UNIVERSITY OF CALIFORNIA SAN DIEGO

Computational and Statistical Methods for Extracting Biological Signal from High-Dimensional Microbiome Data

A dissertation submitted in partial satisfaction

of the requirements for the Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Gibraan Rahman

Committee in charge:

Professor Rob Knight, Chair
Professor Pieter Dorrestein, Co-Chair
Professor Kit Curtius
Professor Nathan Lewis
Professor Siavash Mirarab

2023

The Dissertation of Gibraan Rahman is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

I would like to dedicate this work to my family, without whom none of this would be possible.

EPIGRAPH

“All models are wrong, some are useful.”

— George Box (1919-2013)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Firstly, I would like to acknowledge my advisor, Rob Knight, who has been an ardent supporter of me and my work. I first came to Rob with essentially no knowledge of microbiome science, programming, or statistics, and yet he enthusiastically welcomed me into his lab to learn from the best. Throughout my time in the Knight Lab, Rob has continually provided me with guidance and mentorship navigating the complex worlds of academia, science, and life.

And, of course, I would like to acknowledge the members of the Knight Lab network, whose advice, guidance, and support was indispensable in my Ph.D journey. I cannot express enough how grateful I am to have been able to spend the last several years with such a talented and hardworking group of researchers.

Next, I would like to thank the entire Bioinformatics and Systems Biology program. I came to UCSD feeling out of my depth and the students provided an immediate and persistent support system that I have been able to fall back on throughout my time here. I have met some of my closest friends here and cannot imagine a better place to have pursued my degree.

Finally, I would like to thank my family. Throughout my life they have fully supported me in my various endeavors. Their ceaseless encouragement has kept me going during times when I needed it the most.

Chapter 1, in full, is a reprint of the material as it appears in "Applications and comparison of dimensionality reduction methods for microbiome data." George Armstrong, Gibraan Rahman, Cameron Martino, Daniel McDonald, Antonio Gonzalez, Gal Mishne, and Rob Knight. *Frontiers in Bioinformatics*. The dissertation author is the co-first author of this paper in conjunction with Dr. George Armstrong.

Chapter 2 has been submitted for publication of the materials as it may appear in *Genes*, "Determination of effect sizes for power analysis for microbiome studies using large microbiome databases." Gibraan Rahman, Daniel McDonald, Antonio Gonzalez, Yoshiki Vázquez-Baeza,

Lingjing Jiang, Climent Casals-Pascual, Shyamal Peddada, Daniel Hakim, Amanda Hazel Dilmore, Brent Nowinski, and Rob Knight. The dissertation author was the primary investigator and first author of this paper.

Chapter 3 has been submitted for publication of the materials as it may appear in *Cell Reports*, "Paired microbiome and metabolome analyses associate bile acid changes with colorectal cancer progression" Ting Fu,, Tao Huan, Gibraan Rahman, Hui Zhi, Zhenjiang Xu, Tae Gyu Oh, Jian Guo, Sally Coulter, Anupriya Tripathi, Cameron Martino, Justin L McCarville, Qiyun Zhu, Fritz Cayabyab, Mingxiao He, Shipei Xing, Ruth T. Yu, Annette Atkins, Christopher Liddle, Janelle Ayres, Manuela Raffatellu, Pieter C. Dorrestein, Michael Downes, Rob Knight3, and Ronald M. Evans. The dissertation author is the co-first author of this paper in conjuction with Dr. Ting Fu and Dr. Tao Huan.

Chapter 4 has been submitted for publication of the material as it may appear in *Nature Microbiology*, "BIRDMAn: A Bayesian differential abundance framework that enables robust inference of host-microbe associations" Gibraan Rahman, James T. Morton, Cameron Martino, Gregory D. Sepich-Poore, Celeste Allaband, Caitlin Guccione, Yang Chen, Daniel Hakim, Mehrbod Estaki, and Rob Knight. The dissertation author was the primary investigator and first author of this paper.

VITA

2018 B.S. in Biomedical Engineering with an emphasis in Computational Biomedical Engineering, University of Texas at Austin

2023 Ph. D. in Bioinformatics and Systems Biology, University of California San Diego

PUBLICATIONS

(*) Denotes co-first authorship

BIRDMAn: A Bayesian differential abundance framework that enables robust inference of host-microbe associations

bioRxiv 2023

doi: https://doi.org/10.1101/2023.01.30.526328

Gibraan Rahman, James T Morton, Cameron Martino, Gregory D Sepich-Poore, Celeste Allaband, Caitlin Guccione, Yang Chen, Daniel Hakim, Mehrbod Estaki, Rob Knight

Insulin-regulated serine and lipid metabolism drive peripheral neuropathy

Nature 2023

doi: https://doi.org/10.1038/s41586-022-05637-6

Michal K Handzlik, Jivani M Gengatharan, Katie E Frizzi, Grace H McGregor, Cameron Martino, **Gibraan Rahman**, Antonio Gonzalez, Ana M Moreno, Courtney R Green, Lucie S Guernsey, Terry Lin, Patrick Tseng, Yoichiro Ideguchi, Regis J Fallon, Amandine Chaix, Satchidananda Panda, Prashant Mali, Martina Wallace, Rob Knight, Marin L Gantner, Nigel A Calcutt, Christian M Metallo

Paired microbiome and metabolome analyses associate bile acid changes with colorectal cancer progression

(Submitted) 2022

Ting Fu*, Tao Huan*, **Gibraan Rahman***, Hui Zhi, Zhenjiang Xu, Tae Gyu Oh, Jian Guo, Sally Coulter, Anupriya Tripathi, Cameron Martino, Justin L McCarville, Qiyun Zhu, Fritz Cayabyab, Mingxiao He, Shipei Xing, Fernando Vargas, Ruth T. Yu, Annette R Atkins, Christopher Liddle, Janelle Ayres, Manuela Raffatellu, Pieter C. Dorrestein, Michael Downes, Rob Knight, Ronald M. Evans

Effects of a Ketogenic and Low Fat Diet on the Human Metabolome, Microbiome and Food-ome in Adults at Risk for Alzheimer's Disease

medRxiv 2022

doi: https://doi.org/10.1101/2022.08.30.22279087

Amanda Hazel Dilmore*, Cameron Martino*, Bryan J. Neth, Kiana A West, Jasmine Zemlin, **Gibraan Rahman**, Morgan Panitchpakdi, Michael J Meehan, Kelly C Weldon, Colette Blach, Leyla Schimmel, Rima Kaddurah-Daouk, Pieter Dorrestein, Rob Knight, Suzanne Craft, Alzheimer's Gut Microbiome Project Consortium

Scalable power analysis and effect size exploration of microbiome community differences with Evident

bioRxiv 2022

doi: https://doi.org/10.1101/2022.05.19.492684

Gibraan Rahman, Daniel McDonald, Antonio Gonzalez, Yoshiki Vázquez-Baeza, Lingjing Jiang, Climent Casals-Pascual, Shyamal Peddada, Daniel Hakim, Amanda Hazel Dilmore, Brent Nowinski, Rob Knight

Compositionally aware phylogenetic beta-diversity measures better resolve microbiomes associated with phenotype

mSystems 2022

doi: https://doi.org/10.1128/msystems.00050-22

Cameron Martino, Daniel McDonald, Kalen Cantrell, Amanda Hazel Dilmore, Yoshiki Vázquez-Baeza, Liat Shenhav, Justin P. Shaffer, **Gibraan Rahman**, George Armstrong, Celeste Allaband, Se Jin Song, Rob Knight

Multi-omic analysis along the gut-brain axis points to a functional architecture of autism

bioRxiv 2022

doi: https://doi.org/10.1101/2022.02.25.482050

James T. Morton, Dong-Min Jin, Robert H. Mills, Yan Shao, **Gibraan Rahman**, Kirsten Berding, Brittany D. Needham, María Fernanda Zurita, Maude David, Olga V. Averina, Alexey S. Kovtun, Antonio Noto, Michele Mussap, Mingbang Wang, Daniel N. Frank, Ellen Li, Wenhao Zhou, Vassilios Fanos, Valery N. Danilenko, Dennis P. Wall, Paúl Cárdenas, ManuelE. Baldeón, Ramnik J. Xavier, Sarkis K. Mazmanian, Rob Knight, Jack A. Gilbert, Sharon M. Donovan, Trevor D. Lawley, Bob Carpenter, Richard Bonneau, Gaspar Taroncher-Oldenburg

Bacterial metatranscriptomes in wastewater can differentiate virally infected human populations

bioRxiv 2022

doi: https://doi.org/10.1101/2022.02.23.481658

Rodolfo A Salido, Cameron Martino, Smruthi Karthikeyan, Shi Huang, **Gibraan Rahman**, Antonio Gonzalez, Livia S. Zaramela, Kristen L Beck, Shrikant Bhute, Kalen Cantrell, Anna Paola Carrieri, Sawyer Farmer, Niina Haiminen, Greg Humphrey, Ho-Cheol Kim, Laxmi Parida, Alex Richter, Yoshiki Vázquez-Baeza, Karsten Zengler, Austin D. Swafford, Andrew Bartko, Rob Knight

Applications and comparison of dimensionality reduction methods for microbiome data

Frontiers in Bioinformatics 2022

doi: https://doi.org/10.3389/fbinf.2022.821861

George Armstrong*, **Gibraan Rahman***, Cameron Martino, Daniel McDonald, Antonio Gonzalez, Gal Mishne, Rob Knight

SARS-CoV-2 Distribution in Residential Housing Suggests Contact Deposition and Correlates with *Rothia* sp.

medRxiv 2021

doi: https://dx.doi.org/10.1101%2F2021.12.06.21267101

Victor Cantú*, Rodolfo Salido*, Shi Huang, **Gibraan Rahman**, Rebecca Tsai, Holly Valentine, Celestine Magallanes, Stefan Aigner, Nathan Baer, Tom Barber, Pedro Belda-Ferre, Maryan Betty, MacKenzie Bryant, Martín Casas Maya, Anelizze Castro-Martínez, Marisol Chacón, Willi Cheung, Evelyn Crescini, Peter De Hoff, Emily Eisner, Sawyer Farmer, Abbas Hakim, Laura Kohn, Alma Lastrella, Elijah Lawrence, Sydney Morgan, Toan Ngo, Alhakam Nouri, R. Ostrander, Ashley Plascencia, Christopher Ruiz, Shashank Sathe, Phoebe Seaver, Tara

Schwartz, Elizabeth Smoot, Thomas Valles, Gene Yeo, Louise Laurent, Rebecca Fielding-Miller, Rob Knight

Uniform Manifold Approximation and Projection (UMAP) Reveals Composite Patterns and Resolves Visualization Artifacts in Microbiome Data

mSystems 2021

doi: https://doi.org/10.1128/mSystems.00691-21

George Armstrong, Cameron Martino, **Gibraan Rahman**, Antonio Gonzalez, Yoshiki Vázquez-Baeza, Gal Mishne, Rob Knight

Reduced gut microbiome diversity in people with HIV who have distal neuropathic pain

The Journal of Pain 2021

doi: https://doi.org/10.1016/j.jpain.2021.08.006

Ronald J.Ellis, Robert K.Heaton, Sara Gianella, **Gibraan Rahman**, Rob Knight

SARS-CoV-2 detection status associates with bacterial community composition in patients and the hospital environment

Microbiome 2021

doi: https://doi.org/10.1186/s40168-021-01083-0

Clarisse Marotz, Pedro Belda-Ferre, Farhana Ali, Promi Das, Shi Huang, Kalen Cantrell, Lingjing Jiang, Cameron Martino, Rachel E. Diner, **Gibraan Rahman**, Daniel McDonald, George Armstrong, Sho Kodera, Sonya Donato, Gertrude Ecklu-Mensah, Neil Gottel, Mariana C. Salas Garcia, Leslie Y. Chiang, Rodolfo A. Salido, Justin P. Shaffer, Mac Kenzie Bryant, Karenina Sanders, Greg Humphrey, Gail Ackermann, Niina Haiminen, Kristen L. Beck, Ho-Cheol Kim, Anna Paola Carrieri, Laxmi Parida, Yoshiki Vázquez-Baeza, Francesca J. Torriani, Rob Knight, Jack Gilbert, Daniel A. Sweeney & Sarah M. Allard

Non-alcoholic Steatohepatitis and HCC in a Hyperphagic Mouse Accelerated by Western Diet

Cellular and Molecular Gastroenterology and Hepatology 2021

doi: https://doi.org/10.1016/j.jcmgh.2021.05.010

Souradipta Ganguly, German Aleman Muench, Linshan Shang, Sara Brin Rosenthal, **Gibraan Rahman**, Ruoyu Wang, Yanhan Wang, Hyeok Choon Kwon, Anthony M. Diomino, Tatiana Kisseleva, Pejman Soorosh, Mojgan Hosseini, Rob Knight, Bernd Schnabl, David A. Brenner, Debanjan Dhar

EMPress enables tree-guided, interactive, and exploratory analyses of multi-omic datasets

mSystems 2021

doi: https://doi.org/10.1128/mSystems.01216-20

Kalen Cantrell*, Marcus W. Fedarko*, **Gibraan Rahman**, Daniel McDonald, Yimeng Yang, Thant Zaw, Antonio Gonzalez, Stefan Janssen, Mehrbod Estaki, Niina Haiminen, Kristen L. Beck, Qiyun Zhu, Erfan Sayyari, Jamie Morton, George Armstrong, Anupriya Tripathi, Julia M. Gauglitz, Clarisse Marotz, Nathaniel L. Matteson, Cameron Martino, Jon G. Sanders, Anna Paola Carrieri, Se Jin Song, Austin D. Swafford, Pieter Dorrestein, Kristian G. Andersen, Laxmi Parida, Ho-Cheol Kim, Yoshiki Vázquez-Baeza, Rob Knight

Visualizing 'omic feature rankings and log-ratios using Qurro

NAR Genomics and Bioinformatics 2020

doi: https://doi.org/10.1093/nargab/lqaa023

Marcus W. Fedarko, Cameron Martino, James T. Morton, Antonio González, **Gibraan Rahman**, Clarisse A. Marotz, Jeremiah J. Minich, Eric E. Allen, Rob Knight

ABSTRACT OF THE DISSERTATION

Computational and Statistical Methods for Extracting Biological Signal from High-Dimensional Microbiome Data

by

Gibraan Rahman

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2023

Professor Rob Knight, Chair

Professor Pieter Dorrestein, Co-Chair

Next-generation sequencing (NGS) has effected an explosion of research into the relationship between genetic information and a variety of biological conditions. One of the most exciting areas of study is how the trillions of microbial species that we share this Earth with affect our health. However, the process of extracting useful biological insights from this breadth of data is far from trivial. There are numerous statistical and computational considerations in addition to the already complex and messy biological problems. In this thesis, I describe my work on developing and implementing software to tackle the complex world of statistical microbiome analysis.

In the first part of this thesis, we review the applications and challenges of performing dimensionality reduction on microbiome data comprising thousands of microbial taxa. When dealing with this high dimensionality, it is imperative to be able to get an overview of the community structure in a lower dimensional space that can be both visualized and interpreted. We review the statistical considerations for dimensionality reduction and the existing tools and algorithms that can and cannot address them. This includes discussions about sparsity, compositionality, and phylogenetic signal. We also make recommendations about tools and algorithms to consider for different use-cases.

In the second part of this thesis, we present a new software, Evident, designed to assist researchers with statistical analysis of microbiome effect sizes and power analysis. Effect sizes of statistical tests are not widely reported in microbiome datasets, limiting the interpretability of community differences such as alpha and beta diversity. As more large microbiome studies are produced, researchers have the opportunity to mine existing datasets to get a sense of the effect size for different biological conditions. These, in turn, can be used to perform power analysis prior to designing an experiment, allowing researchers to better allocate resources. We show how Evident is scalable to dozens of datasets and provides easy calculation and exploration of effect sizes and power analysis from existing data.

