J. Egozcue. Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology*, 8:2224, 2017. ISSN 1664-302X. doi: 10.3389/fmicb.2017.02224. URL https://www.frontiersin.org/article/10.3389/fmicb.2017.02224.

Michael Greenacre, Marina Martínez-Álvaro, and Agustín Blasco. Compositional Data Analysis of Microbiome and Any-Omics Datasets: A Validation of the Additive Logratio Transformation. *Frontiers in Microbiology*, 12:2625, 2021. ISSN 1664-302X. doi: 10.3389/fmicb.2021.727398. URL https://www.frontiersin.org/article/10.3389/fmicb.2021.727398.

J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal. Isometric Logratio Transformations for Compositional Data Analysis. *Mathematical Geology*, 35(3):279–300, April 2003. ISSN 1573-8868. doi: 10.1023/A:1023818214614. URL https://doi.org/10.1023/A:1023818214614.

Michael Greenacre. Variable Selection in Compositional Data Analysis Using Pairwise Logratios. *Mathematical Geosciences*, 51(5):649–682, July 2019. ISSN 1874-8953. doi: 10.1007/s11004-018-9754-x. URL https://doi.org/10.1007/s11004-018-9754-x.

Gregory B. Gloor and Gregor Reid. Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. *Canadian Journal of Microbiology*, 62(8):692–703, August 2016. ISSN 0008-4166. doi: 10.1139/cjm-2015-0821. URL `https://cdnsciencepub.com/doi/full/10.1139/cjm-2015-0821`. Publisher: NRC Research Press.

Thomas P Quinn, Ionas Erb, Greg Gloor, Cedric Notredame, Mark F Richardson, and Tamsyn M Crowley. A field guide for the compositional analysis of any-omics data. *GigaScience*, 8(9):giz107, September 2019. ISSN 2047-217X. doi: 10.1093/gigascience/giz107. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6755255/`.

Andrew D Fernandes, Jennifer NS Reid, Jean M Macklaim, Thomas A McMurrough, David R Edgell, and Gregory B Gloor. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, 2:15, May 2014. ISSN 2049-2618. doi: 10.1186/2049-2618-2-15. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4030730/`.

Siddhartha Mandal, Will Van Treuren, Richard A. White, Merete Eggesbø, Rob Knight, and Shyamal D. Peddada. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial Ecology in Health and Disease*, 26(1):27663, December 2015a. ISSN null. doi: 10.3402/mehd.v26.27663. URL `https://www.tandfonline.com/doi/abs/10.3402/mehd.v26.27663`. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.3402/mehd.v26.27663.

Josep-Antoni Martín-Fernández, Karel Hron, Matthias Templ, Peter Filzmoser, and Javier Palarea-Albaladejo. Bayesian-multiplicative treatment of count zeros in compositional data sets. *Statistical Modelling*, 15(2):134–158, April 2015. ISSN 1471-082X. doi: 10.1177/1471082X14535524. URL `https://doi.org/10.1177/1471082X14535524`. Publisher: SAGE Publications India.

J A Martın-Fernandez, C Barcelo-Vidal, and V Pawlowsky-Glahn. Dealing with Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation. *Mathematical Geology*, page 26, 2003.

Rebeca Martín, Sylvie Miquel, Philippe Langella, and Luis G. Bermúdez-Humarán. The role of metagenomics in understanding the human microbiome in health and disease. *Virulence*, 5(3):413–423, April 2014. ISSN 2150-5594. doi: 10.4161/viru.27864. URL `https://doi.org/10.4161/viru.27864`. Publisher: Taylor & Francis _eprint: https://doi.org/10.4161/viru.27864.

Ravi Ranjan, Asha Rani, Ahmed Metwally, Halvor S. McGee, and David L. Perkins. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochemical and Biophysical Research Communications*, 469(4):967–977, January 2016. ISSN 1090-2104. doi: 10.1016/j.bbrc.2015.12.083.

Vancheswaran Gopalakrishnan, Beth A. Helmink, Christine N. Spencer, Alexandre Reuben, and Jennifer A. Wargo. The Influence of the Gut Microbiome on Cancer, Immunity, and Cancer Immunotherapy. *Cancer Cell*, 33(4):570–580, April 2018. ISSN 1535-6108. doi: 10.1016/j.ccell.2018.03.015. URL `https://www.sciencedirect.com/science/article/pii/S153561081830120X`.

Chaysavanh Manichanh, Natalia Borruel, Francesc Casellas, and Francisco Guarner. The gut microbiota in IBD. *Nature Reviews Gastroenterology & Hepatology*, 9(10):599–608, October 2012. ISSN 1759-5053. doi: 10.1038/nrgastro.2012.152. URL http://www.nature.com/articles/nrgastro.2012.152. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 10 Primary_atype: Reviews Publisher: Nature Publishing Group Subject_term: Dysbiosis;Inflammatory bowel disease;Microbiota;Pathogenesis Subject_term_id: dysbiosis;inflammatory-bowel-disease;microbiota;pathogenesis.

Nan Qin, Fengling Yang, Ang Li, Edi Prifti, Yanfei Chen, Li Shao, Jing Guo, Emmanuelle Le Chatelier, Jian Yao, Lingjiao Wu, Jiawei Zhou, Shujun Ni, Lin Liu, Nicolas Pons, Jean Michel Batto, Sean P. Kennedy, Pierre Leonard, Chunhui Yuan, Wenchao Ding, Yuanting Chen, Xinjun Hu, Beiwen Zheng, Guirong Qian, Wei Xu, S. Dusko Ehrlich, Shusen Zheng, and Lanjuan Li. Alterations of the human gut microbiome in liver cirrhosis. *Nature*, 513(7516):59–64, September 2014. ISSN 1476-4687. doi: 10.1038/nature13568. URL http://www.nature.com/articles/nature13568. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7516 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Biomarkers;Genome informatics;Microbiology Subject_term_id: biomarkers;genome-informatics;microbiology.

Jane A. Foster and Karen-Anne McVey Neufeld. Gut–brain axis: how the microbiome influences anxiety and depression. *Trends in Neurosciences*, 36(5):305–312, May 2013. ISSN 01662236. doi: 10.1016/j.tins.2013.01.005. URL https://linkinghub.elsevier.com/retrieve/pii/S0166223613000088.

Heidi H. Kong, Julia Oh, Clay Deming, Sean Conlan, Elizabeth A. Grice, Melony A. Beatson, Effie Nomicos, Eric C. Polley, Hirsh D. Komarow, Patrick R. Murray, Maria L. Turner, and Julia A. Segre. Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis. *Genome Research*, 22(5):850–859, May 2012. ISSN 1088-9051. doi: 10.1101/gr.131029.111. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3337431/.

Floyd E. Dewhirst, Tuste Chen, Jacques Izard, Bruce J. Paster, Anne C. R. Tanner, Wen-Han Yu, Abirami Lakshmanan, and William G. Wade. The human oral microbiome. *Journal of Bacteriology*, 192(19): 5002–5017, October 2010. ISSN 1098-5530. doi: 10.1128/JB.00542-10.

Michael T. Wilson and Daniel L. Hamilos. The Nasal and Sinus Microbiome in Health and Disease. *Current Allergy and Asthma Reports*, 14(12):485, October 2014. ISSN 1534-6315. doi: 10.1007/s11882-014-0485-x. URL https://doi.org/10.1007/s11882-014-0485-x.

Joseph P. Zackular, Mary A. M. Rogers, Mack T. Ruffin, and Patrick D. Schloss. The Human Gut Microbiome as a Screening Tool for Colorectal Cancer. *Cancer Prevention Research*, 7(11):1112–1121, November 2014. ISSN 1940-6207, 1940-6215. doi: 10.1158/1940-6207.CAPR-14-0129. URL https://cancerpreventionresearch.aacrjournals.org/content/7/11/1112. Publisher: American Association for Cancer Research Section: Research Article.

Robert Schlaberg. Microbiome Diagnostics. *Clinical Chemistry*, 66(1):68–76, January 2020. ISSN 0009-9147. doi: 10.1373/clinchem.2019.303248. URL https://doi.org/10.1373/clinchem.2019.303248.

J. Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D. Bushman, Elizabeth K. Costello, Noah Fierer, Antonio Gonzalez Peña, Julia K. Goodrich, Jeffrey I. Gordon, Gavin A. Huttley, Scott T. Kelley, Dan Knights, Jeremy E. Koenig, Ruth E. Ley, Catherine A. Lozupone, Daniel McDonald, Brian D. Muegge, Meg Pirrung, Jens Reeder, Joel R. Sevinsky, Peter J. Turnbaugh, William A. Walters, Jeremy Widmann, Tanya Yatsunenko, Jesse Zaneveld, and Rob Knight. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5):335–336, May

2010. ISSN 1548-7105. doi: 10.1038/nmeth.f.303. URL http://www.nature.com/articles/nmeth.f.303. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 5 Primary_atype: Correspondence Publisher: Nature Publishing Group Subject_term: Microbial ecology;Next-generation sequencing;Software Subject_term_id: microbial-ecology;next-generation-sequencing;software.

Schloss Patrick D., Westcott Sarah L., Ryabin Thomas, Hall Justine R., Hartmann Martin, Hollister Emily B., Lesniewski Ryan A., Oakley Brian B., Parks Donovan H., Robinson Courtney J., Sahl Jason W., Stres Blaz, Thallinger Gerhard G., Van Horn David J., and Weber Carolyn F. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology*, 75(23):7537–7541, December 2009. doi: 10.1128/AEM.01541-09. URL http://journals.asm.org/doi/10.1128/AEM.01541-09. Publisher: American Society for Microbiology.

Duy Tin Truong, Eric A. Franzosa, Timothy L. Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods*, 12(10):902–903, October 2015. ISSN 1548-7105. doi: 10.1038/nmeth.3589. URL http://www.nature.com/articles/nmeth.3589. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 10 Primary_atype: Correspondence Publisher: Nature Publishing Group Subject_term: Classification and taxonomy;Metagenomics;Software Subject_term_id: classification-and-taxonomy;metagenomics;software.

Derrick E. Wood, Jennifer Lu, and Ben Langmead. Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20(1):257, November 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1891-0. URL https://doi.org/10.1186/s13059-019-1891-0.

Sophie Weiss, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, Jesse R. Zaneveld, Yoshiki Vázquez-Baeza, Amanda Birmingham, Embriette R. Hyde, and Rob Knight. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5, March 2017. ISSN 2049-2618. doi: 10.1186/s40168-017-0237-y. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5335496/.

Michael Greenacre and Paul Lewi. Distributional Equivalence and Subcompositional Coherence in the Analysis of Compositional Data, Contingency Tables and Ratio-Scale Measurements. *Journal of Classification*, 26(1):29–54, April 2009. ISSN 1432-1343. doi: 10.1007/s00357-009-9027-y. URL https://doi.org/10.1007/s00357-009-9027-y.

David R Lovell, Xin-Yi Chua, and Annette McGrath. Counts: an outstanding challenge for log-ratio analysis of compositional data in the molecular biosciences. *NAR Genomics and Bioinformatics*, 2(2), June 2020. ISSN 2631-9268. doi: 10.1093/nargab/lqaa040. URL https://doi.org/10.1093/nargab/lqaa040.

Siddhartha Mandal, Will Van Treuren, Richard A. White, Merete Eggesbø, Rob Knight, and Shyamal D. Peddada. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial Ecology in Health and Disease*, 26:10.3402/mehd.v26.27663, May 2015b. ISSN 0891-060X. doi: 10.3402/mehd.v26.27663. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4450248/.

Huang Lin and Shyamal Das Peddada. Analysis of compositions of microbiomes with bias correction. *Nature Communications*, 11(1):3514, July 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17041-7. URL http://www.nature.com/articles/s41467-020-17041-7. Bandiera_abtest: a

Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Computational biology and bioinformatics;Ecology;Microbiology Subject_term_id: computational-biology-and-bioinformatics;ecology;microbiology.

Mehdi Layeghifard, David M. Hwang, and David S. Guttman. Disentangling Interactions in the Microbiome: A Network Perspective. *Trends in Microbiology*, 25(3):217–228, March 2017. ISSN 0966842X. doi: 10.1016/j.tim.2016.11.008. URL `https://linkinghub.elsevier.com/retrieve/pii/S0966842X16301858`.

Nancy A. Obuchowski. Multivariate statistical methods. *AJR. American journal of roentgenology*, 185(2): 299–309, August 2005. ISSN 0361-803X. doi: 10.2214/ajr.185.2.01850299.

Marti J. Anderson. Permutational Multivariate Analysis of Variance (PERMANOVA). In *Wiley StatsRef: Statistics Reference Online*, pages 1–15. American Cancer Society, 2017. ISBN 978-1-118-44511-2. Type: Book Section.

K. R. Clarke. Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology*, 18(1):117–143, 1993. ISSN 1442-9993. doi: https://doi.org/10.1111/j.1442-9993.1993.tb00438.x. URL `http://onlinelibrary.wiley.com/doi/abs/10.1111/j.1442-9993.1993.tb00438.x`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1442-9993.1993.tb00438.x.

Maria L. Rizzo and Gábor J. Székely. Energy distance. *Wiley Interdisciplinary Reviews: Computational Statistics*, 8(1):27–38, January 2016. ISSN 19395108. doi: 10.1002/wics.1375. URL `http://doi.wiley.com/10.1002/wics.1375`.

Knut Baumann. Cross-validation as the objective function for variable-selection techniques. *TrAC Trends in Analytical Chemistry*, 22(6):395–406, June 2003. ISSN 0165-9936. doi: 10.1016/S0165-9936(03)00607-1. URL `https://www.sciencedirect.com/science/article/pii/S0165993603006071`.

Fredrik Lindgren, Björn Hansen, Walter Karcher, Michael Sjöström, and Lennart Eriksson. Model validation by permutation tests: Applications to variable selection. *Journal of Chemometrics*, 10(5-6): 521–532, 1996. ISSN 1099-128X. doi: 10.1002/(SICI)1099-128X(199609)10:5/6⟨521::AID-CEM448⟩3.0.CO;2-J. URL `https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291099-128X%28199609%2910%3A5/6%3C521%3A%3AAID-CEM448%3E3.0.CO%3B2-J`. _eprint: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291099-128X%28199609%2910%3A5/6%3C521%3A%3AAID-CEM448%3E3.0.CO%3B2-J.

Chong Wu, Jun Chen, Junghi Kim, and Wei Pan. An adaptive association test for microbiome data. *Genome Medicine*, 8(1):56, May 2016. ISSN 1756-994X. doi: 10.1186/s13073-016-0302-3. URL `https://doi.org/10.1186/s13073-016-0302-3`.

Hyunwook Koh and Ni Zhao. A powerful microbial group association test based on the higher criticism analysis for sparse microbial association signals. *Microbiome*, 8(1):63, May 2020. ISSN 2049-2618. doi: 10.1186/s40168-020-00834-9.

Susan Wei, Chihoon Lee, Lindsay Wichers, and J. S. Marron. Direction-Projection-Permutation for High-Dimensional Hypothesis Tests. *Journal of Computational and Graphical Statistics*, 25(2):549–569, April 2016. ISSN 1061-8600. doi: 10.1080/10618600.2015.1027773. URL `https://doi.org/10.1080/10618600.2015.1027773`. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10618600.2015.1027773.

Javier Palarea-Albaladejo and Josep Antoni Martín-Fernández. zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems*, 143:85–96, April 2015. ISSN 0169-7439. doi: 10.1016/j.chemolab.2015.02.019. URL `https://www.sciencedirect.com/science/article/pii/S0169743915000490`.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008. ISSN 1467-9868. doi: https://doi.org/10.1111/j.1467-9868.2008.00674.x. URL `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2008.00674.x`. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2008.00674.x.

Maria L. Rizzo and Gábor J. Székely. DISCO analysis: A nonparametric extension of analysis of variance. *The Annals of Applied Statistics*, 4(2):1034–1055, June 2010. ISSN 1932-6157, 1941-7330. doi: 10.1214/09-AOAS245. URL `http://projecteuclid.org/journals/annals-of-applied-statistics/volume-4/issue-2/DISCO-analysis-A-nonparametric-extension-of-analysis-of-variance/10.1214/09-AOAS245.full`. Publisher: Institute of Mathematical Statistics.

Marti J. Anderson. Distance-based tests for homogeneity of multivariate dispersions. *Biometrics*, 62(1): 245–253, March 2006. ISSN 0006-341X. doi: 10.1111/j.1541-0420.2005.00440.x.

Michael D. Ernst. Permutation Methods: A Basis for Exact Inference. *Statistical Science*, 19(4):676–685, 2004. ISSN 0883-4237, 2168-8745. doi: 10.1214/088342304000000396. Type: Journal Article.