In the third part of this thesis, we describe a novel investigation into the joint microbiome and metabolome axis in colorectal cancer. In most cases of sporadic colorectal cancers (CRC), tumorigenesis is a multistep process driven by genomic alterations in concert with dietary influences. In addition, mounting evidence has implicated the gut microbiome as an effector in the development and progression of CRC. While large meta-analyses have provided mechanistic insight into disease progression in CRC patients, study heterogeneity has limited causal associations. To address this limitation, multi-omics studies on genetically controlled cohorts of mice were performed to distinguish genetic and dietary influences. Diet was identified as the major driver of microbial and metabolomic differences, with reductions in alpha diversity and widespread changes in cecal metabolites seen in HFD-fed mice. Similarly, the levels of non-classic amino acid conjugated forms of the bile acid cholic acid (AA-CAs) increased with HFD. We show that these AA-CAs signal through the nuclear receptor FXR and membrane receptor TGR5 to functionally impact intestinal stem cell growth. In addition, the poor intestinal permeability of these AA-CAs supports their localization in the gut. Moreover, two cryptic microbial strains, *Ileibacterium valens* and *Ruminococcus gnavus*, were shown to have the capacity to synthesize these AA-CAs. This multi-omics dataset from CRC mouse models supports diet-induced shifts in the microbiome and metabolome in disease progression with potential utility in directing future diagnostic and therapeutic developments.

In the fourth chapter, we demonstrate a new framework for performing differential abundance analysis using customized statistical modeling. As we learn more and more about the relationship between the microbiome and biological conditions, experimental protocols are becoming more and more complex. For example, meta-analyses, interventions, longitudinal studies, etc. are being used to better understand the dynamic nature of the microbiome. However, statistical methods to analyze these relationships are lacking – especially in the field of differential abundance. Finding biomarkers associated with conditions of interest must be performed with

statistical care when dealing with these kinds of experimental designs. We present BIRDMAn, a software package integrating probabilistic programming with Stan to build custom models for analyzing microbiome data. We show that, on both simulated and real datasets, BIRDMAn is able to extract novel biological signals that are missed by existing methods.

These chapters, taken together, advance our knowledge of statistical analysis of microbiome data and provide tools and references for researchers looking to perform analysis on their own data.

Chapter 1. Applications and comparison of dimensionality reduction methods for microbiome data

Abstract

Dimensionality reduction techniques are a key component of most microbiome studies, providing both the ability to tractably visualize complex microbiome datasets and the starting point for additional, more formal, statistical analyses. In this review, we discuss the motivation for applying dimensionality reduction techniques, the special characteristics of microbiome data such as sparsity and compositionality that make this difficult, the different categories of strategies that are available for dimensionality reduction, and examples from the literature of how they have been successfully applied (together with pitfalls to avoid). We conclude by describing the need for further development in the field, in particular combining the power of phylogenetic analysis with the ability to handle sparsity, compositionality, and non-normality, as well as discussing current techniques that should be applied more widely in future analyses.

1.1. Introduction: what is dimensionality reduction and why do we do it?

To a first approximation, life on Earth consists of complex microbial communities, with "familiar" multicellular organisms such as plants and animals being rounding errors in terms of cell count and biomass. The genetic repertoire of such a community is called a "microbiome" (Turnbaugh et al., 2007), although the term "microbiome" is often also loosely applied to the collection of microbes that make up the community. In either sense, microbiomes are typically incredibly complex, containing vast numbers of species and genes, and how samples relate, even in well-studied contexts, are not predetermined. For example, in the Earth Microbiome Project (EMP) (Thompson et al., 2017) and the work leading up to it (Lozupone and Knight, 2007; Ley et

al., 2008; Caporaso et al., 2011), an ontology constructed from the microbe's perspective based on community similarities and differences revealed many surprises, such as a deep separation between free-living and host-associated samples, and between saline and non-saline samples. Accordingly, to truly understand the microbial perspective, we must get acquainted with the structure of the data in human-interpretable formats. This is especially important when we need to separate new biological discoveries from technical artifacts, such as distinguishing clusters related to different habitats on the human body from artifacts caused by different sequencing methodologies such as PCR primers (The Human Microbiome Project Consortium, 2012).

When microbiome sequencing data (Fig. 1.1a) are arranged into count tables (Fig. 1.1b), such as those that count 16S amplicon sequence variants (ASVs) or the microbial genes present in a sample, the number of features being counted across all of the samples often vastly outnumbers the number of samples observed. This phenomenon of having many features, and particularly having far more features than samples, is a hallmark of high-dimensionality. For example, the EMP (Thompson et al., 2017) contained 23,828 samples and represented 307,572 ASVs, where each of these ASVs is considered a dimension of the resulting count table. This degree of high feature dimensionality creates difficulties for interpreting data and calculating meaningful statistics, since humans cannot visualize more than 3 dimensions, many of the features are noisy or redundant, the number of hypotheses that explain the data is far greater than the number of observations, and the number of features can cause run-time issues for downstream analysis. These are all common consequences of the "curse of dimensionality". Dimensionality reduction transforms a high-dimensional dataset into a representation with fewer dimensions, while retaining the key relationships among samples from the full dataset, making analysis tractable. Accordingly, dimensionality reduction is a core step in microbiome analyses, both for creating human-understandable visualizations of the data and as the basis for further analysis. The EMP used dimensionality reduction to produce plots of the 23,828 samples using

3 coordinates (in contrast to the 307,572 ASVs) that demonstrate the large difference between host-associated and non-host-associated microbiomes, and between saline and non-saline free-living microbiomes (Fig. 1.1c). These differences in microbial communities were subsequently statistically validated. This example is particularly salient because it shows the value of preserving the structure of the data while using much less information to represent it. Owing to its importance, dimensionality reduction methods are included in many analysis packages, including QIIME 2 (Bolyen et al., 2019), mothur (Schloss et al., 2009), and phyloseq (McMurdie and Holmes, 2013), as well as online software such as Qiita (Gonzalez et al., 2018) and MG-RAST (Keegan et al., 2016).

Figure 1.1: Overview of dimensionality reduction pipeline. Nucleotide sequences (a) from a biological experiment are organized in a feature table (b) containing the abundance of each feature in each sample. (c) Beta diversity plots showing unweighted UniFrac coordinates of EMP data annotated by EMPO levels 2 and 3 modified from Thompson et. al. (2017) (CC BY 4.0).

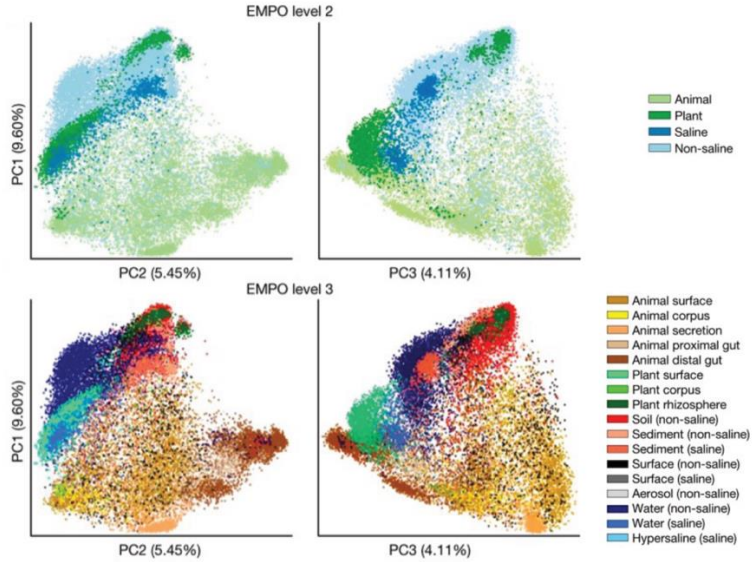
Sequences (a)

CAACGAATA
ATATGAAGA
CAATAATAC
ATACGGAGC
ATACTAGAA
CAACGAATA
GATCGAATA
.
.
.

Table (b)

	Feature 1	Feature 2	Feature 3	
Sample 1	2	0	1	. . .
Sample 2	0	1	0	. . .
Sample 3	1	0	0	. . .
.	.	.	.	
.	.	.	.	
.	.	.	.	

Dimensionality Reduction (c)



In this review, we describe how the characteristics of microbiome data complicate dimensionality reduction. We then discuss common strategies for dimensionality reduction, examining in detail whether and how they address each of the aspects that, in conjunction, confound microbiome analysis. Tried-and-true techniques, although useful, often have conceptual and practical problems that limit their utility in the microbiome, due to the inability to handle the data's most salient traits simultaneously. In this light, we then focus on examples of how dimensionality reduction techniques have been used in the literature, highlighting biological findings that have been revealed by each, while also discussing what may have been obscured. We then discuss common artifacts of widely used dimensionality reduction techniques, including specific pitfalls that users of these techniques must avoid in order to draw conclusions that are robust, reproducible, and well-supported by their data. We end with guidance on how dimensionality reduction should be used responsibly by practitioners in the field, and with an outlook describing how additional techniques that are seldom used today might provide valuable advances

1.2. Specific features of microbiome data that complicate dimensionality reduction

"Microbiome data" most often refers to sequencing results from two primary methodologies. The first class of microbiome sequencing is known as "amplicon sequencing" where a specific gene or region of a gene is targeted in each sample. 16S, 18S, and ITS sequencing approaches all fall under this class of methods. Variants of the targeted nucleotide sequences are used as a proxy for discrete microbial taxa. These unique sequences can be clustered by sequence similarity into "operational taxonomic units" (OTUs) or used by themselves as individual units after denoisers, such as DADA2 & Deblur, resolve the individual sequence variants from error-prone sequences (Callahan et al. 2016; Amir et al. 2017). These filtered sequences are often called amplicon sequence variants (ASVs) (Callahan et al., 2017) or sub-

OTUs (sOTUs). The second class of microbiome sequencing is shotgun or whole metagenome sequencing. In this method, the DNA from a sample is collected and sequenced broadly. The reads are then mapped to a reference database to determine the corresponding units, which can range from taxonomic identities to gene families or genes from a specific reference genome.

The result of these sequence analysis pipelines is typically a “feature table” that counts the microbial “units” or features (OTU, ASV, MAG, etc.) associated with each sample. Additionally, information about the relationship between features, such as taxonomic identity or gene family, can optionally be used to “collapse” the feature table to a lower resolution sum of its units. At this point, the data are generally ready to pursue exploratory analysis with dimensionality reduction.

However, there are several features common to microbiome data that can make standard dimensionality reduction techniques difficult to apply or to interpret. Each method must therefore handle each of these key issues or be benchmarked carefully to determine that these issues do not strongly affect the results in ways that are problematic for biological interpretation.

High dimensionality. In this context, “dimensionality” refers to the number of features in a feature table. Microbiome data typically have far more features than samples. Across studies ranging from tens of samples to tens of thousands of samples, the number of features for taxonomic data typically exceeds the number of samples by 20-fold or more. With gene-oriented data, the number of genes represented in a metagenomic study typically exceeds samples by several orders of magnitude. This can lead many statistical methods to overfit or to produce artifactual results.

Sparsity. Most microbes are not found in most samples, even of the same biospecimen type, for example, most human stool specimens from the same population have relatively low shared taxa (Allaband et al., 2019). As a result, a feature table containing counts of each microbe in each sample often has many zeros corresponding to unobserved microbes. Most 16S microbiome datasets do not have even as many as 10% of the possible entries observed in most

of the specimens. Feature tables with this over-abundance of unobserved counts are said to be "sparse", posing problems for statistical analysis. Moreover, the proportion of observed values tends to decrease as additional samples are sequenced, often leading to tables with density well below 1% (Hamady and Knight, 2009; McDonald et al., 2012).

Compositionality. In any high-throughput sequencing experiment, we impose an implicit limitation and randomness to the number of reads from a given sample due to many factors, including the random sub-sampling occurring in the process of collecting samples as well as uncontrolled variation in how efficiently each sample is amplified and incorporated into molecular libraries for sequencing. This limitation, termed "compositionality", should always be kept in mind when performing any microbiome analysis on abundance data. The total number of sequences per sample can affect the distances between samples (Weiss et al. 2017). Strategies such as rarefaction and relative abundance normalization are common for normalizing differences in sequencing depth. However, the relative amount of one feature in the sample is not independent from the counts of the other features. A difference in just one feature of the original sample can induce an observation that many other features are also changing (Morton et al. 2019) and neither rarefaction nor relative abundance sampling solve this issue. Due to this effect, many dimensionality reduction methods, such as PCA, will emphasize false correlations in the data.

Repeated measures. One of the most challenging experimental aspects to account for in dimensionality reduction is repeated measures data, e.g., multiple timepoints from the same subject where the variation between subjects may be greater than the variation between timepoints (Wu et al., 2011). In the context of dimensionality reduction, subjects or sites with multiple samples represented (such as in longitudinal studies or replicate analysis) provide an additional source of variation that can inhibit interpretation of the experimental effect of interest; the samples from a single subject can be highly correlated, resulting in between-subject differences dominating the ordination (e.g., (Song et al., 2016)).

Feature interpretation. Analysis of high-dimensional microbiome data is often motivated to find microbial biomarkers associated with observed differences in sample communities (Fedarko et al., 2020). This line of inquiry is of interest for diagnosis and/or prognosis of disease status, dysbiosis, and a host of other biological questions. Although this task is often addressed with differential abundance methods, those methods make specific statistical assumptions and may not correspond to the group separation observed in an exploratory analysis performed with any dimensionality reduction method (Lin and Peddada, 2020). Thus, methods that offer a quantitative justification of their representation in terms of the microbial features are often desirable. However, methods with feature importance that are not specifically designed for the microbiome often fail to account for compositionality, which can include many false positives due to the induced correlation of features, and sparsity, where important but infrequently observed features will not be detected (false negatives).

Complex patterns. Microbiome data are often assumed to contain clusters or gradients (Kuczynski et al., 2010). For example, multiple samples swabbed from one's own keyboard are more likely to be similar to each other than samples from another individual's keyboard (Fierer et al., 2010), and the microbial composition of soils is expected to vary continuously with soil pH (Lauber et al., 2009). However, with larger and larger datasets with many covariates and metadata on these being collected, more complex patterns can be detected (Debelius et al., 2016), such as grouping by both biological and technical factors in the case of the Human Microbiome Project (The Human Microbiome Project Consortium, 2012). Furthermore, many conventional dimensionality reduction methods, such as principal component analysis (PCA), assume the data lie in a linear subspace, and this assumption is violated by microbiome data (Ginter and Thorndike, 1979; Greig-Smith, 1980; Potvin and Roff, 1993; Tabachnick and Fidell, 2013).

1.3. Strategies for dimensionality reduction in the microbiome

The problems that complicate dimensionality reduction in microbiome data are scattered throughout the analysis pipeline. Difficulties can arise immediately from the raw sequence count data. Many can be corrected before the dimensionality reduction step, with careful preprocessing, though this can raise other issues. Furthermore, beta-diversity analysis, which seeks to quantify the pairwise differences in microbial communities among all samples with dissimilarity metrics (tailored to microbiome data), is often helpful for addressing many of the aforementioned circumstances (Pielou, 1966). Algorithms that are able to incorporate these metrics are particularly valuable, and this can be done in a variety of ways. Finally, additional constraints can be placed on dimensionality reduction algorithms to account for study design or provide additional information about the correspondence between the features and the reduced dimensions. In this section, we discuss each of these strategies in depth.

Compositionally Aware: Comparisons between and among samples must consider how sampling and sequencing depth can affect projection into low-dimensional space. Traditionally, compositionality has been addressed using logarithmic transformations of feature ratios. Transformations such as the additive log-ratio (ALR), centered log-ratio (CLR), and isometric log-ratio (ILR) can convert abundance data to the space of real numbers such that analysis and interpretation are less skewed by false positives (Aitchison and Greenacre, 2002; Pawlowsky-Glahn and Buccianti, 2011). After transformation, the Euclidean distance can be taken directly on the log-ratio transformed data (referred to as Aitchison distance) (Aitchison and Greenacre, 2002). Dimensionality reduction methods that incorporate log-ratio transformations attempt to preserve high-dimensional dissimilarities while taking into account the latent non-independence of microbial counts.

Pseudocounts and Imputation: High-dimensional microbiome data is almost always plagued by problems of “sparsity”, or an overabundance of zeroes. The data transformations to

address compositionality (as outlined above) are often based on logarithmic functions which are undefined at zero. The simplest solution is to add a small positive pseudocount to each entry of the feature table so that logarithmic functions can be applied. However, downstream analyses based on this approach are sensitive to the choice of pseudocount (Kumar et al., 2018) and there does not exist a standardized way to choose such a value. Other options include imputation of zeros (Martín-Fernández et al., 2003) through inference of the latent vector space. Fundamentally, zero handling is complicated by the inherent unknowability of the zero generating processes for each zero instance. In Silverman et al. (2020), they characterize the three different types of zero-generating processes (ZGP) as sampling, biological, and technical and demonstrate how the results of different zero-handling processes are affected by the (unknowable) mix of ZGPs in a given dataset. Recently Martino et al. (2019) introduced a version of the CLR transform that only computes the geometric mean on the non-zero components of a given sample. This avoids the problem of logarithms being undefined at 0 and thus dimensionality reduction through this method is robust to the high levels of sparsity in microbiome data.

Incorporating Phylogeny: Organisms identified using microbiome data can be related to one another through hierarchical structures that describe their evolutionary relationships. Typically, these structures take the form of either a taxonomy or a phylogeny. A taxonomy is a description of the organism relationships, generally derived subjectively using multiple biological criteria. A phylogeny, in contrast, is an inference of a tree, commonly with branch lengths, derived from quantitative algorithms that are typically applied to microbial, nucleic acid, or protein sequence data. Taxonomies have the advantage of being more directly interpretable because hierarchical structures correspond to a defined organization and classification pattern curated by experts in the field. However, these assignments and hierarchies are often putative and subject to change as more information about microbial taxa emerges. In contrast, phylogenies are derived from quantitative measures of sequence similarity from sample reads. These data structures are

more easily incorporated into statistical analyses but often at the cost of less interpretability as the hierarchical structures do not necessarily map to pre-defined microbial relationships. These evolutionary relationships, particularly phylogenies, add information to microbiome analysis, because related organisms are more likely to exhibit similar phenotypes (although counterexamples do exist, especially closely related taxa such as *Escherichia* and *Shigella*, which are very similar genetically but produce different clinical phenotypes).

When comparing the similarity of pairs of microbial communities, it is possible to utilize these hierarchical structures, and derive a metric that computes a distance as a function of shared evolutionary history (Lozupone and Knight, 2005). Specifically, communities that are very similar will share most of their evolutionary history, whereas those that are very dissimilar will have relatively little in common. A popular form of phylogenetically-aware distances is the suite of UniFrac metrics, which includes both quantitative (Lozupone et al., 2007) and qualitative (Lozupone and Knight, 2005) forms. Numerous extensions to UniFrac have been developed (Chang et al., 2011; Chen et al., 2012), including variants that account explicitly for the compositional nature of microbiome data (Wong et al., 2016). Because these metrics all utilize not only exactly observed features, but also the relationships among features, they can better account for the sparsity of microbiome data which manifests at the tips of a phylogenetic tree (because most microbes are not observed in most environments). In contrast, a metric like the Euclidean distance is limited to only the information at the tips of these hierarchies, and, worse, assumes that all features at the tips are equally related to one another (so that in a tree consisting of a mouse, a rat, and a squid, there is no allowance for the fact that the two rodents are much more similar to each other than they are to the squid). Neither phylogenetic nor non-phylogenetic beta-diversity measures explicitly model differences in sequencing depth per sample, although these differences in depth can be standardized through rarefaction (Weiss et al., 2017).

Operates on Generalized Beta-Diversity Matrix: Many of the issues outlined above can be

easily addressed at the sample dissimilarity level rather than directly through dimensionality reduction algorithms. A number of dissimilarity/distance metrics have been developed to account for factors such as phylogenetic data incorporation, compositionality, or sparsity that output a sample by sample matrix estimating high-dimensional dissimilarity. Dimensionality reduction methods that operate on arbitrary dissimilarity metrics are attractive options because the complex handling of the various feature table issues can be split into the choice of dissimilarity metric and the choice of dimensionality reduction algorithm. This adds a layer of flexibility for researchers to analyze their data depending on their needs. Methods based on multidimensional scaling approaches such as PCoA (Kruskal and Wish, 1978) and nMDS (Kruskal, 1964) attempt to preserve as much as possible the pairwise distances between subjects. Other methods such as t-distributed stochastic neighbor embedding (t-SNE) (van der Maaten and Hinton, 2008) and Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) are non-linear dimensionality reduction techniques that aim to find a low-dimensional representation such that similar data points are placed closed together and dissimilar points are pushed apart. A caveat of these methods is that they can be very sensitive to the choice of dissimilarity used. Patterns that may appear from one measure of dissimilarity may not be as apparent in a different measure. As an example, phylogenetic metrics such as UniFrac may differ from non-phylogenetic metrics such as Bray-Curtis depending on the strength of phylogenetic contribution (Shankar et al., 2017). The choice of dissimilarity metric should therefore be considered carefully, as different dimensionality reduction techniques yield visually and statistically very different results on the same data (Kuczynski et al., 2011).

Linear vs. Non-Linear Methods: Principal coordinates analysis (PCoA) and PCA are popular dimensionality reduction techniques that fall under the "linear" category. Linear techniques attempt to reduce or transform the data such that an approximation of the original data can be reconstructed by a weighted sum of the resulting coordinates. These methods typically

involve computing decompositions/factorizations of the data that are highly computationally efficient and work well on data that is naturally linear. Various other techniques, such as robust Aitchison PCA (RPCA) (Martino et al., 2019), and nonnegative matrix factorization (NMF) (Lee and Seung, 1999) also fall under this class of techniques.

Other methods fall under the "non-linear" category, which perform more complex transformations that often excel at preserving different patterns that may not be linear. This category includes methods such as the non-metric multidimensional scaling (nMDS), t-SNE, and UMAP. These methods can more succinctly represent complex patterns, but possibly at the expense of additional computation. Furthermore, these models tend to have randomness (such as from initialization) and more hyperparameters that the output can be highly sensitive to, so it is usually necessary to run these algorithms multiple times to ensure the conclusions are reproducible. Other non-linear methods that have seen less frequent use in microbiome data (and bioinformatics generally) include kernel PCA (Scholkopf et al., 1999), locally linear embeddings (Roweis and Saul, 2000), Laplacian eigenmaps (Belkin and Niyogi, 2001), and ISOMAP (Tenenbaum et al., 2000).

Unlike its close, linear counterpart PCoA, nMDS performs the ordination onto a pre-specified number of dimensions and operates on the ranks of the dissimilarities, rather than the dissimilarities themselves. This rank-based approach can be beneficial for representing data that departs from the assumptions of linearity. Other non-linear methods, such as t-SNE and UMAP, also transform the data onto a pre-specified number of dimensions and operate by assuming the high-dimensional data follow a non-linear structure that can be represented with fewer dimensions.

Repeated Measures: If the biological variable of interest occurs at the subject level, repeated samples (such as through a longitudinal study design) can artificially inflate how tight a cluster appears in low-dimensional space. Dimensionality reduction methods for microbiome

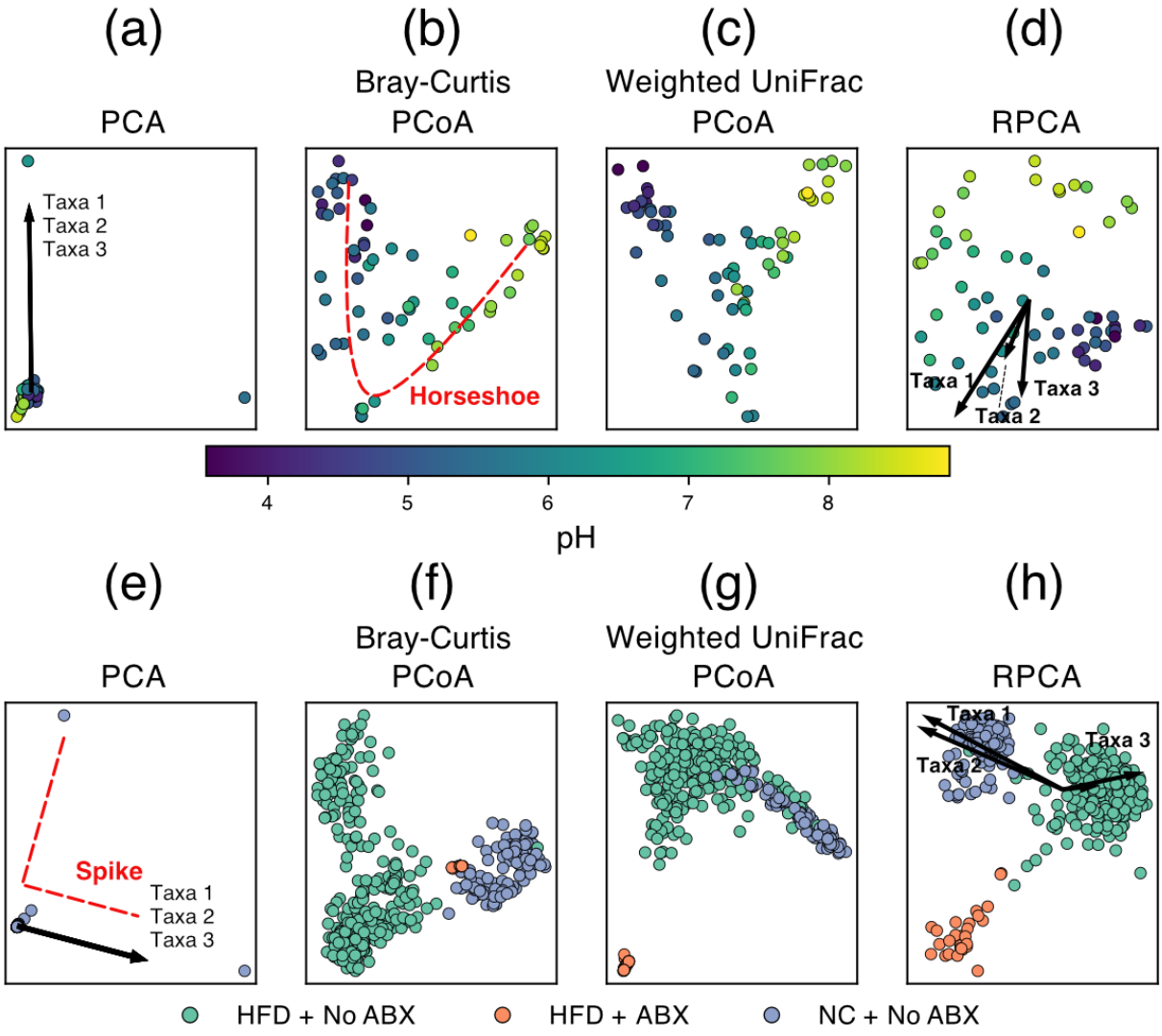
need to be designed for the purpose of handling this kind of data, with the intent to represent the relationships between explanatory variables while accounting for the inherent similarity between samples from the same subject. Methods to account for repeated measures can incorporate the relationship between individual samples and subjects by subject-aware decomposition of the data (Martino et al., 2021). There has also been discussion about incorporating prior sample relationship information into ordinations through Bayesian methods (Ren et al., 2017). Nevertheless, methods that incorporate repeated measures remain an underexplored area in dimensionality reduction literature.

Feature Importance: When the lower-dimensional representation of microbial communities shows separation between sample groups, a natural next question is what microbes or groups of microbes are driving such a separation. Dimensionality reduction methods that return a quantitative relationship between individual microbial features and the latent lower-dimensional space are a powerful class of methods that can demystify the construction of the lower-dimensional axes. However, certain methods that attempt to find high-dimensional patterns, such as non-linear methods, do not have an explicit interpretable correspondence between the output coordinates and the input features.

The most relevant category of methods for visualizing feature importance is the biplot ordination family of approaches. Biplots display both the samples and the driving variable vectors in reduced dimension space (**Fig. 1.2a, d, e, h**). For example, PCA naturally quantifies the contribution of each microbe to the principal component axes through matrix factorization into linear combinations of features. RPCA modifies this approach to account for compositionality and sparsity while retaining interpretable feature loadings (Martino et al., 2019). Another set of ecologically motivated matrix factorization methods is the correspondence analysis (CA) family. The general CA method can be thought of as an implementation of PCA that operates on count data. It is also possible to explicitly incorporate sample metadata into these dimensionality

reduction methods. Researchers are often interested in the explanatory power of their sample metadata (site, pH, subject, etc.). Certain dimensionality reduction methods can take as input both a feature table and a table of sample metadata to jointly estimate the low-dimensional representation of samples as well as the relative contribution of the provided metadata vectors. The general goal of these methods is to determine whether and/or which explanatory variables may be driving the differences in microbial communities among samples. Canonical correspondence analysis (CCA) is an extension of CA that incorporates sample variables of interest to determine which covariates are associated with the placement of samples and feature vectors in low-dimensional space (ter Braak, 1985). The results of CCA can be visualized as a "tri-plot" where samples are simultaneously visualized with the relative contribution of features and explanatory variables near related samples. These contributions can motivate subsequent statistical analysis of associations between sample metadata and specific microbial taxa.

Figure 1.2: Examples of dimensionality reduction techniques applied to publicly available microbiome data. (Top) Beta-diversity plots of soil samples colored by pH from (Lauber et al., 2009). (Bottom) Beta-diversity plots of murine fecal samples colored by diet and antibiotics usage from (Shalapour et al., 2017). (HFD = high-fat diet, NC = normal chow, ABX = antibiotics). PCA plots (a, e) show extremely high sample overlap due to outliers and characteristic "spike" artifacts. The top three taxa driving variation also overlap as shown by arrow superposition. (b) "Horseshoe" pattern emerges for samples following ecological gradients such as pH. RPCA plots (d, h) show the top three taxa driving separation of groups. (f) and (g) show strong overlap of HFD + ABX samples resolved by (h).


(a)
PCA
Taxa 1
Taxa 2
Taxa 3
(b)
Bray-Curtis
PCoA
Horseshoe
(c)
Weighted UniFrac
PCoA
(d)
RPCA
Taxa 1
Taxa 2
Taxa 3
4
5
6
7
8
pH
(e)
PCA
Spike
Taxa 1
Taxa 2
Taxa 3
(f)
Bray-Curtis
PCoA
(g)
Weighted UniFrac
PCoA
(h)
RPCA
Taxa 1
Taxa 2
Taxa 3
HFD + No ABX
HFD + ABX
NC + No ABX

1.4. Uses of dimensionality reduction for microbiome data

Over the past decade, PCoA has seen an increase in use in microbiome analyses, and it is the primary ordination method for beta-diversity included by default in workflows such as QIIME2 (Bolyen et al., 2019). It is typically used for exploratory visualization, as it excels at rendering biologically relevant patterns, such as clusters and gradients (Kuczynski et al., 2010). When used as an exploratory tool, observed patterns are often followed with statistical analysis on the original feature tables or dissimilarity matrices (Galloway-Peña and Hanson, 2020), such as ANOSIM (Clarke and Ainsworth, 1993), PERMANOVA (aka Adonis) (Anderson, 2017), ANCOM (Mandal et al., 2015), or bioenv (Clarke and Ainsworth, 1993). It should also be noted that some of these statistical techniques use the full table or distance matrix, not the reduced dimension matrix as visualized (at least by default) and may therefore introduce incongruent results between the statistics and the visualization.