Julia K. Goodrich, Jillian L. Waters, Angela C. Poole, Jessica L. Sutter, Omry Koren, Ran Blekhman, Michelle Beaumont, William Van Treuren, Rob Knight, Jordana T. Bell, Timothy D. Spector, Andrew G. Clark, and Ruth E. Ley. Human Genetics Shape the Gut Microbiome. *Cell*, 159(4):789–799, November 2014. ISSN 0092-8674. doi: 10.1016/j.cell.2014.09.053. URL `https://www.sciencedirect.com/science/article/pii/S0092867414012410`.

Claire Duvallet, Sean Gibbons, Thomas Gurry, Rafael Irizarry, and Eric Alm. MicrobiomeHD: the human gut microbiome in health and disease, August 2017. URL `https://zenodo.org/record/1146764`. Type: dataset.

David Zeevi, Tal Korem, Niv Zmora, David Israeli, Daphna Rothschild, Adina Weinberger, Orly Ben-Yacov, Dar Lador, Tali Avnit-Sagi, Maya Lotan-Pompan, Jotham Suez, Jemal Ali Mahdi, Elad Matot, Gal Malka, Noa Kosower, Michal Rein, Gili Zilberman-Schapira, Lenka Dohnalová, Meirav Pevsner-Fischer, Rony Bikovsky, Zamir Halpern, Eran Elinav, and Eran Segal. Personalized Nutrition by Prediction of Glycemic Responses. *Cell*, 163(5):1079–1094, November 2015. ISSN 1097-4172. doi: 10.1016/j.cell.2015.11.001.

Edoardo Pasolli, Lucas Schiffer, Paolo Manghi, Audrey Renson, Valerie Obenchain, Duy Tin Truong, Francesco Beghini, Faizan Malik, Marcel Ramos, Jennifer B. Dowd, Curtis Huttenhower, Martin Morgan, Nicola Segata, and Levi Waldron. Accessible, curated metagenomic data through ExperimentHub. *Nature Methods*, 14(11):1023–1024, November 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4468. URL `https://www.nature.com/articles/nmeth.4468`.

Matteo Calgaro, Chiara Romualdi, Levi Waldron, Davide Risso, and Nicola Vitulo. Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data. *Genome Biology*, 21(1):1–31, December 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-02104-1. URL `http://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02104-1`. Number: 1 Publisher: BioMed Central.

Miron B. Kursa, Aleksander Jankowski, and Witold R. Rudnicki. Boruta – A System for Feature Selection. *Fundamenta Informaticae*, 101(4):271–285, 2010. ISSN 01692968. doi: 10.3233/FI-2010-288. URL `https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/FI-2010-288`.

Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. URL `http://www.jstor.org.libproxy.lib.unc.edu/stable/2346178`. Publisher: [Royal Statistical Society, Wiley].

JOHN T. KENT. Information gain and a general measure of correlation. *Biometrika*, 70(1):163–173, April 1983. ISSN 0006-3444. doi: 10.1093/biomet/70.1.163. URL `https://doi.org/10.1093/biomet/70.1.163`.

Pablo M. Granitto, Cesare Furlanello, Franco Biasioli, and Flavia Gasperi. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, 83(2):83–90, September 2006. ISSN 0169-7439. doi: 10.1016/j.chemolab.2006.01.007. URL `https://www.sciencedirect.com/science/article/pii/S0169743906000232`.

Marti J. Anderson and Daniel C. I. Walsh. PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing? *Ecological Monographs*, 83(4): 557–574, 2013. ISSN 1557-7015. doi: https://doi.org/10.1890/12-2010.1. URL `http://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/12-2010.1`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1890/12-2010.1.

Joseph N. Paulson, Brent L. Williams, Christine Hehnly, Nischay Mishra, Shamim A. Sinnar, Lijun Zhang, Paddy Ssentongo, Edith Mbabazi-Kabachelor, Dona S. S. Wijetunge, Benjamin von Bredow, Ronnie Mulondo, Julius Kiwanuka, Francis Bajunirwe, Joel Bazira, Lisa M. Bebell, Kathy Burgoine, Mara Couto-Rodriguez, Jessica E. Ericson, Tim Erickson, Matthew Ferrari, Melissa Gladstone, Cheng Guo, Murali Haran, Mady Hornig, Albert M. Isaacs, Brian Nsubuga Kaaya, Sheila M. Kangere, Abhaya V. Kulkarni, Elias Kumbakumba, Xiaoxiao Li, David D. Limbrick, Joshua Magombe, Sarah U. Morton, John Mugamba, James Ng, Peter Olupot-Olupot, Justin Onen, Mallory R. Peterson, Farrah Roy, Kathryn Sheldon, Reid Townsend, Andrew D. Weeks, Andrew J. Whalen, John Quackenbush, Peter Ssenyonga, Michael Y. Galperin, Mathieu Almeida, Hannah Atkins, Benjamin C. Warf, W. Ian Lipkin, James R. Broach, and Steven J. Schiff. Paenibacillus infection with frequent viral coinfection contributes to postinfectious hydrocephalus in Ugandan infants. *Science Translational Medicine*, 12(563):eaba0565, September 2020. ISSN 1946-6242. doi: 10.1126/scitranslmed.aba0565.

Francislon S Oliveira, John Brestelli, Shon Cade, Jie Zheng, John Iodice, Steve Fischer, Cristina Aurrecoechea, Jessica C Kissinger, Brian P Brunk, Christian J Stoeckert, Jr, Gabriel R Fernandes, David S Roos, and Daniel P Beiting. MicrobiomeDB: a systems biology platform for integrating, mining and analyzing microbiome experiments. *Nucleic Acids Research*, 46(D1):D684–D691, January 2018. ISSN 0305-1048. doi: 10.1093/nar/gkx1027. URL `https://doi.org/10.1093/nar/gkx1027`.

Nicholas A. Bokulich, Jennifer Chung, Thomas Battaglia, Nora Henderson, Melanie Jay, Huilin Li, Arnon D. Lieber, Fen Wu, Guillermo I. Perez-Perez, Yu Chen, William Schweizer, Xuhui Zheng, Monica Contreras, Maria Gloria Dominguez-Bello, and Martin J. Blaser. Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Science Translational Medicine*, 8(343):343ra82–343ra82, June 2016. ISSN 1946-6234, 1946-6242. doi: 10.1126/scitranslmed.aad7121. URL `https://stm.sciencemag.org/lookup/doi/10.1126/scitranslmed.aad7121`.

Antonio Gonzalez, Jose A. Navas-Molina, Tomasz Kosciolek, Daniel McDonald, Yoshiki Vázquez-Baeza, Gail Ackermann, Jeff DeReus, Stefan Janssen, Austin D. Swafford, Stephanie B. Orchanian, Jon G. Sanders, Joshua Shorenstein, Hannes Holste, Semar Petrus, Adam Robbins-Pianka, Colin J. Brislawn, Mingxun Wang, Jai Ram Rideout, Evan Bolyen, Matthew Dillon, J. Gregory Caporaso, Pieter C. Dorrestein, and Rob Knight. Qiita: rapid, web-enabled microbiome meta-analysis. *Nature Methods*, 15 (10):796–798, October 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0141-9. URL `https://www.nature.com/articles/s41592-018-0141-9`. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 10 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Microbial communities;Data integration;Data mining;Data publication and archiving;Databases Subject_term_id: communities;data-integration;data-mining;data-publication-and-archiving;databases.

Lita M. Proctor, Heather H. Creasy, Jennifer M. Fettweis, Jason Lloyd-Price, Anup Mahurkar, Wenyu Zhou, Gregory A. Buck, Michael P. Snyder, Jerome F. Strauss, George M. Weinstock, Owen White, Curtis Huttenhower, and The Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project. *Nature*, 569(7758):641–648, May 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1238-8. URL `https://www.nature.com/articles/s41586-019-1238-8`. Number: 7758 Publisher: Nature Publishing Group.