Exploratory visualizations have revealed microbial-associated patterns in applications ranging from host-associated gut microbiomes to soil, ocean, and other environmental microbiome contexts. For example, studies have applied PCoA to demonstrate differences between host groups, such as differences between humans', chimpanzees', and gorillas' gut microbial taxa (Campbell et al., 2020), or the correspondence between human gut microbiomes and westernization (Yatsunenko et al., 2012; Campbell et al., 2020). Host microbiome-disease associations have also been identified using PCoA, such as in the case of colorectal cancer (Young et al., 2021) in humans and metritis in cows (Galvão et al., 2019). Uses also extend to host-environment relationships, such as demonstrating the differences between oyster digestive glands, oyster shells, and their surrounding soils (Arfken et al., 2017). The microbiome-shaping roles of environmental factors such as salinity in shaping free-living environments (Lozupone and Knight, 2007), pH in arctic soils (Malard et al., 2019) and depth in the ocean (Sunagawa et al., 2015) have also been elucidated with PCoA. In many of these cases, the PCoA visualizations

demonstrated a separation between groups that was subsequently followed by statistical validation with PERMANOVA or ANOSIM.

In numerous other instances, PCoA has also been used to make claims that extend beyond exploratory group differences followed by statistical analysis. For example, Halfvarson et al. (2017) fit a plane to the healthy subjects in the first three coordinates of a PCoA and then used the distance to this plane to associate dissimilarities in the microbiome with the severity of irritable bowel disease (IBD) (Halfvarson et al., 2017); this approach has subsequently been replicated (Gonzalez et al., 2018). Others have used regression of participant and microbiome characteristics (e.g., age and alpha diversity, respectively) on PCoA coordinates to determine whether the given factors have a significant relationship with microbial community composition in the context of dietary interventions (Lang et al., 2018). In one case, while providing visualization with PCoA and statistical confirmation with ANOSIM, Vangay et al. (2018) additionally plotted ellipses for visualizing cluster centers/spread in their PCoA coordinates (Vangay et al., 2018). In another instance, Metcalf et al. (2017) showed the correspondence of dissimilarities between the 16S rRNA profiles and chloroplast marker profiles by performing a Procrustes analysis on the separate ordinations of the different data types (Metcalf et al., 2017).

We note that the choice of dissimilarity metric can have a significant impact on the low-rank embedding depending on the dataset. Shi et. al. (2021) review the effect of high and low-abundance operational taxonomic units have on unsupervised clustering of Bray-Curtis and unweighted UniFrac (Shi et al., 2021). Marshall et. al (2019) compare Bray-Curtis ordination with weighted UniFrac on marine sediment samples and note that the most relevant clustering variable differed depending on the dissimilarity used (Marshall et al., 2019). These results imply that interpretation of low-dimensional embeddings and the putative driving variables must be performed in the context of the choice of dissimilarity. Metrics such as Bray-Curtis and weighted UniFrac take into consideration the abundance of individual microbes in each sample which can

be important for datasets with many rare taxa. In contrast, some dissimilarity metrics such as Jaccard and unweighted UniFrac are only defined on binarized data, which may mask this property. Furthermore, phylogenetic metrics such as the UniFrac suite of metrics are best when the evolutionary relationships among microbial features is of interest in the context of sample communities. These metrics may also be more appropriate than other methods for datasets with particularly high sparsity. Note that metrics such as Bray-Curtis dissimilarity, which are not strictly distance metrics, can be symmetrized to yield a valid distance matrix.

PCA is arguably the most widely used and popular form of dimensionality reduction, which does not allow generalized beta-diversity distances (e.g., PCoA or UMAP), but does allow for the direct interpretation of feature importances relative to sample separations in the ordination. However, due to compositionality and sparsity, PCA often leads to spurious results on microbiome data (Hamady and Knight, 2009; Morton et al., 2017). Aitchison PCA attempts to fix these issues by using log transformation, but imputation is required (because the log of zero is undefined). Therefore, (Martino et al., 2019) proposed the adoption of RPCA for dimensionality reduction. This method has been shown to discriminate between sample groups in a wide array of biological contexts, including fecal microbiota transplants (Goloshchapov et al., 2019), cancer (Bali et al., 2021), and HIV (Parbie et al., 2021). Moreover, the generalized version of this technique accounts for repeated measures, allowing for large improvements in the ability to discriminate subjects by phenotypes across time or space (Martino et al., 2021). This advantage has been crucial in the statistical analysis of complicated longitudinal experimental designs such as early infant development models (Song et al., 2021). Feature loadings from these PCA-based methods can be used to inform selection of microbial features for log-ratio analysis (Morton et al., 2019; Fedarko et al., 2020), leading to novel biomarker discovery.

For feature interpretation, CCA is the most commonly used CA-based method for analyzing high dimensional microbiome data, due to its ability to incorporate sample metadata

into the low-rank embeddings. This strategy has shown success in differentiating clinical outcomes following stem cell transplantation (Ingham et al., 2019) as well as diarrhea status in children (Dinleyici et al., 2018). CCA has also shown success in projecting environmental samples into lower-dimensional space such as in rhizosphere microbial communities (Benitez et al., 2017; Pérez-Jaramillo et al., 2017), and aerosol samples (Souza et al., 2021). Another approach designed for microbial feature interpretation has been posed by (Xu et al., 2021), explicitly modeling the ZGP through a zero-inflation model. This method attempts to optimize a statistical model for jointly estimating the "true" zero-generating probability as well as the Poisson rate of each microbial count.

Of non-linear methods, nMDS has historically been more widely used in microbiome data analysis, in part because it can incorporate an arbitrary dissimilarity measure. Furthermore, since nMDS is a rank-based approach, it is less likely than linear methods to be highly influenced by outliers in beta-diversity dissimilarities. Recent uses have involved using nMDS to show differences in the gastric microbiome between samples from patients with gastric cancer cases against the control of gastric dyspepsia (recurrent indigestion without apparent cause) (Castaño-Rodríguez et al., 2017) and demonstrating differences in the gut microbiome based on diabetes status (Das et al., 2021). In both of these cases, the visual distinction between groups was supported by PERMANOVA.

Other non-linear methods have been increasingly used for analyzing other types of sequencing data, especially in the single-cell genomics field, but have not yet been widely deployed in the microbiome. The most popular of these methods for visualization, t-SNE and UMAP, are starting to see more use in the microbiome field. (Xu et al., 2020) developed a method to classify microbiome samples using t-SNE embeddings. We recently reviewed the usage and provided recommendations for implementing UMAP for microbiome data (Armstrong et al., 2021).

UMAP with an input beta-diversity dissimilarity matrix can reveal biological signals that may be difficult to see with traditional methods such as PCoA.

1.5 Artifacts and cautionary tales in dimensionality reduction

Dimensionality reduction is incredibly useful and has led to many interesting biological conclusions. However, when using dimensionality reduction techniques, one must be careful how results are interpreted. There are known examples of patterns that are induced by the properties of the data alone (rather than the relationships among specific samples or groups of samples), and others that are a product of the method itself. Here, we discuss several known issues, as well as insights into evaluating the degree to which an ordination represents the actual data.

One of the most well-known artifacts in microbial ecology is the horseshoe effect (Podani and Miklós, 2002), wherein the ordination has a curvilinear pattern along what otherwise appears to be a linear gradient. This pattern can occur when a variable, such as soil pH (Lauber et al., 2009) or length of time of corpse decay (Metcalf et al., 2016) corresponds with drastic changes in microbiome composition on a continuous scale. Since the characteristic "bend" in the horseshoe typically occurs along the second coordinate of a PCoA (**Fig. 1.2b**), it can obfuscate additional gradients/associations along that axis. Recent research in the topic has also identified that indeed, it is unlikely the horseshoe appears from a real effect, and instead it is a product of the limitations of many distance metrics to capture distance along a gradient when no features are shared between many of the samples (i.e., saturation) (Morton et al., 2017), which can be an issue with many common metrics, such as Euclidean, Jaccard, and Bray-Curtis distances (Morton et al., 2017). As a result, a possible remedy for the artifact is to use a distance metric that considers the relationships between features, such that two samples that share no features do not necessarily have the same dissimilarity as two different samples that share no features, e.g, UniFrac or weighted UniFrac. If a change in metric does not resolve the issue, it may be possible to avoid

the horseshoe artifact by using RPCA or a non-linear method (e.g., UMAP). "Spikes" are another artifact, more prevalent on cluster-structured data, where outliers dominate the embedding and it fails to separate into clusters in the visualization (Vázquez-Baeza et al., 2017). Spikes also appear to be mitigated with an appropriate choice in distance metric, such as UniFrac (Hamady and Knight, 2009). In both cases, since the issues are with representing the distances between distant or extreme samples, non-linear methods (such as UMAP or nMDS) that dampen the effect of outliers provide a potential workaround to reveal secondary gradients or the obfuscated cluster structures (Armstrong et al., 2021). Though it is possible that the benefits offered by non-linear methods for the horseshoe effect are limited by the aspect ratio of the gradient (Kohli et al., 2021), and potentially the parameters of the algorithms.

Dimensionality reduction is also commonly used in other bioinformatic disciplines. Particularly, single-cell transcriptomics has used dimensionality reduction prolifically, with many publications using PCA, t-SNE, or UMAP visualizations. Furthermore, single-cell RNA-seq data shares many properties with microbiome data, including sparsity/zero-inflation, sequencing depth differences, and even phylogenetic relationships (Lähnemann et al., 2020). This connection is further strengthened by the fact that researchers in both disciplines investigate similar types of questions, albeit with different underlying data. Microbiome researchers often ask whether there is a difference between different treatments or disease-statuses (David et al., 2013; Lloréns-Rico et al., 2021), and which microbes contribute to those differences (i.e., differential abundance analysis). Similarly, transcriptomics may investigate parallel scenarios (Ocasio et al., 2019; Taavitsainen et al., 2021), where the goal is to discover transcripts whose expression stratifies the desired groups (i.e., differential expression).

Despite these similarities, the most popular methods for dimensionality reduction in microbiome and single-cell publications differ significantly, with PCoA being more prevalent among microbiome publications, and t-SNE (or variants (Linderman et al., 2019)) and UMAP more

prevalent in single-cell publications (Kobak and Berens, 2019). Given the similarities in hypotheses and the properties of the data, but use of different methods, it is reasonable to suppose that methods such as t-SNE and UMAP have potential utility in the microbiome. However, global distances are not necessarily preserved in these methods, therefore distances between different clusters should not be interpreted as demonstrating similarity or dissimilarity. Consequently, recent research concerning the representation of single-cell RNA-seq data should also be taken into account when applying these methods to microbiome data.

First, t-SNE and UMAP are fairly complex algorithms that have many hyperparameters that can be adjusted, so it is important to be able to evaluate the faithfulness of the embeddings they produce. The evaluation of dimensionality reduction has been performed with many different measures, each of which has its own characteristics. Some measures reward embeddings that adequately preserve the local-scale structures in the embedding but do not necessarily penalize inaccurate representations of large distances in the original high-dimensional data, like the KNN evaluation measure (Kobak and Berens, 2019), which takes the average accuracy of the k=10 nearest neighbors in the reduced dimensions compared to the original space. Others, such as the correlation (either Pearson or Spearman) between distances in the original space and reduced dimensions have been used (Becht et al., 2019; Kobak and Berens, 2019; Kobak and Linderman, 2021). The correlation measure generalizes whether the two representations overall are similar, i.e. close points in the original space are close in the low-dimensional space, and similar for far points. However, high correlation does not guarantee that the fine-scale structures have been preserved. Additionally, measures that use sample metadata about known classes can be used, such as the KNC measure (Kobak and Berens, 2019), which measures whether the closest class/category centers to a given category are preserved in the embedding. KNC emphasizes the preservation of relationships between classes, but not necessarily structures within the classes or between distant classes. These measures have been used to evaluate the quality of several

dimensionality reduction methods across a variety of parameter settings on complex datasets. Notably, Kobak and Berens (2019) demonstrated on several single-cell transcriptomics datasets, that t-SNE with the default value for "perplexity" performed well at representing the relationships between nearby points (KNN), but poorly at representing the large-scale patterns (KNC and correlation). However, when they increased the perplexity parameter, they achieved improved KNC and correlation at the expense of a decreased KNN score. Kobak and Linderman (2021) observed with correlation that the best method (between t-SNE and UMAP) can vary by dataset. So, in practice, it may be necessary to compare multiple dimensionality reduction methods (and parameter settings) on a dataset using the measure that best suits the question, e.g., use the correlation measure when seeking a visualization of earth microbiomes by environment to show which environments are similar to each other.

Furthermore, since UMAP and t-SNE are algorithms that require configurable (possibly random) initializations, particular attention has been paid to their reproducibility. A metric to evaluate reproducibility comes from (Becht et al., 2019), which measures the preservation of pairwise distances in the embeddings by comparing an embedding on a subset of the points to the location of those points in the embedding of the entire dataset. In its original application, the reproducibility measure was used to demonstrate UMAP providing more reproducible results than t-SNE and variants of t-SNE. However, (Kobak and Linderman, 2021) showed that with appropriate (spectral) initialization, t-SNE can perform just as well by this metric as UMAP. While reproducibility is important, this metric should be applied carefully, because it fails to account for rotations in the embedding. Another important concern related to reproducibility is whether even random noise will yield apparent clusters. This phenomenon has been observed with t-SNE (Wattenberg et al., 2016), and whether other dimensionality reduction techniques are also susceptible to this effect warrants further systematic investigation. However, because these benchmarks are all performed within transcriptomics, further validation is needed to determine

whether the conclusions generalize to microbiome data. These measures provide a starting point for evaluating the application of non-linear dimensionality reduction techniques on microbiome data.

Finally, literature from mathematics and computer science that has not been as widely applied to dimensionality reduction in bioinformatics may also be relevant. Of particular interest is the study of distortion, which is applicable when the goal of the embedding is to preserve distances, like one might expect for an exploratory analysis. Similar to the previously described correlation measure, distortion measures summarize the extent to which the distances in high dimensions match the distances in low-dimensions, however, distortion is defined in terms of the expansions and contractions of distances between points. Furthermore, there are many ways to summarize the expansions and contractions, including the worst-case, average-case and local-case, which are all detailed more in (Vankadara and von Luxburg, 2018).

1.6. Discussion

The above examples illustrate that dimensionality reduction is an extremely powerful technique that has enhanced a wide range of microbiome studies. However, with great power comes great responsibility. It is unlikely that any one method will excel at representing all datasets, so responsible users of dimensionality reduction should try out several techniques, ideally guided by characteristics of the data rather than as a fishing expedition to see whether any one of many techniques produce results that “look good” (which may even happen in random data for some techniques and parameters) or that fulfill pre-conceived hypotheses and biases. We need standard protocols and software interfaces for choosing the algorithm that suits your data best, rather than the algorithm that shows what you want to see if you squint at it correctly. Methods are needed both for diagnosing the issues that may be most prevalent in your data and affecting

your representation, and for rationally choosing among different methods that could be applied to a given dataset. Developing these methods is a key priority for the field.

Dimensionality reduction for the purposes of visualization has somewhat different goals from dimensionality reduction for other purposes and developing a better appreciation of this distinction is important for practice in the field. The goal of dimensionality reduction for visualization is primarily for exploratory overview by human observers (do groups differ from one another, is there overall structure such as gradients in the data). As such, visualization is usually done with three dimensions (more can be examined through parallel plots), while the intrinsic dimensionality of the data may be higher. Visualization is typically only the first step in the data analysis pipeline, and is followed by downstream analysis, such as multivariate analysis/regression (PERMANOVA, ANOSIM, PERMDISP) either on the original distances or on a dimensionality-reduced version of the data (which can be higher than three dimensions). These results can also be used to motivate supervised differential abundance modeling, such as to determine which groups separate and then determine which microbes are driving these separations.

Dimensionality reduction is thus often an early step in a multi-step pipeline. What downstream analyses is dimensionality reduction a step towards, and how are these accomplished? Feature loadings (i.e. the importance of particular taxa or genes) can be interpreted using log ratios from tools such as DEICODE (Martino et al., 2019), which can then be visualized in Qurro (Fedarko et al., 2020). Classification can be accomplished using machine learning techniques such as random forests, allowing estimates of classifier accuracy and group stability, and also allowing tests of the reusability of these models, e.g. applying a model of human inflammatory bowel disease to dogs (Vázquez-Baeza et al., 2016) or models of aging between different human populations (Huang et al., 2020). A popular strategy is to use a lower-dimensional embedding for traditional statistical analysis, such as using PCA or PCoA coordinates as inputs

for regression, classification, clustering, and other analyses. However, as we have seen, many dimensionality reduction methods induce various kinds of artifacts or distortions, and cannot generalize well beyond the data on which the model was initially optimized on, including PCoA, nMDS, RPCA/CTF, and UMAP/t-SNE. Consequently, analyses on these coordinates should be performed with caution. Furthermore, since the parameters and software versions used with these methods have the potential to be highly influential to their results, we recommend that these always be reported for dimensionality reduction methods.

Given the large number of publications that have used dimensionality reduction on microbiome data, we can start to draw conclusions about which dimensionality reduction strategies should be more widely used, and which less widely used. On larger, sparser, compositional datasets, we recommend against the use of conventional PCA, Bray-Curtis and Jaccard distances, and pseudocounts. Conventional PCA presents the clearest case of a method that should not be used on microbiome data due the sparsity and compositional nature of the data. UniFrac and weighted UniFrac are essentially phylogenetically informed versions of Jaccard and Bray-Curtis beta-diversity metrics respectively. Due to the current default generation of a phylogeny in most 16S and shotgun analyses, there is no reason not to use the phylogenetic counterparts, which have been shown to have better discriminatory power. Pseudocounts should not be used because the choice of pseudocount impacts the lower-dimensional embedding, and there is no clear method for determining which pseudocount value is best.

In contrast, CTF and non-linear methods should be used more in microbiome contexts. As the cost of acquiring microbiome data continues to decrease, experimental designs are getting increasingly complex, and include repeated measures, longitudinal studies, batch effects, etc. We therefore need methods that can determine which biological signals are relevant among all these confounding factors. Additionally, we are increasingly recognizing that many relationships between/among samples are non-linear. Using non-linear methods can potentially explain more

of such datasets with fewer dimensions, although additional benchmarking is required to understand the performance of these methods.

Our analyses suggest some important gaps in the field that could be important areas for future development. There are no dimensionality reduction methods yet that are both able to incorporate phylogeny and are compositionally aware. Several methods, such as Robust PCA and CTF, control for the sparsity, non-normality, compositionality, and are adaptable to specific study-designs of microbiome data but do not incorporate phylogenetic information. In contrast, phylogenetic techniques do not account for sparsity and compositionality, and some also perform poorly with non-normality. A unified method that is appropriate for any microbiome study is therefore still in the future, despite many important recent advances. The ability to perform this task using a generalizable dissimilarity measure would be particularly useful, because it would allow for full utilization of PCoA and non-linear methods including nMDS and UMAP.

Taken together, we conclude that dimensionality reduction is a key part of many, if not most, of the highest-impact microbiome studies performed to date. We can expect this situation to continue into the future, especially as larger study designs and datasets continue to accumulate, and additional method development advances increase the speed and range of applicability of these techniques.

1.7. Acknowledgements

This work was supported in part by grants NSF 2038509, NIH U24CA248454, NIH 1DP1AT010885, and by CRISP, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA. https://crisp.engineering.virginia.edu/

Chapter 1, in full, is a reprint of the material as it appears in "Applications and comparison of dimensionality reduction methods for microbiome data." George Armstrong, Gibraan Rahman, Cameron Martino, Daniel McDonald, Antonio Gonzalez, Gal Mishne, and Rob Knight. *Frontiers in Bioinformatics*. The dissertation author is the co-first author of this paper in conjunction with Dr. George Armstrong.

1.8. References

1. Aitchison, J., and Greenacre, M. J. (2002). Biplots of compositional data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 51, 375–392.

2. Allaband, C., McDonald, D., Vázquez-Baeza, Y., Minich, J. J., Tripathi, A., Brenner, D. A., et al. (2019). Microbiome 101: Studying, Analyzing, and Interpreting Gut Microbiome Data for Clinicians. *Clin. Gastroenterol. Hepatol.* 17, 218.

3. Anderson, M. J. (2017). Permutational Multivariate Analysis of Variance (PERMANOVA). *Wiley StatsRef: Statistics Reference Online*, 1–15. doi:10.1002/9781118445112.stat07841.

4. Arfken, A., Song, B., Bowman, J. S., and Piehler, M. (2017). Denitrification potential of the eastern oyster microbiome using a 16S rRNA gene based metabolic inference approach. *PLoS One* 12, e0185071.

5. Armstrong, G., Martino, C., Rahman, G., Gonzalez, A., Vázquez-Baeza, Y., Mishne, G., et al. (2021). Uniform Manifold Approximation and Projection (UMAP) Reveals Composite Patterns and Resolves Visualization Artifacts in Microbiome Data. *mSystems* 6, e0069121.

6. Bali, P., Coker, J., Lozano-Pope, I., Zengler, K., and Obonyo, M. (2021). Microbiome Signatures in a Fast- and Slow-Progressing Gastric Cancer Murine Model and Their Contribution to Gastric Carcinogenesis. *Microorganisms* 9, 189. doi:10.3390/microorganisms9010189.

7. Barker, M., and Rayens, W. (2003). Partial least squares for discrimination. *J. Chemom.* 17, 166–173.

8. Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., et al. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology* 37, 38–44. doi:10.1038/nbt.4314.

9. Belkin, M., and Niyogi, P. (2001). Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. in *NIPS'01: Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic* doi:10.7551/mitpress/1120.003.0080.

10. Bellman, R. E. (2015). *Adaptive Control Processes*. Princeton University Press.

11. Benitez, M.-S., Osborne, S. L., and Lehman, R. M. (2017). Previous crop and rotation history effects on maize seedling health and associated rhizosphere microbiome. *Sci. Rep.* 7, 15709.

12. Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37, 852–857.

13. Callahan, B. J., McMurdie, P. J., and Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* 11, 2639–2643.

14. Campbell, T. P., Sun, X., Patel, V. H., Sanz, C., Morgan, D., and Dantas, G. (2020). The microbiome and resistome of chimpanzees, gorillas, and humans across host lifestyle and geography. *ISME J.* 14, 1584–1599.

15. Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., et al. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. U. S. A.* 108 Suppl 1, 4516–4522.

16. Castaño-Rodríguez, N., Goh, K.-L., Fock, K. M., Mitchell, H. M., and Kaakoush, N. O. (2017). Dysbiosis of the microbiome in gastric carcinogenesis. *Scientific Reports* 7. doi:10.1038/s41598-017-16289-2.

17. Chang, Q., Luan, Y., and Sun, F. (2011). Variance adjusted weighted UniFrac: a powerful beta diversity measure for comparing communities based on phylogeny. *BMC Bioinformatics* 12, 118.

18. Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., et al. (2012). Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics* 28, 2106–2113.

19. Clarke, K. R., and Ainsworth, M. (1993). A method of linking multivariate community structure to environmental variables. *Marine Ecology Progress Series* 92, 205–219. doi:10.3354/meps092205.

20. Das, T., Jayasudha, R., Chakravarthy, S., Prashanthi, G. S., Bhargava, A., Tyagi, M., et al. (2021). Alterations in the gut bacterial microbiome in people with type 2 diabetes mellitus and diabetic retinopathy. *Sci. Rep.* 11, 2738.

21. David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., et al. (2013). Diet rapidly and reproducibly alters the human gut microbiome. *Nature* 505, 559–563.

22. Debelius, J., Song, S. J., Vazquez-Baeza, Y., Xu, Z. Z., Gonzalez, A., and Knight, R. (2016). Tiny microbes, enormous impacts: what matters in gut microbiome studies? *Genome Biol.* 17. doi:10.1186/s13059-016-1086-x.

23. Dinleyici, E. C., Martínez-Martínez, D., Kara, A., Karbuz, A., Dalgic, N., Metin, O., et al. (2018). Time Series Analysis of the Microbiota of Children Suffering From Acute Infectious Diarrhea and Their Recovery After Treatment. *Front. Microbiol.* 9, 1230.

24. Galvão, K. N., Higgins, C. H., Zinicola, M., Jeon, S. J., Korzec, H., and Bicalho, R. C. (2019). Effect of pegbovigrastim administration on the microbiome found in the vagina of cows postpartum. *J. Dairy Sci.* 102, 3439–3451.

25. Fedarko, M. W., Martino, C., Morton, J. T., González, A., Rahman, G., Marotz, C. A., et al. (2020). Visualizing 'omic feature rankings and log-ratios using Qurro. *NAR Genom Bioinform* 2. doi:10.1093/nargab/lqaa023.

26. Fierer, N., Lauber, C. L., Zhou, N., McDonald, D., Costello, E. K., and Knight, R. (2010). Forensic identification using skin bacterial communities. *Proc. Natl. Acad. Sci. U. S. A.* 107, 6477–6481.

27. Galloway-Peña, J., and Hanson, B. (2020). Tools for Analysis of the Microbiome. *Digestive Diseases and Sciences* 65, 674–685. doi:10.1007/s10620-020-06091-y.

28. Ginter, J. L., and Thorndike, R. M. (1979). Correlational Procedures for Research. *Journal of Marketing Research* 16, 600. doi:10.2307/3150840.

29. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. *Front. Microbiol.* 0. doi:10.3389/fmicb.2017.02224.

30. Goloshchapov, O. V., Olekhnovich, E. I., Sidorenko, S. V., Moiseev, I. S., Kucher, M. A., Fedorov, D. E., et al. (2019). Long-term impact of fecal transplantation in healthy volunteers. *BMC Microbiol.* 19, 312.

31. Gonzalez, A., Navas-Molina, J. A., Kosciolek, T., McDonald, D., Vázquez-Baeza, Y., Ackermann, G., et al. (2018). Qiita: rapid, web-enabled microbiome meta-analysis. *Nat. Methods* 15, 796–798.

32. Greig-Smith, P. (1980). The development of numerical classification and ordination. *Vegetatio* 42, 1–9.

33. Halfvarson, J., Brislawn, C. J., Lamendella, R., Vázquez-Baeza, Y., Walters, W. A., Bramer, L. M., et al. (2017). Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat Microbiol* 2, 17004.

34. Hamady, M., and Knight, R. (2009). Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Res.* 19, 1141–1152.

35. Huang, S., Haiminen, N., Carrieri, A. P., Hu, R., Jiang, L., Parida, L., et al. (2020). Human Skin, Oral, and Gut Microbiomes Predict Chronological Age. *mSystems* 5. doi:10.1128/mSystems.00630-19.

36. Ingham, A. C., Kielsen, K., Cilieborg, M. S., Lund, O., Holmes, S., Aarestrup, F. M., et al. (2019). Specific gut microbiome members are associated with distinct immune markers in pediatric allogeneic hematopoietic stem cell transplantation. *Microbiome* 7, 131.

37. Keegan, K. P., Glass, E. M., and Meyer, F. (2016). MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. *Methods Mol. Biol.* 1399. doi:10.1007/978-1-4939-3369-3_13.

38. Kobak, D., and Berens, P. (2019). The art of using t-SNE for single-cell transcriptomics. *Nat. Commun.* 10, 5416.

39. Kobak, D., and Linderman, G. C. (2021). Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nat. Biotechnol.* 39, 156–157.

40. Kohli, D., Cloninger, A., and Mishne, G. (2021). LDLE: Low Distortion Local Eigenmaps. *J. Mach. Learn. Res.* Available at: https://arxiv.org/abs/2101.11055.

41. Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29, 1–27. doi:10.1007/bf02289565.

42. Kruskal, J., and Wish, M. (1978). Multidimensional Scaling. doi:10.4135/9781412985130.

43. Kuczynski, J., Lauber, C. L., Walters, W. A., Parfrey, L. W., Clemente, J. C., Gevers, D., et al. (2011). Experimental and analytical tools for studying the human microbiome. *Nat. Rev. Genet.* 13, 47–58.

44. Kuczynski, J., Liu, Z., Lozupone, C., McDonald, D., Fierer, N., and Knight, R. (2010). Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nat. Methods* 7, 813–819.

45. Kumar, M. S., Slud, E. V., Okrah, K., Hicks, S. C., Hannenhalli, S., and Corrada Bravo, H. (2018). Analysis and correction of compositional bias in sparse sequencing count data. *BMC Genomics* 19, 799.

46. Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., et al. (2020). Eleven grand challenges in single-cell data science. *Genome Biol.* 21, 1–35.

47. Lang, J. M., Pan, C., Cantor, R. M., Tang, W. H. W., Garcia-Garcia, J. C., Kurtz, I., et al. (2018). Impact of Individual Traits, Saturated Fat, and Protein Source on the Gut Microbiome. *MBio* 9. doi:10.1128/mBio.01604-18.

48. Lauber, C. L., Hamady, M., Knight, R., and Fierer, N. (2009). Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl. Environ. Microbiol.* 75. doi:10.1128/AEM.00335-09.

49. Lee, D. D., and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 788–791.

50. Ley, R. E., Lozupone, C. A., Hamady, M., Knight, R., and Gordon, J. I. (2008). Worlds within worlds: evolution of the vertebrate gut microbiota. *Nat. Rev. Microbiol.* 6, 776–788.

51. Linderman, G. C., Rachh, M., Hoskins, J. G., Steinerberger, S., and Kluger, Y. (2019). Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nat. Methods* 16, 243–245.

52. Lin, H., and Peddada, S. D. (2020). Analysis of microbial compositions: a review of normalization and differential abundance analysis. *NPJ Biofilms Microbiomes* 6, 60.

53. Lloréns-Rico, V., Gregory, A. C., Van Weyenbergh, J., Jansen, S., Van Buyten, T., Qian, J., et al. (2021). Clinical practices underlie COVID-19 patient respiratory microbiome composition and its interactions with the host. *Nat. Commun.* 12, 1–12.

54. Lozupone, C. A., Hamady, M., Kelley, S. T., and Knight, R. (2007). Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.* 73, 1576–1585.

55. Lozupone, C. A., and Knight, R. (2007). Global patterns in bacterial diversity. *Proc. Natl. Acad. Sci. U. S. A.* 104, 11436–11440.

56. Lozupone, C., and Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* 71, 8228–8235.

57. Malard, L. A., Anwar, M. Z., Jacobsen, C. S., and Pearce, D. A. (2019). Biogeographical patterns in soil bacterial communities across the Arctic region. *FEMS Microbiol. Ecol.* 95. doi:10.1093/femsec/fiz128.

58. Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R., and Peddada, S. D. (2015). Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. Ecol. Health Dis.* 26, 27663.

59. Marshall, I. P. G., Ren, G., Jaussi, M., Lomstein, B. A., Jørgensen, B. B., Røy, H., et al. (2019). Environmental filtering determines family-level structure of sulfate-reducing microbial communities in subsurface marine sediments. *ISME J.* 13, 1920–1932.

60. Martín-Fernández, J. A., Barceló-Vidal, C., and Pawlowsky-Glahn, V. (2003). Dealing with Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation. *Math. Geol.* 35, 253–278.

61. Martino, C., Morton, J. T., Marotz, C. A., Thompson, L. R., Tripathi, A., Knight, R., et al. (2019). A Novel Sparse Compositional Technique Reveals Microbial Perturbations. *mSystems* 4. doi:10.1128/mSystems.00016-19.

62. Martino, C., Shenhav, L., Marotz, C. A., Armstrong, G., McDonald, D., Vázquez-Baeza, Y., et al. (2021). Context-aware dimensionality reduction deconvolutes gut microbial community dynamics. *Nat. Biotechnol.* 39, 165–168.

63. McDonald, D., Clemente, J. C., Kuczynski, J., Rideout, J. R., Stombaugh, J., Wendel, D., et al. (2012). The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *Gigascience* 1. doi:10.1186/2047-217X-1-7.

64. McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Available at: http://arxiv.org/abs/1802.03426 [Accessed November 21, 2021].

65. McMurdie, P. J., and Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 8, e61217.

66. Metcalf, J. L., Song, S. J., Morton, J. T., Weiss, S., Seguin-Orlando, A., Joly, F., et al. (2017). Evaluating the impact of domestication and captivity on the horse gut microbiome. *Sci. Rep.* 7, 15497.