Jason Lloyd-Price, Cesar Arze, Ashwin N. Ananthakrishnan, Melanie Schirmer, Julian Avila-Pacheco, Tiffany W. Poon, Elizabeth Andrews, Nadim J. Ajami, Kevin S. Bonham, Colin J. Brislawn, David Casero, Holly Courtney, Antonio Gonzalez, Thomas G. Graeber, A. Brantley Hall, Kathleen Lake, Carol J. Landers, Himel Mallick, Damian R. Plichta, Mahadev Prasad, Gholamali Rahnavard, Jenny Sauk, Dmitry Shungin, Yoshiki Vázquez-Baeza, Richard A. White, Jonathan Braun, Lee A. Denson, Janet K. Jansson, Rob Knight, Subra Kugathasan, Dermot P. B. McGovern, Joseph F. Petrosino, Thaddeus S. Stappenbeck, Harland S. Winter, Clary B. Clish, Eric A. Franzosa, Hera Vlamakis, Ramnik J. Xavier, and Curtis Huttenhower. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, 569(7758):655–662, May 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1237-9. URL `http://www.nature.com/articles/s41586-019-1237-9`. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 7758 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Crohn's disease;Microbiome;Systems analysis;Ulcerative colitis Subject_term_id: crohns-disease;microbiome;systems-analysis;ulcerative-colitis.

Tommi Vatanen, Aleksandar D. Kostic, Eva d'Hennezel, Heli Siljander, Eric A. Franzosa, Moran Yassour, Raivo Kolde, Hera Vlamakis, Timothy D. Arthur, Anu-Maaria Hämäläinen, Aleksandr Peet, Vallo Tillmann, Raivo Uibo, Sergei Mokurov, Natalya Dorshakova, Jorma Ilonen, Suvi M. Virtanen, Susanne J. Szabo, Jeffrey A. Porter, Harri Lähdesmäki, Curtis Huttenhower, Dirk Gevers, Thomas W. Cullen, Mikael Knip, and Ramnik J. Xavier. Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans. *Cell*, 165(4):842–853, May 2016. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2016.04.007. URL https://www.cell.com/cell/abstract/S0092-8674(16)30398-1. Publisher: Elsevier.

Antoni Susin, Yiwen Wang, Kim-Anh Lê Cao, and M. Luz Calle. Variable selection in microbiome compositional data analysis. *NAR Genomics and Bioinformatics*, 2(2), June 2020. doi: 10.1093/nargab/lqaa029. URL https://academic.oup.com/nargab/article/2/2/lqaa029/5836692. Publisher: Oxford Academic.

Laura Judith Marcos-Zambrano, Kanita Karaduzovic-Hadziabdic, Tatjana Loncar Turukalo, Piotr Przymus, Vladimir Trajkovik, Oliver Aasmets, Magali Berland, Aleksandra Gruca, Jasminka Hasic, Karel Hron, Thomas Klammsteiner, Mikhail Kolev, Leo Lahti, Marta B. Lopes, Victor Moreno, Irina Naskinova, Elin Org, Inês Paciência, Georgios Papoutsoglou, Rajesh Shigdel, Blaz Stres, Baiba Vilne, Malik Yousef, Eftim Zdravevski, Ioannis Tsamardinos, Enrique Carrillo de Santa Pau, Marcus J. Claesson, Isabel Moreno-Indias, and Jaak Truu. Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment. *Frontiers in Microbiology*, 12:313, 2021. ISSN 1664-302X. doi: 10.3389/fmicb.2021.634511. URL https://www.frontiersin.org/article/10.3389/fmicb.2021.634511.

Edoardo Pasolli, Duy Tin Truong, Faizan Malik, Levi Waldron, and Nicola Segata. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. *PLOS Computational Biology*, 12(7):e1004977, July 2016. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004977. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004977.

J. Rivera-Pinto, J. J. Egozcue, V. Pawlowsky-Glahn, R. Paredes, M. Noguera-Julian, and M. L. Calle. Balances: a New Perspective for Microbiome Analysis. *mSystems*, 3(4):e00053–18, August 2018. ISSN 2379-5077. doi: 10.1128/mSystems.00053-18. URL https://msystems.asm.org/content/3/4/e00053-18.

Matthew R. Olm, Nicholas Bhattacharya, Alexander Crits-Christoph, Brian A. Firek, Robyn Baker, Yun S. Song, Michael J. Morowitz, and Jillian F. Banfield. Necrotizing enterocolitis is preceded by increased gut bacterial replication, Klebsiella, and fimbriae-encoding bacteria. *Science Advances*, 5(12):eaax5727, December 2019. ISSN 2375-2548. doi: 10.1126/sciadv.aax5727.

Qiang Feng, Suisha Liang, Huijue Jia, Andreas Stadlmayr, Longqing Tang, Zhou Lan, Dongya Zhang, Huihua Xia, Xiaoying Xu, Zhuye Jie, Lili Su, Xiaoping Li, Xin Li, Junhua Li, Liang Xiao, Ursula Huber-Schönauer, David Niederseer, Xun Xu, Jumana Yousuf Al-Aama, Huanming Yang, Jian Wang, Karsten Kristiansen, Manimozhiyan Arumugam, Herbert Tilg, Christian Datz, and Jun Wang. Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nature Communications*, 6: 6528, March 2015. ISSN 2041-1723. doi: 10.1038/ncomms7528.

Ankit Gupta, Darshan B. Dhakan, Abhijit Maji, Rituja Saxena, Vishnu Prasoodanan P K, Shruti Mahajan, Joby Pulikkan, Jacob Kurian, Andres M. Gomez, Joy Scaria, Katherine R. Amato, Ashok K. Sharma, and Vineet K. Sharma. Association of Flavonifractor plautii, a Flavonoid-Degrading Bacterium, with the Gut Microbiome of Colorectal Cancer Patients in India. *mSystems*, 4(6):e00438–19, November 2019. ISSN 2379-5077. doi: 10.1128/mSystems.00438-19.

Geoffrey D. Hannigan, Melissa B. Duhaime, Mack T. Ruffin, Charlie C. Koumpouras, and Patrick D. Schloss. Diagnostic Potential and Interactive Dynamics of the Colorectal Cancer Virome. *mBio*, 9(6): e02248–18, November 2018. ISSN 2150-7511. doi: 10.1128/mBio.02248-18.

Jakob Wirbel, Paul Theodor Pyl, Ece Kartal, Konrad Zych, Alireza Kashani, Alessio Milanese, Jonas S. Fleck, Anita Y. Voigt, Albert Palleja, Ruby Ponnudurai, Shinichi Sunagawa, Luis Pedro Coelho, Petra Schrotz-King, Emily Vogtmann, Nina Habermann, Emma Niméus, Andrew M. Thomas, Paolo Manghi, Sara Gandini, Davide Serrano, Sayaka Mizutani, Hirotsugu Shiroma, Satoshi Shiba, Tatsuhiro Shibata, Shinichi Yachida, Takuji Yamada, Levi Waldron, Alessio Naccarati, Nicola Segata, Rashmi Sinha, Cornelia M. Ulrich, Hermann Brenner, Manimozhiyan Arumugam, Peer Bork, and Georg Zeller. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nature Medicine*, 25(4):679–689, April 2019. ISSN 1546-170X. doi: 10.1038/s41591-019-0406-6.

Andrew Maltez Thomas, Paolo Manghi, Francesco Asnicar, Edoardo Pasolli, Federica Armanini, Moreno Zolfo, Francesco Beghini, Serena Manara, Nicolai Karcher, Chiara Pozzi, Sara Gandini, Davide Serrano, Sonia Tarallo, Antonio Francavilla, Gaetano Gallo, Mario Trompetto, Giulio Ferrero, Sayaka Mizutani, Hirotsugu Shiroma, Satoshi Shiba, Tatsuhiro Shibata, Shinichi Yachida, Takuji Yamada, Jakob Wirbel, Petra Schrotz-King, Cornelia M. Ulrich, Hermann Brenner, Manimozhiyan Arumugam, Peer Bork, Georg Zeller, Francesca Cordero, Emmanuel Dias-Neto, João Carlos Setubal, Adrian Tett, Barbara Pardini, Maria Rescigno, Levi Waldron, Alessio Naccarati, and Nicola Segata. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nature Medicine*, 25(4):667–678, April 2019. ISSN 1546-170X. doi: 10.1038/s41591-019-0405-7.

Emily Vogtmann, Xing Hua, Georg Zeller, Shinichi Sunagawa, Anita Y. Voigt, Rajna Hercog, James J. Goedert, Jianxin Shi, Peer Bork, and Rashmi Sinha. Colorectal Cancer and the Human Gut Microbiome: Reproducibility with Whole-Genome Shotgun Sequencing. *PloS One*, 11(5):e0155362, 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0155362.