67. Metcalf, J. L., Xu, Z. Z., Weiss, S., Lax, S., Van Treuren, W., Hyde, E. R., et al. (2016). Microbial community assembly and metabolic function during mammalian corpse decomposition. *Science* 351, 158–162.

68. Morton, J. T., Marotz, C., Washburne, A., Silverman, J., Zaramela, L. S., Edlund, A., et al. (2019). Establishing microbial composition measurement standards with reference frames. *Nat. Commun.* 10, 1–11.

69. Morton, J. T., Toran, L., Edlund, A., Metcalf, J. L., Lauber, C., and Knight, R. (2017). Uncovering the Horseshoe Effect in Microbial Analyses. *mSystems* 2. doi:10.1128/mSystems.00166-16.

70. Ocasio, J., Babcock, B., Malawsky, D., Weir, S. J., Loo, L., Simon, J. M., et al. (2019). scRNA-seq in medulloblastoma shows cellular heterogeneity and lineage expansion support resistance to SHH inhibitor therapy. *Nat. Commun.* 10, 1–17.

71. Paliy, O., and Shankar, V. (2016). Application of multivariate statistical techniques in microbial ecology. *Mol. Ecol.* 25, 1032–1057.

72. Parbie, P. K., Mizutani, T., Ishizaka, A., Kawana-Tachikawa, A., Runtuwene, L. R., Seki, S., et al. (2021). Dysbiotic Fecal Microbiome in HIV-1 Infected Individuals in Ghana. *Front. Cell. Infect. Microbiol.* 11, 646467.

73. Pawlowsky-Glahn, V., and Buccianti, A. (2011). *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons.

74. Pérez-Jaramillo, J. E., Carrión, V. J., Bosse, M., Ferrão, L. F. V., de Hollander, M., Garcia, A. A. F., et al. (2017). Linking rhizosphere microbiome composition of wild and domesticated Phaseolus vulgaris to genotypic and root phenotypic traits. *ISME J.* 11, 2244–2257.

75. Pielou, E. C. (1966). The measurement of diversity in different types of biological collections. *J. Theor. Biol.* 13, 131–144.

76. Podani, J., and Miklós, I. (2002). Resemblance Coefficients and the Horseshoe Effect in Principal Coordinates Analysis. *Ecology* 83, 3331–3343. doi:10.1890/0012-9658(2002)083[3331:rcathe]2.0.co;2.

77. Potvin, C., and Roff, D. A. (1993). Distribution-Free and Robust Statistical Methods: Viable Alternatives to Parametric Statistics. *Ecology* 74, 1617–1628. doi:10.2307/1939920.

78. Ren, B., Bacallado, S., Favaro, S., Holmes, S., and Trippa, L. (2017). Bayesian Nonparametric Ordination for the Analysis of Microbial Communities. *J. Am. Stat. Assoc.* 112, 1430–1442.

79. Roweis, S. T., and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* 290. doi:10.1126/science.290.5500.2323.

80. Ruiz-Perez, D., Guan, H., Madhivanan, P., Mathee, K., and Narasimhan, G. (2020). So you think you can PLS-DA? *BMC Bioinformatics* 21, 1–10.

81. Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 7537–7541.

82. Scholkopf, B., Smola, A., and Müller, K.-R. (1999). Kernel principal component analysis. in *ADVANCES IN KERNEL METHODS - SUPPORT VECTOR LEARNING* Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.128.7613 [Accessed November 23, 2021].

83. Shalapour, S., Lin, X.-J., Bastian, I. N., Brain, J., Burt, A. D., Aksenov, A. A., et al. (2017). Inflammation-induced IgA+ cells dismantle anti-liver cancer immunity. *Nature* 551, 340–345.

84. Shankar, V., Agans, R., and Paliy, O. (2017). Advantages of phylogenetic distance based constrained ordination analyses for the examination of microbial communities. *Sci. Rep.* 7, 6481.

85. Shi, Y., Zhang, L., Peterson, C., Do, K.-A., and Jenq, R. (2021). Performance Determinants of Unsupervised Clustering Methods for Microbiome Data. *bioArxiv*. doi:10.1101/2021.04.08.439060.

86. Silverman, J. D., Roche, K., Mukherjee, S., and David, L. A. (2020). Naught all zeros in sequence count data are the same. *Comput. Struct. Biotechnol. J.* 18, 2789–2798.

87. Song, S. J., Amir, A., Metcalf, J. L., Amato, K. R., Xu, Z. Z., Humphrey, G., et al. (2016). Preservation Methods Differ in Fecal Microbiome Stability, Affecting Suitability for Field Studies. *mSystems* 1. doi:10.1128/mSystems.00021-16.

88. Song, S. J., Wang, J., Martino, C., Jiang, L., Thompson, W. K., Shenhav, L., et al. (2021). Naturalization of the microbiota developmental trajectory of Cesarean-born neonates after vaginal seeding. *Med* 2, 951–964.e5. doi:10.1016/j.medj.2021.05.003.

89. Souza, F. F. C., Mathai, P. P., Pauliquevis, T., Balsanelli, E., Pedrosa, F. O., Souza, E. M., et al. (2021). Influence of seasonality on the aerosol microbiome of the Amazon rainforest. *Sci. Total Environ.* 760, 144092.

90. Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., et al. (2015). Ocean plankton. Structure and function of the global ocean microbiome. *Science* 348, 1261359.

91. Taavitsainen, S., Engedal, N., Cao, S., Handle, F., Erickson, A., Prekovic, S., et al. (2021). Single-cell ATAC and RNA sequencing reveal pre-existing and persistent cells associated with prostate cancer relapse. *Nat. Commun.* 12, 1–16.

92. Tabachnick, B. G., and Fidell, L. S. (2013). *Using Multivariate Statistics*.

93. Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* 290. doi:10.1126/science.290.5500.2319.

94. ter Braak, C. J. F. (1985). *Canonical Correspondence Analysis: A New Eigenvector Technique for Multivariate Direct Gradient Analysis*.

95. The Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214.

96. Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., et al. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* 551, 457–463.

97. Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The human microbiome project. *Nature* 449, 804–810.

98. van der Maaten, L., and Hinton, G. (2008). Visualizing Data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

99. Vangay, P., Johnson, A. J., Ward, T. L., Al-Ghalith, G. A., Shields-Cutler, R. R., Hillmann, B. M., et al. (2018). US Immigration Westernizes the Human Gut Microbiome. *Cell* 175, 962–972.e10.

100. Vankadara, L. C., and von Luxburg, U. (2018). Measures of distortion for machine learning. *Adv. Neural Inf. Process. Syst.* 31. Available at: https://proceedings.neurips.cc/paper/2018/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf [Accessed November 20, 2021].

101. Vázquez-Baeza, Y., Gonzalez, A., Smarr, L., McDonald, D., Morton, J. T., Navas-Molina, J. A., et al. (2017). Bringing the Dynamic Microbiome to Life with Animations. *Cell Host Microbe* 21, 7–10.

102. Vázquez-Baeza, Y., Hyde, E. R., Suchodolski, J. S., and Knight, R. (2016). Dog and human inflammatory bowel disease rely on overlapping yet distinct dysbiosis networks. *Nature microbiology* 1. doi:10.1038/nmicrobiol.2016.177.

103. Wattenberg, M., Viégas, F., and Johnson, I. (2016). How to Use t-SNE Effectively. *Distill* 1, e2.

104. Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 5, 1–18.

105. Wong, R. G., Wu, J. R., and Gloor, G. B. (2016). Expanding the UniFrac Toolbox. *PLoS One* 11, e0161196.

106. Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S. A., et al. (2011). Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. *Science* 334, 105.

107. Xu, T., Demmer, R. T., and Li, G. (2021). Zero-inflated Poisson factor model with application to microbiome read counts. *Biometrics* 77, 91–101.

108. Xu, X., Xie, Z., Yang, Z., Li, D., and Xu, X. (2020). A t-SNE Based Classification Approach to Compositional Microbiome Data. *Front. Genet.* 11, 620143.

109. Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., et al. (2012). Human gut microbiome viewed across age and geography. *Nature* 486, 222–227.

110. Young, C., Wood, H. M., Seshadri, R. A., Van Nang, P., Vaccaro, C., Melendez, L. C., et al. (2021). The colorectal cancer-associated faecal microbiome of developing countries resembles that of developed countries. *Genome Med.* 13, 1–13.

Chapter 2. Determination of effect sizes for power analysis for microbiome studies using large microbiome databases

Abstract

Herein, we present a tool called Evident that can be used for deriving effect sizes for a broad spectrum of metadata variables, such as mode of birth, antibiotics, socioeconomics, etc., to provide power calculations for a new study. Evident can be used to mine existing databases of large microbiome studies (such as the American Gut Project, FINRISK, and TEDDY) to analyze the effect sizes for planning future microbiome studies via power analysis. For each metavariable, the Evident software is flexible to compute effect sizes for many commonly used measures of microbiome analyses, including α diversity, β diversity, and log-ratio analysis. In this work, we describe why effect size and power analysis are necessary for computational microbiome analysis and show how Evident can help researchers perform these procedures. Additionally, we describe how Evident is easy for researchers to use and provide an example of efficient analyses using a dataset of thousands of samples and dozens of metadata categories.

2.1. Introduction

Power analysis for a univariate (or multivariate) outcome variable is not new. Numerous statistical packages are available (e.g., SAS) for a variety of experimental designs and outcome variables. For a given level of significance, a common challenge with any power analysis is the understanding of the underlying variability in the data and the value of the parameter of interest in the alternative hypothesis. Once the statistical parameter of interest is identified, researchers often conduct a pilot study to estimate mean differences and standard deviations and use these values, termed effect sizes, as the basis for conducting power analysis, i.e., sample size

calculations for the larger study proposed in their research program. The larger the effect size, the stronger the statistical difference, and the fewer samples are needed for high statistical power.

This type of power analysis is important because of the limited resources available for experimental designs. Ensuring that researchers do not spend more resources than required to achieve a given statistical power is paramount. The problem is more complicated when it comes to microbiome studies because there are a variety of parameters one can base their designs on. Almost all parameters of interest, such as measures of α or β diversity are (nonlinear) functions of relative abundances of various taxa. Estimation of relative abundances using small pilot studies (i.e., $N < 100$) is not always satisfactory because the observed count data contain a large number of zeros. The preliminary estimates from a pilot study are potentially subject to large bias and uncertainties. Consequently, the determination of the effect size for a given parameter, say α diversity defined by Shannon's entropy, is a difficult task. This article takes the first step towards addressing this challenging problem by making use of the recently created large databases such as the American Gut Project, TEDDY, and FINRISK. These are very rich databases that continue to grow. They contain microbiome data on several thousands of individuals along with hundreds of commonly measured metadata and thousands of represented taxa. For each variable in the metadata, say, mode of birth, the user-friendly software Evident derives the effect size for a parameter of a researcher's interest, such as Shannon's entropy. Using this parameter, a researcher can then conduct a simulation study to derive power functions for different sample sizes.

Since microbiome datasets such as AGP, TEDDY, and FINRISK are very large and contain a large number of metavariables, we expect Evident to be a useful tool for deriving effect sizes for variables of common interest. Importantly, Evident takes user-inputted study data for the generation of results, so researchers can customize their analyses as they see fit. As new databases get constructed, Evident will access those to derive better and more refined effect size

estimates that will be useful for planning microbiome studies. Evident is available both as a standalone Python package as well as a QIIME 2 plugin [8]. Currently, effect size analysis and power analysis for microbiome science can be performed using programming languages such as Python and R. However, these approaches are not designed for use with many metavariables. As a result, researchers must write custom code to iterate through the full dataset. With Evident, researchers can seamlessly explore the effect size of community differences in dozens of metadata columns at once and easily perform power analysis. The interactive component of Evident additionally makes this process easy to use and share. This scalability, flexibility, user-friendliness, and integration with existing microbiome software make Evident easier to slot into existing microbiome workflows over existing methods.

2.2. Materials and Methods

Figure 2.1a shows an overview of the Evident workflow. As input, Evident takes a sample metadata file and a data file of interest (for example, α diversity). In this cartoon example, we show the main Evident workflow is (1) calculating effect size for a metavariable of interest between two or more groups (2) performing parametric power analysis on varying sample sizes, levels of significance, and/or effect sizes (3) plotting the accompanying power curve(s). Both univariate per-sample data (such as α diversity) and multivariate data (as a distance matrix such as β diversity) are supported. For univariate measures, the differences in means among groups are considered. For multivariate measures, the difference in means among within-group pairwise distances is considered. We also note that, at the moment, Evident implements effect size computations of univariable analyses (without explicit handling of confounders) following the approach of existing work [10–13].

Figure 2.1: Evident workflow and interactive visualizations. a, Graphical overview of Evident usage. Sample metadata with categorical groups are used to determine differences among samples. Effect size calculation can be performed and used to generate power curves (in this example using classification status from Casals-Pascual et al., 2020) at multiple statistical significance levels and sample sizes. **b,c** Screenshots of interactive webpage for dynamic exploration of effect sizes and power analysis. Summarized effect sizes of all columns can be used to inform interactive power analysis on multiple groups (b). The underlying grouped data can be visualized with boxplots and, optionally, the raw data as scatter plots (c). Data shown is from McClorry et al. (Qiita study ID: 11402)9.

a

Sample Metadata

Sample	Sex	Diet
S1	M	ND
S2	F	ND
S3	F	HFD
S4	M	HFD

+

Alpha Diversity

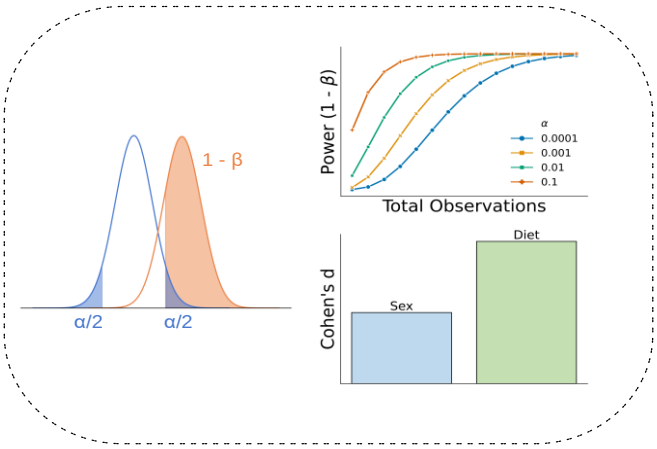
Sample	Faith's PD
S1	2.56
S2	1.14
S3	3.75
S4	1.96

or

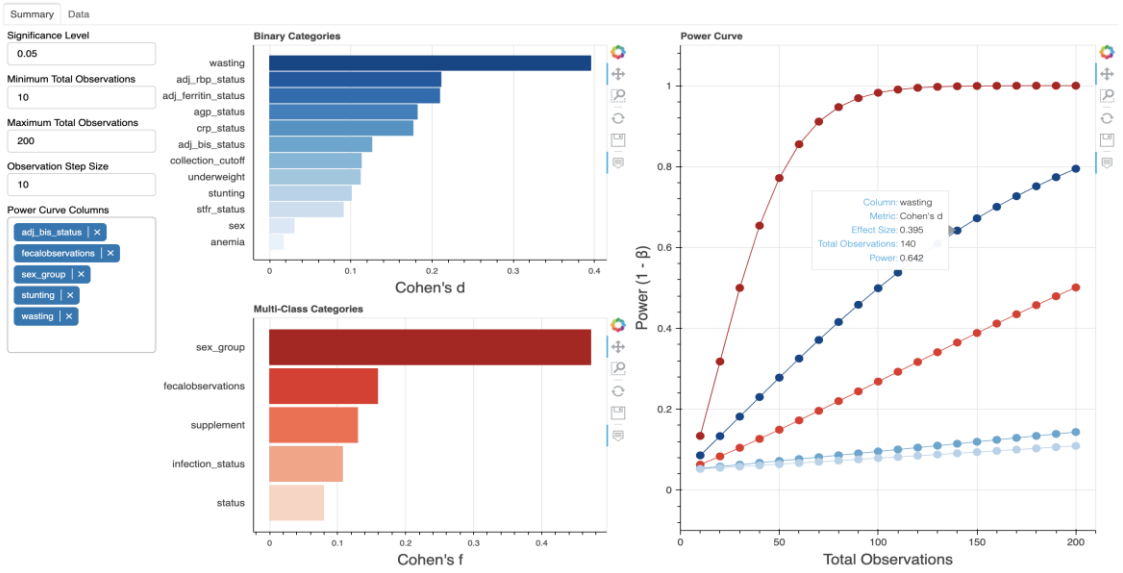
Beta Diversity

	S1	S2	S3	S4
S1				
S2				
S3				
S4				

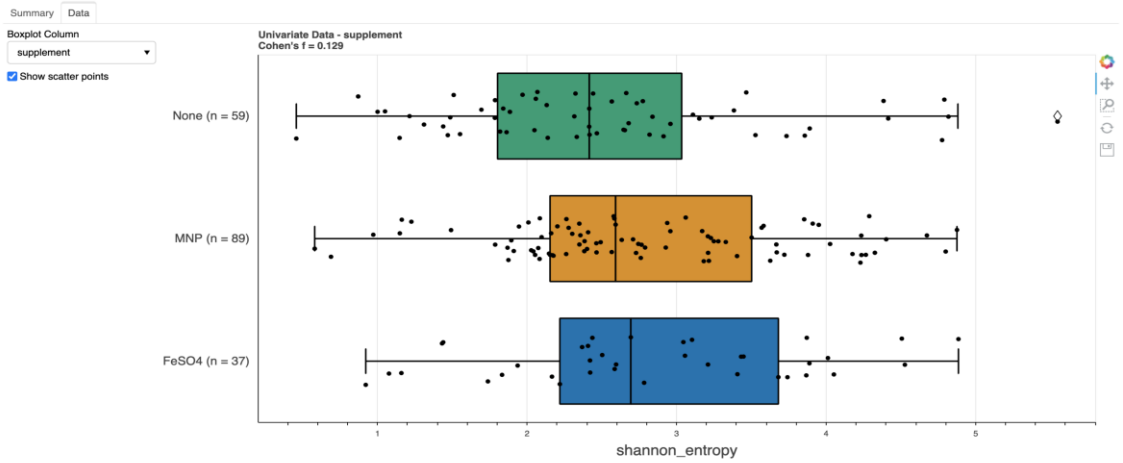
Evident



b



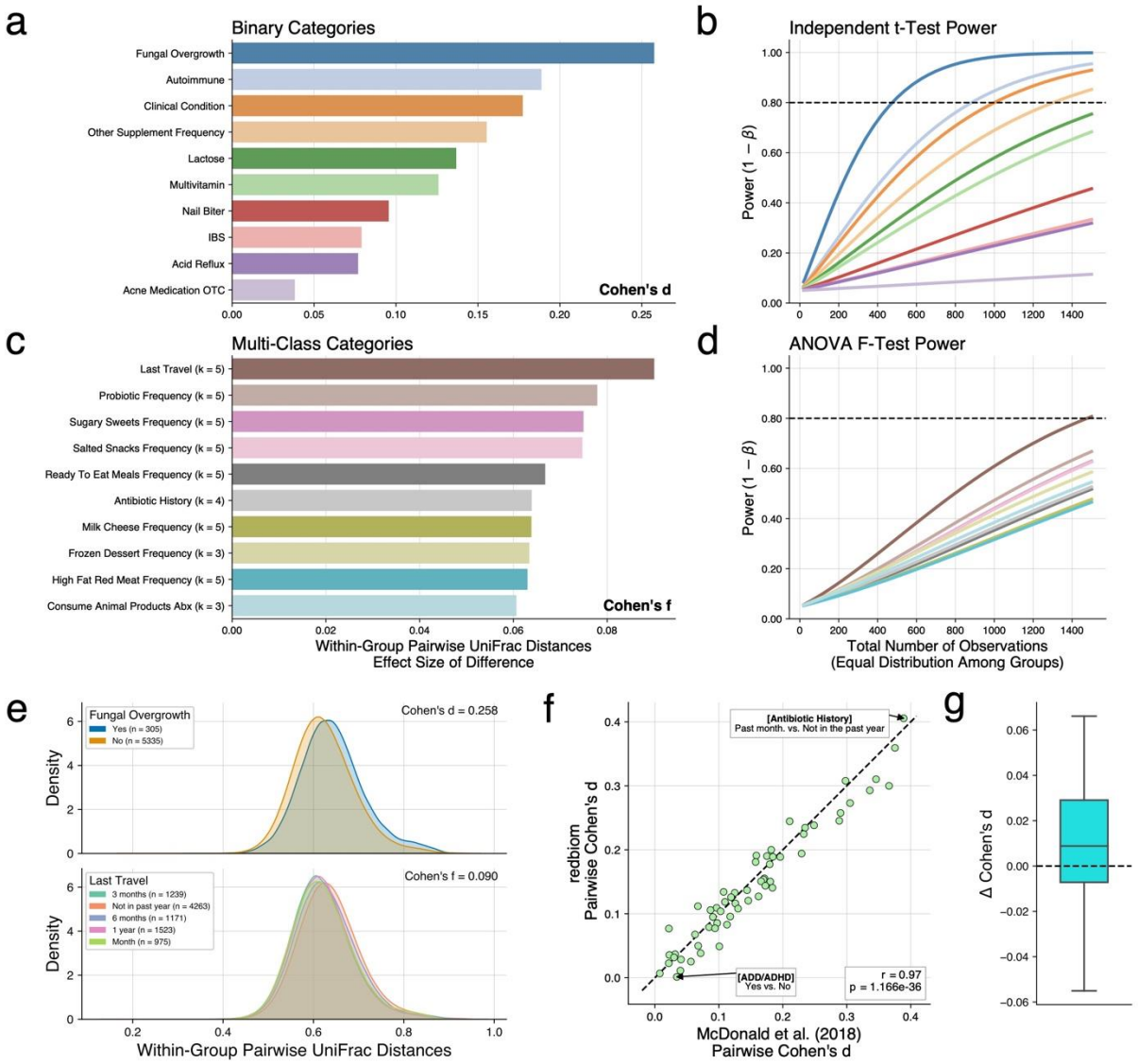
c



Evident supports both binary categories and multi-class categories. For binary categories, Cohen's d is calculated between the two levels. For multi-class categories, Cohen's f is calculated among the levels [15]. Users also have the option of performing pairwise effect size calculations between levels of a multi-class category rather than comparing all groups together. Effect size calculations can be performed on multiple categories at once with simple parallelization by providing the number of CPUs to use. For example, this architecture allows us to decrease the runtime of effect size calculations for 9495 samples comprising 61 categories from over 12 min to 3.5 min using 8 CPUs in parallel.

Evident also provides an interactive component by which users can dynamically explore sample groupings. In Figure 2.1b,c, we show a screenshot of a web app that users can access with Evident. Metadata categories are pre-sorted by effect size, allowing efficient determination of interesting categories. Power analysis is implemented dynamically—multiple categories can be visualized simultaneously for a specified significance level and number of observations. Researchers can look at the "elbow" of the power curves to determine an optimal number of samples to achieve the desired statistical power for their experiments.

Figure 2.2: Analysis of American Gut Project data. a) Top 10 binary categories by group-wise effect size. b) Two-sample independent t-test power analysis of selected binary category effect sizes for significance level of 0.05. c) Top 10 multi-class categories by group-wise effect size. d) One-way ANOVA F-test power analysis of selected multi-class category effect sizes at significance level of 0.05. e) Distributions of within-group pairwise UniFrac distances for highest effect size binary category (top) and multi-class category (bottom). f) Comparison of pairwise effect sizes between reprocessed data from redbiom and published effect sizes from McDonald et al. Reprocessing results are not identical due to inherent randomness in rarefaction. g) Boxplot of differences in effect sizes between published and reprocessed effect sizes.


a
Binary Categories
Fungal Overgrowth
Autoimmune
Clinical Condition
Other Supplement Frequency
Lactose
Multivitamin
Nail Biter
IBS
Acid Reflux
Acne Medication OTC
Cohen's d
0.00
0.05
0.10
0.15
0.20
0.25
b
Independent t-Test Power
Power (1 − β)
0.00
0.20
0.40
0.60
0.80
1.00
0
200
400
600
800
1000
1200
1400
c
Multi-Class Categories
Last Travel (k = 5)
Probiotic Frequency (k = 5)
Sugary Sweets Frequency (k = 5)
Salted Snacks Frequency (k = 5)
Ready To Eat Meals Frequency (k = 5)
Antibiotic History (k = 4)
Milk Cheese Frequency (k = 5)
Frozen Dessert Frequency (k = 3)
High Fat Red Meat Frequency (k = 5)
Consume Animal Products Abx (k = 3)
Cohen's f
0.00
0.02
0.04
0.06
0.08
Within-Group Pairwise UniFrac Distances
Effect Size of Difference
d
ANOVA F-Test Power
Power (1 − β)
0.00
0.20
0.40
0.60
0.80
1.00
0
200
400
600
800
1000
1200
1400
Total Number of Observations
(Equal Distribution Among Groups)
e
Fungal Overgrowth
Yes (n = 305)
No (n = 5335)
Cohen's d = 0.258
Density
0
2
4
6
Last Travel
3 months (n = 1239)
Not in past year (n = 4263)
6 months (n = 1171)
1 year (n = 1523)
Month (n = 975)
Cohen's f = 0.090
Density
0
2
4
6
0.2
0.4
0.6
0.8
1.0
Within-Group Pairwise UniFrac Distances
f
[Antibiotic History]
Past month. vs. Not in the past year
[ADD/ADHD]
Yes vs. No
r = 0.97
p = 1.166e-36
redbiom
Pairwise Cohen's d
0.0
0.1
0.2
0.3
0.4
McDonald et al. (2018)
Pairwise Cohen's d
g
Δ Cohen's d
−0.06
−0.04
−0.02
0.00
0.02
0.04
0.06

Statistical Methodology

Let $X_1, X_2, \ldots, X_l$ denote *l* metavariables available in some database. Without loss of generality, in the following, we shall describe the methodology used in Evident for X_1. For simplicity of exposition, we shall drop the subscript 1 from X_1. Furthermore, to fix ideas of the methodology and simplicity of exposition, we shall assume X is binary, such as mode of delivery. The outcome variable is denoted by Y, such as Shannon entropy, a measure of α diversity of an infant's gut microbiome. The relative abundance of the j^{th} taxon, $j = 1, 2, \ldots, q$, in the k^{th} infant belonging to the $X = i^{th}$ group $i = 1, 2, \ldots, G$, (e.g., mode of delivery), is denoted by p_{ijk}. For example, $X = 1$ represents babies born vaginally, and $X = 2$ represents babies born by C-Section. We assume that there are q taxa measured on each infant (some may be zeros) and there are N_i infants in the i^{th} the group in the large database. Thus, the Shannon entropy for the k^{th} subject belonging to the $X = i^{th}$ group is given by $Y_{ik} = -\sum_{j=1}^{q} p_{ijk} \ln(p_{ijk})$. In this definition, $p_{ijk} \ln(p_{ijk}) \to 0, \text{as } p_{ijk} \to 0$. Since we are working with very large databases, such as AGP, we assume that each n_i is sufficiently large.

The Evident methodology for determining the effect size needed for conducting power analysis and sample size calculations for a future infant gut microbiome study using Shannon entropy to describe microbial diversity is described in the following steps.

Step 1 (Average population diversity): For each value of X, for each subject in the database, using the available microbiome data, compute the desired parameter of interest, for example, the average Shannon entropy for α diversity, $\mu_i = -(\frac{1}{N_i}) \sum_{k=1}^{N_i} Y_{ik}, i = 1, 2$. As noted above, we assume that each N_i is sufficiently large so that μ_i represents the average Shannon entropy, for the i^{th} the population of infants.

Step 2 (Variance of population diversity): Similar to the population mean μ_i, for each $i = 1, 2, \ldots G$ we compute the population variance of α diversity, denoted by $\sigma_i^2 = \left(\frac{1}{N_i}\right) \sum_{k=1}^{N_i} (Y_{ik} -$

$\mu_i)^2$. Again, each n_i is sufficiently large so that σ_i^2 represents the population variance of Shannon entropy for the i^{th} population of infants. Under the simplifying assumption of homoscedasticity (i.e., all populations have same the variance), we average the two empirical variances to obtain the pooled variance, i.e., $\sigma_{pool}^2 = \frac{\sum_{i=1}^{G} N_i \sigma_i^2}{\sum_{i=1}^{G} N_i}$. Again, since the sample sizes are large, we regard the pooled variance as the true population variance for all our calculations.

Step 3 (Effect size calculations): Assuming that the outcome variable of interest (e.g., α diversity) is normally distributed, we have the following formulas for effect sizes using non-central distribution for the test statistic (for $G = 2$) or non-central F distribution (for $G \geq 2$), respectively:

$$d = \frac{\mu_1 - \mu_2}{\sigma_{pool}}.$$

$$f = \frac{\sum_{i=1}^{G} \left(\frac{N_i}{N}\right)(\mu_i - \bar{\mu})^2}{\sigma_{pool}^2}, \bar{\mu} = \sum_{i=1}^{G} \frac{N_i}{N} \mu_i, N = \sum_{i=1}^{G} N_i.$$

Although equal variances across groups may be an unreasonable assumption, it is a simplifying assumption.

Step 4 (Power and sample size calculations): For a future study, suppose a researcher has a budget for a sample size m_i, $i = 1, 2$, for the i^{th} population of infants, then for a level of significance of α, the power corresponding to the effect size d and sample m_i, can be calculated parametrically, assuming $Y_{ik} \sim^{iid} N(\mu_i, \sigma_{pool})$.

Under the normality assumption, for $G = 2$, Evident calculates power using non-central t-distribution using the effect size parameter d and different choices of samples sizes m_1, m_2. In the case $G > 2$, it uses non-central F distribution with effect size parameter f and different choices of samples sizes $m_1, m_2, \ldots, m_G$.

Interactive Exploration of Community Differences

The interactive visualization provided in Evident is created with Bokeh. Given microbiome data and sample metadata, Evident creates a Bokeh app that dynamically calculates effect sizes and power analysis for the chosen parameters. This view also shows the raw data values as boxplots with optional scatter points.

Analysis of AGP Data

A sample ID list was generated from the original distance matrix used in the AGP study. 100 nucleotide 16S rRNA gene amplicon (16S) data targeting the V4 hypervariable region for these samples were downloaded from the AGP study on Qiita (study ID: 10317) using redbiom [28,29]. Both preparation and sample metadata were also retrieved with redbiom. Due to multiple preparations containing data from some samples, we performed disambiguation by keeping the samples with the highest sequencing depth.

We then processed the feature table and metadata according to the original study. The original workflow used the default parameters in Deblur to remove features with fewer than 10 occurrences in the data [30]. Because Qiita does not perform this filtering by default, we performed this filtering manually. To remove sequences associated with sample bloom, we performed bloom filtering [31]. We then rarefied the feature table to 1250 sequences as in the original analysis.

We processed the sample metadata in accordance with the original study. Because of differences in self-reporting protocols from 2018, metadata categories associated with reported Vioscreen responses as well as those associated with alcohol consumption were removed. The following categories were removed due to mismatches in sample metadata: roommates, allergies, age_cat, bmi_cat, longitude, latitude, elevation, height_cm, collection_time, and center_project_name. Only the top four annotated countries were considered—US, UK, Australia,

and Canada. All other countries were ignored. Overall, 61 metadata categories common to both the original data and redbiom data were used for further analysis.

Sequences from the feature table were placed into a 99% Greengenes [32] insertion reference tree using SEPP [33]. We then used unweighted UniFrac to generate a sample-by-sample distance matrix [34]. This distance matrix was used as input to Evident along with the disambiguated, processed sample metadata.