Shinichi Yachida, Sayaka Mizutani, Hirotsugu Shiroma, Satoshi Shiba, Takeshi Nakajima, Taku Sakamoto, Hikaru Watanabe, Keigo Masuda, Yuichiro Nishimoto, Masaru Kubo, Fumie Hosoda, Hirofumi Rokutan, Minori Matsumoto, Hiroyuki Takamaru, Masayoshi Yamada, Takahisa Matsuda, Motoki Iwasaki, Taiki Yamaji, Tatsuo Yachida, Tomoyoshi Soga, Ken Kurokawa, Atsushi Toyoda, Yoshitoshi Ogura, Tetsuya Hayashi, Masanori Hatakeyama, Hitoshi Nakagama, Yutaka Saito, Shinji Fukuda, Tatsuhiro Shibata, and Takuji Yamada. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nature Medicine*, 25(6):968–976, June 2019. ISSN 1546-170X. doi: 10.1038/s41591-019-0458-7.

Jun Yu, Qiang Feng, Sunny Hei Wong, Dongya Zhang, Qiao Yi Liang, Youwen Qin, Longqing Tang, Hui Zhao, Jan Stenvang, Yanli Li, Xiaokai Wang, Xiaoqiang Xu, Ning Chen, William Ka Kei Wu, Jumana Al-Aama, Hans Jørgen Nielsen, Pia Kiilerich, Benjamin Anderschou Holbech Jensen, Tung On Yau, Zhou Lan, Huijue Jia, Junhua Li, Liang Xiao, Thomas Yuen Tung Lam, Siew Chien Ng, Alfred Sze-Lok Cheng, Vincent Wai-Sun Wong, Francis Ka Leung Chan, Xun Xu, Huanming Yang, Lise Madsen, Christian Datz, Herbert Tilg, Jian Wang, Nils Brünner, Karsten Kristiansen, Manimozhiyan Arumugam, Joseph Jao-Yiu Sung, and Jun Wang. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut*, 66(1):70–78, January 2017a. ISSN 1468-3288. doi: 10.1136/gutjnl-2015-309800.

Georg Zeller, Julien Tap, Anita Y. Voigt, Shinichi Sunagawa, Jens Roat Kultima, Paul I. Costea, Aurélien Amiot, Jürgen Böhm, Francesco Brunetti, Nina Habermann, Rajna Hercog, Moritz Koch, Alain Luciani, Daniel R. Mende, Martin A. Schneider, Petra Schrotz-King, Christophe Tournigand, Jeanne Tran Van Nhieu, Takuji Yamada, Jürgen Zimmermann, Vladimir Benes, Matthias Kloor, Cornelia M. Ulrich, Magnus von Knebel Doeberitz, Iradj Sobhani, and Peer Bork. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular Systems Biology*, 10:766, November 2014. ISSN 1744-4292. doi: 10.15252/msb.20145645.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. ISSN 0167-8655. doi: 10.1016/j.patrec.2005.10.010. Type: Journal Article.

Andrew Hinton and Peter J. Mucha. A simultaneous feature selection and compositional association test for detecting sparse associations in high-dimensional metagenomic data. Technical report, December 2021. URL `https://www.researchsquare.com/article/rs-703177/v1`. ISSN: 2693-5015 Type: article.

Vera Pawlowsky-Glahn, Juan Jose Egozcue, and Raimon Tolosana-Delgado. Lecture Notes on Compositional Data Analysis. January 2007.

Feng Zhu, Yanmei Ju, Wei Wang, Qi Wang, Ruijin Guo, Qingyan Ma, Qiang Sun, Yajuan Fan, Yuying Xie, Zai Yang, Zhuye Jie, Binbin Zhao, Liang Xiao, Lin Yang, Tao Zhang, Junqin Feng, Liyang Guo, Xiaoyan He, Yunchun Chen, Ce Chen, Chengge Gao, Xun Xu, Huanming Yang, Jian Wang, Yonghui Dang, Lise Madsen, Susanne Brix, Karsten Kristiansen, Huijue Jia, and Xiancang Ma. Metagenome-wide association of gut microbiome features for schizophrenia. *Nature Communications*, 11(1):1612, March 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-15457-9.

Meagan A. Rubel, Arwa Abbas, Louis J. Taylor, Andrew Connell, Ceylan Tanes, Kyle Bittinger, Valantine N. Ndze, Julius Y. Fonsah, Eric Ngwang, André Essiane, Charles Fokunang, Alfred K. Njamnshi, Frederic D. Bushman, and Sarah A. Tishkoff. Lifestyle and the presence of helminths is associated with gut microbiome composition in Cameroonians. *Genome Biology*, 21(1):122, May 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-02020-4.

Suzanne R. Sharpton, Germaine J.M. Yong, Norah A. Terrault, and Susan V. Lynch. Gut Microbial Metabolism and Nonalcoholic Fatty Liver Disease. *Hepatology Communications*, 3(1):29–43, 2019. ISSN 2471-254X. doi: 10.1002/hep4.1284. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/hep4.1284`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hep4.1284.

Alyxandria M. Schubert, Mary A. M. Rogers, Cathrin Ring, Jill Mogle, Joseph P. Petrosino, Vincent B. Young, David M. Aronoff, and Patrick D. Schloss. Microbiome Data Distinguish Patients with Clostridium difficile Infection and Non-C. difficile-Associated Diarrhea from Healthy Controls. *mBio*, 5 (3):e01021–14, May 2014. ISSN 2150-7511. doi: 10.1128/mBio.01021-14. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4010826/`.

TW Wang, K Asman, AS Gentzke, and et al. Tobacco Product Use Among Adults — United States, 2017. *MMWR Morb Mortal Wkly Rep*, 67(45):1225–1232, 2017. doi: http://dx.doi.org/10.15585/mmwr.mm6744a2. Type: Journal Article.

Adam M. Leventhal, Richard Miech, Jessica Barrington-Trimis, Lloyd D. Johnston, Patrick M. O'Malley, and Megan E. Patrick. Flavors of e-Cigarettes Used by Youths in the United States. *Jama*, 2019. ISSN 0098-7484. doi: 10.1001/jama.2019.17968. Type: Journal Article.

T. W. Wang, L. J. Neff, E. Park-Lee, C. Ren, K. A. Cullen, and B. A. King. E-cigarette Use Among Middle and High School Students - United States, 2020. *MMWR Morb Mortal Wkly Rep*, 69(37):1310–1312, 2020. ISSN 0149-2195 (Print) 0149-2195. doi: 10.15585/mmwr.mm6937e1. Type: Journal Article.

K. A. Cullen, A. S. Gentzke, M. D. Sawdey, J. T. Chang, G. M. Anic, T. W. Wang, M. R. Creamer, A. Jamal, B. K. Ambrose, and B. A. King. e-Cigarette Use Among Youth in the United States, 2019. *Jama*, 2019. ISSN 0098-7484. doi: 10.1001/jama.2019.18387. Type: Journal Article.

E. Kiernan, E. S. Click, P. Melstrom, M. E. Evans, M. R. Layer, D. N. Weissman, S. Reagan-Steiner, J. L. Wiltz, S. Hocevar, A. B. Goodman, and E. Twentyman. A Brief Overview of the National Outbreak of e-Cigarette, or Vaping, Product Use-Associated Lung Injury and the Primary Causes. *Chest*, 159(1): 426–431, 2021. ISSN 0012-3692. doi: 10.1016/j.chest.2020.07.068. Type: Journal Article.

K. D. McAlinden, M. S. Eapen, W. Lu, C. Chia, G. Haug, and S. S. Sohal. COVID-19 and vaping: risk for increased susceptibility to SARS-CoV-2 infection? *Eur Respir J*, 56(1), 2020. ISSN 0903-1936 (Print) 0903-1936. doi: 10.1183/13993003.01645-2020. Type: Journal Article.

E. M. Martin, P. W. Clapp, M. E. Rebuli, E. A. Pawlak, E. Glista-Baker, N. L. Benowitz, R. C. Fry, and I. Jaspers. E-cigarette use results in suppression of immune and inflammatory-response genes in nasal epithelial cells similar to cigarette smoke. *Am J Physiol Lung Cell Mol Physiol*, 311(1):L135–44, 2016. ISSN 1040-0605. doi: 10.1152/ajplung.00170.2016. Type: Journal Article.

B. Reidel, G. Radicioni, P. W. Clapp, A. A. Ford, S. Abdelwahab, M. E. Rebuli, P. Haridass, N. E. Alexis, I. Jaspers, and M. Kesimer. E-Cigarette Use Causes a Unique Innate Immune Response in the Lung, Involving Increased Neutrophilic Activation and Altered Mucin Secretion. *Am J Respir Crit Care Med*, 197(4):492–501, 2018. ISSN 1535-4970 (Electronic) 1073-449X (Linking). doi: 10.1164/rccm.201708-1590OC. URL `https://www.ncbi.nlm.nih.gov/pubmed/29053025`. Type: Journal Article.

P. W. Clapp, E. A. Pawlak, J. T. Lackey, J. E. Keating, S. L. Reeber, G. L. Glish, and I. Jaspers. Flavored e-cigarette liquids and cinnamaldehyde impair respiratory innate immune cell function. *Am J Physiol Lung Cell Mol Physiol*, 313(2):L278–L292, 2017. ISSN 1522-1504 (Electronic) 1040-0605 (Linking). doi: 10.1152/ajplung.00452.2016. URL `https://www.ncbi.nlm.nih.gov/pubmed/28495856`. Type: Journal Article.

M. C. Madison, C. T. Landers, B. H. Gu, C. Y. Chang, H. Y. Tung, R. You, M. J. Hong, N. Baghaei, L. Z. Song, P. Porter, N. Putluri, R. Salas, B. E. Gilbert, I. Levental, M. J. Campen, D. B. Corry, and F. Kheradmand. Electronic cigarettes disrupt lung lipid homeostasis and innate immunity independent of nicotine. *J Clin Invest*, 129(10):4290–4304, 2019. ISSN 0021-9738. doi: 10.1172/jci128531. Type: Journal Article.

A. Ghosh, R. D. Coakley, A. J. Ghio, M. S. Muhlebach, Jr. Esther, C. R., N. E. Alexis, and R. Tarran. Chronic E-Cigarette Use Increases Neutrophil Elastase and Matrix Metalloprotease Levels in the Lung. *Am J Respir Crit Care Med*, 2019. ISSN 1073-449x. doi: 10.1164/rccm.201903-0615OC. Type: Journal Article.

T. E. Sussan, S. Gajghate, R. K. Thimmulappa, J. Ma, J. H. Kim, K. Sudini, N. Consolini, S. A. Cormier, S. Lomnicki, F. Hasan, A. Pekosz, and S. Biswal. Exposure to electronic cigarettes impairs pulmonary anti-bacterial and anti-viral defenses in a mouse model. *PLoS One*, 10(2):e0116861, 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0116861. Type: Journal Article.

J. Gerloff, I. K. Sundar, R. Freter, E. R. Sekera, A. E. Friedman, R. Robinson, T. Pagano, and I. Rahman. Inflammatory Response and Barrier Dysfunction by Different e-Cigarette Flavoring Chemicals Identified by Gas Chromatography-Mass Spectrometry in e-Liquids and e-Vapors on Human Lung Epithelial Cells and Fibroblasts. *Appl In Vitro Toxicol*, 3(1):28–40, 2017. ISSN 2332-1512 (Print) 2332-1512 (Linking).

doi: 10.1089/aivt.2016.0030. URL `https://www.ncbi.nlm.nih.gov/pubmed/28337465`. Type: Journal Article.

T. Muthumalage, M. Prinz, K. O. Ansah, J. Gerloff, I. K. Sundar, and I. Rahman. Inflammatory and Oxidative Responses Induced by Exposure to Commonly Used e-Cigarette Flavoring Chemicals and Flavored e-Liquids without Nicotine. *Front Physiol*, 8:1130, 2017. ISSN 1664-042X (Print) 1664-042X (Linking). doi: 10.3389/fphys.2017.01130. URL `https://www.ncbi.nlm.nih.gov/pubmed/29375399`. Type: Journal Article.

R. Z. Behar, W. Luo, K. J. McWhirter, J. F. Pankow, and P. Talbot. Analytical and toxicological evaluation of flavor chemicals in electronic cigarette refill fluids. *Sci Rep*, 8(1):8288, 2018. ISSN 2045-2322 (Electronic) 2045-2322 (Linking). doi: 10.1038/s41598-018-25575-6. URL `https://www.ncbi.nlm.nih.gov/pubmed/29844439`. Type: Journal Article.

P. W. Clapp, K. S. Lavrich, C. A. van Heusden, E. R. Lazarowski, J. L. Carson, and I. Jaspers. Cinnamaldehyde in flavored e-cigarette liquids temporarily suppresses bronchial epithelial cell ciliary motility by dysregulation of mitochondrial function. *Am J Physiol Lung Cell Mol Physiol*, 316(3): L470–l486, 2019. ISSN 1040-0605. doi: 10.1152/ajplung.00304.2018. Type: Journal Article.

E. Hickman, C. A. Herrera, and I. Jaspers. Common E-Cigarette Flavoring Chemicals Impair Neutrophil Phagocytosis and Oxidative Burst. *Chem Res Toxicol*, 32(6):982–985, 2019. ISSN 0893-228x. doi: 10.1021/acs.chemrestox.9b00171. Type: Journal Article.

J. H. Hwang, M. Lyes, K. Sladewski, S. Enany, E. McEachern, D. P. Mathew, S. Das, A. Moshensky, S. Bapat, D. T. Pride, W. M. Ongkeko, and L. E. Crotty Alexander. Electronic cigarette inhalation alters innate immunity and airway cytokines while increasing the virulence of colonizing bacteria. *J Mol Med (Berl)*, 94(6):667–79, 2016. ISSN 0946-2716. doi: 10.1007/s00109-016-1378-3. Type: Journal Article.

L. Miyashita, R. Suri, E. Dearing, I. Mudway, R. E. Dove, D. R. Neill, R. Van Zyl-Smit, A. Kadioglu, and J. Grigg. E-cigarette vapour enhances pneumococcal adherence to airway epithelial cells. *Eur Respir J*, 51(2), 2018. ISSN 0903-1936. doi: 10.1183/13993003.01592-2017. Type: Journal Article.

W. H. Man, W. A. de Steenhuijsen Piters, and D. Bogaert. The microbiota of the respiratory tract: gatekeeper to respiratory health. *Nat Rev Microbiol*, 15(5):259–270, 2017. ISSN 1740-1526. doi: 10.1038/nrmicro.2017.14. Type: Journal Article.

N. D. J. Ubags and B. J. Marsland. Mechanistic insight into the function of the microbiome in lung diseases. *Eur Respir J*, 50(3), 2017. ISSN 0903-1936. doi: 10.1183/13993003.02467-2016. Type: Journal Article.

E. S. Charlson, J. Chen, R. Custers-Allen, K. Bittinger, H. Li, R. Sinha, J. Hwang, F. D. Bushman, and R. G. Collman. Disordered microbial communities in the upper respiratory tract of cigarette smokers. *PLoS One*, 5(12):e15216, 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0015216. Type: Journal Article.

V. R. Ramakrishnan, L. J. Hauser, and D. N. Frank. The sinonasal bacterial microbiome in health and disease. *Curr Opin Otolaryngol Head Neck Surg*, 24(1):20–5, 2016. ISSN 1068-9508. doi: 10.1097/moo.0000000000000221. Type: Journal Article.

Rune Grønseth, Christine Drengenes, Harald G. Wiker, Solveig Tangedal, Yaxin Xue, Gunnar Reksten Husebø, Øistein Svanes, Sverre Lehmann, Marit Aardal, Tuyen Hoang, Tharmini Kalananthan, Einar Marius Hjellestad Martinsen, Elise Orvedal Leiten, Marianne Aanerud, Eli Nordeide, Ingvild Haaland, Inge Jonassen, Per Bakke, and Tomas Eagan. Protected sampling is preferable in bronchoscopic studies of the airway microbiome. *ERJ Open Research*, 3(3):00019–2017, 2017. doi:

10.1183/23120541.00019-2017. URL `http://openres.ersjournals.com/content/3/3/00019-2017.abstract`. Type: Journal Article.

Arianna Di Stadio, Claudio Costantini, Giorgia Renga, Marilena Pariano, Giampietro Ricci, and Luigina Romani. The Microbiota/Host Immune System Interaction in the Nose to Protect from COVID-19. *Life (Basel, Switzerland)*, 10(12):345, 2020. ISSN 2075-1729. doi: 10.3390/life10120345. URL `https://pubmed.ncbi.nlm.nih.gov/33322584https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7763594/`. Type: Journal Article.

C. Rosas-Salazar, K. S. Kimura, M. H. Shilts, B. A. Strickland, M. H. Freeman, B. C. Wessinger, V. Gupta, H. M. Brown, S. V. Rajagopala, J. H. Turner, and S. R. Das. SARS-CoV-2 Infection and Viral Load are Associated with the Upper Respiratory Tract Microbiome. *J Allergy Clin Immunol*, 2021. ISSN 0091-6749 (Print) 0091-6749. doi: 10.1016/j.jaci.2021.02.001. Type: Journal Article.

Jack A. Gilbert, Martin J. Blaser, J. Gregory Caporaso, Janet K. Jansson, Susan V. Lynch, and Rob Knight. Current understanding of the human microbiome. *Nature Medicine*, 24(4):392–400, 2018. ISSN 1546-170X. doi: 10.1038/nm.4517. URL `https://doi.org/10.1038/nm.4517`. Type: Journal Article.

Hongzhe Li. Statistical and Computational Methods in Microbiome and Metagenomics. In *Handbook of Statistical Genomics*, pages 977–550. 2019. doi: https://doi.org/10.1002/9781119487845.ch35. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119487845.ch35`. Type: Book Section.

Duo Jiang, Courtney R. Armour, Chenxiao Hu, Meng Mei, Chuan Tian, Thomas J. Sharpton, and Yuan Jiang. Microbiome Multi-Omics Network Analysis: Statistical Considerations, Limitations, and Opportunities. *Frontiers in Genetics*, 10(995), 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.00995. URL `https://www.frontiersin.org/article/10.3389/fgene.2019.00995`. Type: Journal Article.

M. E. Rebuli, A. M. Speen, P. W. Clapp, and I. Jaspers. Novel applications for a noninvasive sampling method of the nasal mucosa. *Am J Physiol Lung Cell Mol Physiol*, 312(2):L288–l296, 2017. ISSN 1040-0605. doi: 10.1152/ajplung.00476.2016. Type: Journal Article.

M. S. Muhlebach, B. T. Zorn, C. R. Esther, J. E. Hatch, C. P. Murray, L. Turkovic, S. C. Ranganathan, R. C. Boucher, S. M. Stick, and M. C. Wolfgang. Initial acquisition and succession of the cystic fibrosis lung microbiome is associated with disease progression in infants and preschool children. *PLoS Pathog*, 14(1): e1006798, 2018. ISSN 1553-7366 (Print) 1553-7366. doi: 10.1371/journal.ppat.1006798. Type: Journal Article.

K. M. Horvath, M. Herbst, H. Zhou, H. Zhang, T. L. Noah, and I. Jaspers. Nasal lavage natural killer cell function is suppressed in smokers after live attenuated influenza virus. *Respir Res*, 12:102, 2011. ISSN 1465-9921. doi: 10.1186/1465-9921-12-102. Type: Journal Article.

N. M. Davis, D. M. Proctor, S. P. Holmes, D. A. Relman, and B. J. Callahan. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome*, 6(1):226, 2018. ISSN 2049-2618. doi: 10.1186/s40168-018-0605-2. Type: Journal Article.

C. Drengenes, H. G. Wiker, T. Kalananthan, E. Nordeide, T. M. L. Eagan, and R. Nielsen. Laboratory contamination in airway microbiome studies. *BMC Microbiol*, 19(1):187, 2019. ISSN 1471-2180. doi: 10.1186/s12866-019-1560-1. Type: Journal Article.

Matthew C. B. Tsilimigras and Anthony A. Fodor. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Annals of Epidemiology*, 26(5):330–335, 2016. ISSN 1047-2797. doi: 10.1016/j.annepidem.2016.03.002. Type: Journal Article.

Pierre Legendre and Louis Legendre. Canonical analysis. In *Developments in Environmental Modelling*, volume 24, pages 625–710. Elsevier, 2012. ISBN 978-0-444-53868-0. Type: Book Section.

Jari Oksanen, F. Guillaume Blanchet, Roeland Kindt, P. Legendre, Peter Minchin, Bob O'Hara, Gavin Simpson, Peter Solymos, Hank Stevens, and Helene Wagner. Vegan: Community Ecology Package. *R Package Version 2.2-1*, 2:1–2, 2015. Type: Journal Article.

Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. *Gephi: An Open Source Software for Exploring and Manipulating Networks*. 2009. Type: Book.

Gabor Csardi and Tamas Nepusz. The Igraph Software Package for Complex Network Research. *InterJournal*, Complex Systems:1695, 2005. Type: Journal Article.

Matthew Barker and William Rayens. Partial least squares for discrimination. *Journal of Chemometrics*, 17 (3):166–173, 2003. ISSN 1099-128X. doi: https://doi.org/10.1002/cem.785. Type: Journal Article.

Richard G. Brereton and Gavin R. Lloyd. Partial least squares discriminant analysis: taking the magic away. *Journal of Chemometrics*, 28(4):213–225, 2014. ISSN 1099-128X. doi: https://doi.org/10.1002/cem.2609. Type: Journal Article.

Max Kuhn. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28(5), 2008. ISSN 1548-7660. doi: 10.18637/jss.v028.i05. Type: Journal Article.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Use R! Springer International Publishing, 2 edition, 2016. ISBN 978-3-319-24275-0. Type: Book.

David J. Hand and Robert J. Till. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, 45(2):171–186, 2001. ISSN 1573-0565. doi: 10.1023/A:1010920819831. Type: Journal Article.

M. Stone. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, 1974. ISSN 2517-6161. doi: https://doi.org/10.1111/j.2517-6161.1974.tb00994.x. Type: Journal Article.

Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12(1):77, 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-77. Type: Journal Article.

Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1): 289–300, 1995. ISSN 0035-9246. Type: Journal Article.

C. Kumpitsch, K. Koskinen, V. Schöpf, and C. Moissl-Eichinger. The microbiome of the upper respiratory tract in health and disease. *BMC Biol*, 17(1):87, 2019. ISSN 1741-7007. doi: 10.1186/s12915-019-0703-z. Type: Journal Article.

Ilke De Boeck, Stijn Wittouck, Sander Wuyts, Eline F. M. Oerlemans, Marianne F. L. van den Broek, Dieter Vandenheuvel, Olivier Vanderveken, and Sarah Lebeer. Comparing the Healthy Nose and Nasopharynx Microbiota Reveals Continuity As Well As Niche-Specificity. *Frontiers in microbiology*, 8:2372–2372, 2017. ISSN 1664-302X. doi: 10.3389/fmicb.2017.02372. URL https://pubmed.ncbi.nlm.nih.gov/29238339https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5712567/. Type: Journal Article.

P. J. McMurdie and S. Holmes. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One*, 8(4):e61217, 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0061217. Type: Journal Article.

M. E. Rebuli, A. M. Speen, E. M. Martin, K. A. Addo, E. A. Pawlak, E. Glista-Baker, C. Robinette, H. Zhou, T. L. Noah, and I. Jaspers. Wood Smoke Exposure Alters Human Inflammatory Responses to Viral Infection in a Sex-Specific Manner: A Randomized, Placebo-Controlled Study. *Am J Respir Crit Care Med*, 2018. ISSN 1073-449x. doi: 10.1164/rccm.201807-1287OC. Type: Journal Article.

Y. Cho, G. Abu-Ali, H. Tashiro, T. A. Brown, R. S. Osgood, D. I. Kasahara, C. Huttenhower, and S. A. Shore. Sex Differences in Pulmonary Responses to Ozone in Mice. Role of the Microbiome. *Am J Respir Cell Mol Biol*, 60(2):198–208, 2019. ISSN 1044-1549 (Print) 1044-1549. doi: 10.1165/rcmb.2018-0099OC. Type: Journal Article.

G. Yu, S. Phillips, M. H. Gail, J. J. Goedert, M. S. Humphrys, J. Ravel, Y. Ren, and N. E. Caporaso. The effect of cigarette smoking on the oral and nasal microbiota. *Microbiome*, 5(1):3, 2017b. ISSN 2049-2618. doi: 10.1186/s40168-016-0226-6. Type: Journal Article.

Pipat Piewngam, Yue Zheng, Thuan H. Nguyen, Seth W. Dickey, Hwang-Soo Joo, Amer E. Villaruz, Kyle A. Glose, Emilie L. Fisher, Rachelle L. Hunt, Barry Li, Janice Chiou, Sujiraphong Pharkjaksu, Sunisa Khongthong, Gordon Y. C. Cheung, Pattarachai Kiratisin, and Michael Otto. Pathogen elimination by probiotic Bacillus via signalling interference. *Nature*, 562(7728):532–537, 2018. ISSN 1476-4687. doi: 10.1038/s41586-018-0616-y. URL https://doi.org/10.1038/s41586-018-0616-y. Type: Journal Article.

Cindy M. Liu, Lance B. Price, Bruce A. Hungate, Alison G. Abraham, Lisbeth A. Larsen, Kaare Christensen, Marc Stegger, Robert Skov, and Paal Skytt Andersen. Staphylococcus aureus and the ecology of the nasal microbiome. *Science Advances*, 1(5):e1400216, 2015. doi: 10.1126/sciadv.1400216. URL https://advances.sciencemag.org/content/advances/1/5/e1400216.full.pdf. Type: Journal Article.

Adèle Sakr, Fabienne Brégeon, Jean-Louis Mège, Jean-Marc Rolain, and Olivier Blin. Staphylococcus aureus Nasal Colonization: An Update on Mechanisms, Epidemiology, Risk Factors, and Subsequent Infections. *Frontiers in microbiology*, 9:2419–2419, 2018. ISSN 1664-302X. doi: 10.3389/fmicb.2018.02419. URL https://pubmed.ncbi.nlm.nih.gov/30349525https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6186810/. Type: Journal Article.

J. K. Actor, S. A. Hwang, and M. L. Kruzel. Lactoferrin as a natural immune modulator. *Curr Pharm Des*, 15(17):1956–73, 2009. ISSN 1381-6128 (Print) 1381-6128. doi: 10.2174/138161209788453202. Type: Journal Article.

G. J. Casimir, N. Lefevre, F. Corazza, and J. Duchateau. Sex and inflammation in respiratory diseases: a clinical viewpoint. *Biol Sex Differ*, 4:16, 2013. ISSN 2042-6410 (Print) 2042-6410. doi: 10.1186/2042-6410-4-16. Type: Journal Article.

W. H. Man, M. A. van Houten, M. E. Mérelle, A. M. Vlieger, Mljn Chu, N. J. G. Jansen, E. A. M. Sanders, and D. Bogaert. Bacterial and viral respiratory tract microbiota and host characteristics in children with lower respiratory tract infections: a matched case-control study. *Lancet Respir Med*, 7(5):417–426, 2019. ISSN 2213-2600 (Print) 2213-2600. doi: 10.1016/s2213-2600(18)30449-1. Type: Journal Article.

Ariane Neumann, Lars Björck, and Inga-Maria Frick. Finegoldia magna, an Anaerobic Gram-Positive Bacterium of the Normal Human Microbiota, Induces Inflammation by Activating Neutrophils. *Frontiers in Microbiology*, 11(65), 2020. ISSN 1664-302X. doi: 10.3389/fmicb.2020.00065. URL `https://www.frontiersin.org/article/10.3389/fmicb.2020.00065`. Type: Journal Article.

Michael Hoggard, Brett Wagner Mackenzie, Ravi Jain, Michael W. Taylor, Kristi Biswas, and Richard G. Douglas. Chronic Rhinosinusitis and the Evolving Understanding of Microbial Ecology in Chronic Inflammatory Mucosal Disease. *Clinical Microbiology Reviews*, 30(1):321, 2017. doi: 10.1128/CMR.00060-16. URL `http://cmr.asm.org/content/30/1/321.abstract`. Type: Journal Article.

Chih-Yung Chiu, Yi-Ling Chan, Ming-Han Tsai, Chia-Jung Wang, Meng-Han Chiang, Chun-Che Chiu, and Shih-Chi Su. Cross-talk between airway and gut microbiome links to IgE responses to house dust mites in childhood airway allergies. *Scientific Reports*, 10(1):13449, 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-70528-7. URL `https://doi.org/10.1038/s41598-020-70528-7`. Type: Journal Article.

G. Bacci, P. Paganin, L. Lopez, C. Vanni, C. Dalmastri, C. Cantale, L. Daddiego, G. Perrotta, D. Dolce, P. Morelli, V. Tuccio, A. De Alessandri, E. V. Fiscarelli, G. Taccetti, V. Lucidi, A. Bevivino, and A. Mengoni. Pyrosequencing Unveils Cystic Fibrosis Lung Microbiome Differences Associated with a Severe Lung Function Decline. *PLoS One*, 11(6):e0156807, 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0156807. Type: Journal Article.

Ulrich M. Zanger and Matthias Schwab. Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacology & Therapeutics*, 138 (1):103–141, 2013. ISSN 0163-7258. doi: https://doi.org/10.1016/j.pharmthera.2012.12.007. URL `https://www.sciencedirect.com/science/article/pii/S0163725813000065`. Type: Journal Article.

R. Vemuri, K. E. Sylvia, S. L. Klein, S. C. Forster, M. Plebanski, R. Eri, and K. L. Flanagan. The microgenderome revealed: sex differences in bidirectional interactions between the microbiota, hormones, immunity and disease susceptibility. *Semin Immunopathol*, 41(2):265–275, 2019. ISSN 1863-2297 (Print) 1863-2297. doi: 10.1007/s00281-018-0716-7. Type: Journal Article.

Y. S. Kim, T. Unno, B. Y. Kim, and M. S. Park. Sex Differences in Gut Microbiota. *World J Mens Health*, 38(1):48–60, 2020. ISSN 2287-4208 (Print) 2287-4208. doi: 10.5534/wjmh.190009. Type: Journal Article.

J. H. Shin, Y. H. Park, M. Sim, S. A. Kim, H. Joung, and D. M. Shin. Serum level of sex steroid hormone is associated with diversity and profiles of human gut microbiome. *Res Microbiol*, 170(4-5):192–201, 2019. ISSN 0923-2508. doi: 10.1016/j.resmic.2019.03.003. Type: Journal Article.

M. K. Han, E. Arteaga-Solis, J. Blenis, G. Bourjeily, D. J. Clegg, D. DeMeo, J. Duffy, B. Gaston, N. M. Heller, A. Hemnes, E. P. Henske, R. Jain, T. Lahm, L. H. Lancaster, J. Lee, M. J. Legato, S. McKee, R. Mehra, A. Morris, Y. S. Prakash, M. R. Stampfli, R. Gopal-Srivastava, A. D. Laposky, A. Punturieri,

L. Reineck, X. Tigno, and J. Clayton. Female Sex and Gender in Lung/Sleep Health and Disease. Increased Understanding of Basic Biological, Pathophysiological, and Behavioral Mechanisms Leading to Better Health for Female Patients with Lung Disease. *Am J Respir Crit Care Med*, 198(7):850–858, 2018. ISSN 1073-449X (Print) 1073-449x. doi: 10.1164/rccm.201801-0168WS. Type: Journal Article.

Ilke De Boeck, Stijn Wittouck, Katleen Martens, Jos Claes, Mark Jorissen, Brecht Steelant, Marianne F. L. van den Broek, Sven F. Seys, Peter W. Hellings, Olivier M. Vanderveken, and Sarah Lebeer. Anterior Nares Diversity and Pathobionts Represent Sinus Microbiome in Chronic Rhinosinusitis. *mSphere*, 4(6): e00532–19, 2019. doi: 10.1128/mSphere.00532-19. URL `http://msphere.asm.org/content/4/6/e00532-19.abstract`. Type: Journal Article.

Y. W. Chen, S. W. Li, C. D. Lin, M. Z. Huang, H. J. Lin, C. Y. Chin, Y. R. Lai, C. H. Chiu, C. Y. Yang, and C. H. Lai. Fine Particulate Matter Exposure Alters Pulmonary Microbiota Composition and Aggravates Pneumococcus-Induced Lung Pathogenesis. *Front Cell Dev Biol*, 8:570484, 2020. ISSN 2296-634X (Print) 2296-634x. doi: 10.3389/fcell.2020.570484. Type: Journal Article.

J. Mariani, C. Favero, A. Spinazzè, D. M. Cavallo, M. Carugno, V. Motta, M. Bonzini, A. Cattaneo, A. C. Pesatori, and V. Bollati. Short-term particulate matter exposure influences nasal microbiota in a population of healthy subjects. *Environ Res*, 162:119–126, 2018. ISSN 0013-9351. doi: 10.1016/j.envres.2017.12.016. Type: Journal Article.