We used effect_size_by_category to calculate the whole-group effect sizes for each column in the metadata and pairwise_effect_size_by_category to calculate the group-pairwise effect sizes for multi-class categories. For each whole-group effect size, we computed a power analysis for α values of 0.01, 0.05, and 0.1. Power was calculated on total sample size values from 20 to 1500 in increments of 40 samples. Evident analyses were performed in parallel on a high-performance computing environment. Group-wise and pairwise effect size calculations both took under 4 min for 82 metadata categories on 9495 samples using 8 CPUs (we note the AGP paper used n = 9511 but operated at 125 nt; we observe a slightly reduced number of samples at 100 nt). We also benchmarked group-wise effect size calculations using only a single CPU as a comparison; this process took 12.4 min, meaning the parallelization decreased runtime by approximately 3.5×. Power analysis calculation took 2.7 min for 82 categories using 8 CPUs in parallel.

Analysis of Study of Latinos Data

We downloaded closed-reference (picked against Greengenes 97%) 16S-V4 fecal data from Qiita (study ID: 11666) using redbiom. We used the bmi_v2 column to separate samples into two groups: normal (BMI < 25) and obese (BMI > 40). For each sample, we summed the abundance of *Prevotella* spp. and *Bacteroides* spp. adding a pseudocount to both sums. We then calculated the (log) ratio of the *Prevotella* sum and *Bacteroides* sum.

For power analysis, we first established the “true” difference between the obese and normal samples as 1.06 (d = 0.27). We used the log-ratio data to determine three levels of effect sizes we wanted to evaluate: 0.5 (d = 0.13), 1.0 (d = 0.25), and 1.5 (d = 0.38). To convert the differences into desired effect sizes, we divided each difference by the pooled standard deviation of the original log ratios. We used Evident to compute the power at each of these effect sizes for a significance threshold of 0.05 for total observations varying between 100 and 1000.

2.3. Results

As a demonstration of Evident, we reprocessed 9495 samples from the AGP to compare the published effect sizes in McDonald 2018 with those from a new analysis with Evident [4]. We downloaded the same samples from the original paper and reprocessed the data and metadata in the same manner, focusing on within-group UniFrac [16] distances. First, we computed the group-wise effect sizes for all valid metadata categories. The top ten binary categories and multi-class effect sizes are shown in Figure 2.2a,c, respectively. Using these effect sizes, we performed power analyses for each category at a significance level of 0.05 for a range of sample sizes from 20 to 1500 (Figure 2.2b,d). We plotted the distribution of the highest effect size binary and multi-class categories as reported by our new analysis in Figure 2.2e. Finally, we computed the pairwise effect sizes as performed in the original paper to verify that Evident returns the same values. Figure 2f shows that the effect sizes map extremely closely between the published data and the newly reprocessed data. The values of effect size differences in Figure 2.2g are distributed around 0, indicating that there is very little difference between effect size calculations. This serves as validation that Evident returns the correct effect sizes. We note, however, that the data used in this study is very heterogeneous—coming from multiple countries. It is important to make sure the data used in computing effect sizes are specific to the biological questions of interest. In Supplementary Figure S1, we plot the effect sizes calculated from only US samples and only UK

samples. These effect sizes have a weak correlation (Spearman rho = 0.54), suggesting that country is a strong factor for effect sizes between biological groups. As such, researchers may want to perform further pre-processing such as stratification of data by country and computing the individual effect sizes for each population. We believe more work should be done on evaluating these differences in relation to these heterogenous populations to ensure results are not artificially inflated or deflated.

While we focus on diversity measures in this work, Evident is also usable with any other data such as log ratios of microbial abundances. As an example, we use Evident to extend the work of Morton et al. [17] and Fedarko et al. [18] in using log ratios for, e.g., post-hoc differential abundance analysis. We analyzed the commonly reported (log) ratio of *Prevotella* to *Bacteroides* in the Study of Latinos (SoL) cohort [19]. In Supplementary Figure S2, we plot the log ratio differences between subjects with a BMI < 25 and subjects with a BMI > 40. We also plot a power curve with custom differences in means, showing Evident's flexibility in designing experiments with specific effect sizes in mind.

2.4. Conclusion

It is important for researchers to keep effect sizes in mind when performing computational microbiome analysis. Calculating and reporting effect sizes make it easier for researchers to determine the magnitude of biological effects on microbial communities. Additionally, these effect sizes can be used to inform power analyses for the efficient allocation of resources for new studies. We designed Evident for researchers to easily mine and process existing datasets for this information. Evident can slot into existing microbiome workflows and process numerous metadata categories efficiently and quickly, allowing its application to a broad range of microbiome research questions.

We note that the choice of study used in Evident should be carefully considered when designing and planning new experiments. For example, an existing study using 16S sequencing may not be completely appropriate when planning shotgun metagenomics experiments, or even experiments that will use a different primer set to target a different region of the 16S rRNA gene, because the different methods may recapture different bacteria with different efficiencies and therefore the effect size of the same per-subject or per-sample variable may differ depending on the methodology. More work should be done to evaluate the differences in downstream analyses on samples between 16S and shotgun metagenomics data. Similarly, culture-based microbiome studies may not follow the same statistical properties as NGS data. Researchers should be mindful of these differences when using Evident. Additionally, researchers should be aware of the limitations of the statistical methodology of Evident. For example, if the assumptions of variance homogeneity are not held, the obtained effect sizes will be inaccurate and the subsequent power analyses can overestimate or underestimate the number of samples required to achieve a given level of statistical power. Similarly, the assumption of equal group sizes in proposed experimental designs from power analysis may be naïve in practice. For rare diseases or phenotypes, it may not be feasible to design an experiment in which all groups have the same number of samples. In these cases, performing simulations with unequal group sizes to determine the necessary sample size to be likely to achieve a statistically significant result may be informative.

We encourage microbiome researchers to incorporate Evident into their workflows for both reporting effect sizes of microbial community differences and planning experimental designs. In the future, we hope to enhance flexibility by including quantitative metadata categories (rather than the current qualitative categories) and unbalanced group sample size power analyses.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/genes14061239/s1, Figure S1: Comparison of effect sizes between UK and US samples; Table S1: Log-ratio analysis on Study of Latinos cohort.

Author Contributions: G.R., A.G., D.M., and R.K. conceived the idea for the software and study. G.R., A.G., D.M., Y.V.-B., and L.J. developed the software. G.R., D.M., and A.G. conducted the analysis of the AGP data. B.N. assisted with AGP metadata curation. A.H.D., Y.V.-B., D.M., and D.H. reviewed the software code and provided valuable feedback and bug reports. C.C.-P., L.J., and S.P. contributed to the original code for effect size and power calculation. All authors have read and agreed to the published version of the manuscript.

2.5. Code availability

The latest version of Evident is available at https://github.com/biocore/evident under the BSD-3 license. Evident is installable from PyPI both as a standalone Python 3 package and a QIIME 2 plugin. The scripts used to download and analyze AGP data as well as the processed Evident results are available at https://github.com/knightlab-analyses/evident-analyses. Analysis of AGP data in this study was performed with Evident version 0.4.0.

2.6. Data availability

Data for the demonstration in Figure S1 were downloaded from Qiita (study ID: 11402)[9] at 90 nucleotides using the deblur[30] pipeline. AGP data were downloaded from Qiita (study ID: 10317) using redbiom with context "Deblur_2021.09-Illumina-16S-V4-100nt-50b3a2". The original pairwise effect sizes, sample metadata, and unweighted UniFrac distance matrix were downloaded from the original McDonald et al. study for comparison. SoL data used in Supplemental Figure AB.1.2 were downloaded from Qiita (study ID: 11666) using redbiom with context "Pick_closed-reference_OTUs-Greengenes-Illumina-16S-V4-90nt-44feac."

2.7. Contributions

G.R., A.G, D.M., & R.K. conceived the idea for the software and study. G.R., A.G., D.M., Y.V.B & L.J. developed the software. G.R., D.M., & A.G. conducted the analysis of the AGP data. B.N. assisted with AGP metadata curation. A.H.D., Y.V.B., D.M., & D.H. reviewed the software code and provided valuable feedback and bug reports. C.C., L.J., & S.P. contributed to the original code for effect size and power calculation. All authors contributed to and reviewed the final manuscript.

2.8 Acknowledgements

We would like to thank the members of the Knight Lab for feedback on the scope and details of Evident. We thank Jamie Morton for valuable discussions about effect size. This work was supported in part by the Alfred P. Sloan foundation (G-2017-9838), NIH-NIDDK (P01DK078669), NIH-NCI (U24CA248454), and NIH (1DP1AT010885, U19AG063744). Research of S.P. was supported [in part] by funding from NIEHS intramural program ZIA

ES103389-01. The authors thank Drs. Huang Lin (NIEHS) and Mikyeong Lee (NIEHS) for their valuable comments that improved the presentation of this article.

Chapter 2 has been submitted for publication of the materials as it may appear in *Genes*, “Determination of effect sizes for power analysis for microbiome studies using large microbiome databases.” Gibraan Rahman, Daniel McDonald, Antonio Gonzalez, Yoshiki Vázquez-Baeza, Lingjing Jiang, Climent Casals-Pascual, Shyamal Peddada, Daniel Hakim, Amanda Hazel Dilmore, Brent Nowinski, and Rob Knight. The dissertation author was the primary investigator and first author of this paper.

2.9 References

1. Sullivan, G.M.; Feinn, R. Using Effect Size—Or Why the P Value Is Not Enough. J. Grad. Med. Educ. 2012, 4, 279–282.

2. Baguley, T. Standardized or simple effect size: What should be reported? Br. J. Psychol. 2009, 100, 603–617.

3. Cohen, J. Statistical Power Analysis. Curr. Dir. Psychol. Sci. 1992, 1, 98–101.

4. McDonald, D.; Hyde, E.; Debelius, J.W.; Morton, J.T.; Gonzalez, A.; Ackermann, G.; Alexander, A. American Gut: An Open Platform for Citizen Science Microbiome Research. mSystems 2018, 3, e00031-18.

5. TEDDY Study Group. The Environmental Determinants of Diabetes in the Young (TEDDY) Study. Ann. N. Y. Acad. Sci. 2008, 1150, 1–13.

6. Vartiainen, E.; Jousilahti, P.; Alfthan, G.; Sundvall, J.; Pietinen, P.; Puska, P. Cardiovascular risk factor changes in Finland, 1972–1997. Int. J. Epidemiol. 2000, 29, 49–56.

7. Casals-Pascual, C.; González, A.; Vázquez-Baeza, Y.; Song, S.J.; Jiang, L.; Knight, R. Microbial Diversity in Clinical Microbiome Studies: Sample Size and Statistical Power Considerations. Gastroenterology 2020, 158, 1524–1528.

8. Bolyen, E.; Rideout, J.R.; Dillon, M.R.; Bokulich, N.A.; Abnet, C.C.; Al-Ghalith, G.A.; Alexander, H.; Alm, E.J.; Arumugam, M.; Asnicar, F.; et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. 2019, 37, 852–857.

9. McClorry, S.; Zavaleta, N.; Llanos, A.; Casapía, M.; Lönnerdal, B.; Slupsky, C.M. Anemia in infancy is associated with alterations in systemic metabolism and microbial structure and function in a sex-specific manner: An observational study. Am. J. Clin. Nutr. 2018, 108, 1238–1248.

10. Yang, L.; Chen, J. A comprehensive evaluation of microbial differential abundance analysis methods: Current status and potential solutions. Microbiome 2022, 10, 130.

11. Dwiyanto, J.; Hussain, M.H.; Reidpath, D.; Ong, K.S.; Qasim, A. ;Lee, S.W.H.; Lee, S.M.; Foo, S.C.; Chong, C.W.; Rahman, S. Ethnicity influences the gut microbiota of individuals sharing a geographical location: A cross-sectional study from a middle-income country. Sci. Rep. 2021, 11, 2618.

12. Park, J.; Kato, K.; Murakami, H.; Hosomi, K.; Tanisawa, K.; Nakagata, T.; Ohno, H.; Konishi, K.; Kawashima, H.; Chen, Y.-A.; Mohsen, A.; et al. Comprehensive analysis of gut microbiota of a healthy population and covariates affecting microbial variation in two large Japanese cohorts. BMC Microbiol. 2021, 21, 151.

13. Falony, G.; Joossens, M.; Vieira-Silva, S.; Wang, J.; Darzi, Y.; Faust, K.; Kurilshikov, A.; Bonder, M.J.; Valles-Colomer, M.; Vandeputte, D.; et al. Population-level analysis of gut microbiome variation. Science 2016, 352, 560–564.

14. Kirby, K.N.; Gerlanc, D. Finding Bootstrap Confidence Intervals for Effect Sizes with BootES. APS Obs. 2017, 30.

15. Cohen, J. Statistical Power Analysis for the Behavioral Sciences; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 1988.

16. McDonald, D.; Vázquez-Baeza, Y.; Koslicki, D.; McClelland, J.; Reeve, N.; Xu, Z.; Gonzalez, A.; Knight, R. Striped UniFrac: Enabling microbiome analysis at unprecedented scale. Nat. Methods 2018, 15, 847–848.

17. Morton, J.T.; Marotz, C.; Washburne, A.; Silverman, J.; Zaramela, L.S.; Edlund, A.; Zengler, K.; Knight, R. Establishing microbial composition measurement standards with reference frames. Nat. Commun. 2019, 10, 2719.

18. Fedarko, M.W.; Martino, C.; Morton, J.T.; González, A.; Rahman, G.; A Marotz, C.; Minich, J.J.; E Allen, E.; Knight, R. Visualizing'omic feature rankings and log-ratios using Qurro. NAR Genom. Bioinform. 2020, 2, lqaa023.

19. Kaplan, R.C.; Wang, Z.; Usyk, M.; Sotres-Alvarez, D.; Daviglus, M.L.; Schneiderman, N.; Talavera, G.A.; Gellman, M.D.; Thyagarajan, B.; Moon, J.-Y.; et al. Gut microbiome composition in the Hispanic Community Health Study/Study of Latinos is shaped by geographic relocation, environmental factors, and obesity. Genome Biol. 2019, 20, 219.

20. McKinney, W. Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference, Austin, TX, USA, 28 June–3 July 2010; Volume 6.

21. Harris, C.R.; Millman, K.J.; van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array programming with NumPy. Nature 2020, 585, 357–362.

22. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nat. Methods 2020, 17, 261–272.

23. Li, J.C.-H. Effect size measures in a two-independent-samples case with nonnormal and nonhomogeneous data. Behav. Res. Methods 2016, 48, 1560–1574.

24. Seabold, S.; Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. In Proceedings of the 9th Python in Science Conference, Austin, TX, USA, 28 June–3 July 2010; pp. 92–96. https://doi.org/10.25080/Majora-92bf1922-011.

25. Quené, H.; van den Bergh, H. On multi-level modeling of data from repeated measures designs: A tutorial. Speech Commun. 2004, 43, 103–121.

26. Guo, Y.; Logan, H.L.; Glueck, D.H.; Muller, K.E. Selecting a sample size for studies with repeated measures. BMC Med. Res. Methodol. 2013, 13, 100.

27. Vonesh, E.F.; Schork, M.A. Sample Sizes in the Multivariate Analysis of Repeated Measurements. Biometrics 1986, 42, 601–610.

28. Gonzalez, A.; Navas-Molina, J.A.; Kosciolek, T.; McDonald, D.; Vázquez-Baeza, Y.; Ackermann, G.; Dereus, J.; Janssen, S.; Swafford, A.D.; Orchanian, S.B.; et al. Qiita: Rapid, web-enabled microbiome meta-analysis. Nat. Methods 2018, 15, 796–798.

29. McDonald, D.; Kaehler, B.; Gonzalez, A.; DeReus, J.; Ackermann, G.; Marotz, C.; Huttley, G.; Knight, R. redbiom: A Rapid Sample Discovery and Feature Characterization System. mSystems 2019, 4, e00215-19. https://doi.org/10.1128/mSystems.00215-19.

30. Amir, A.; McDonald, D.; Navas-Molina, J.A.; Kopylova, E.; Morton, J.T.; Zech Xu, Z.; Kightley, E.P.; Thompson, L.R.; Hyde, E.R.; Gonzalez, A.; et al. Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. mSystems 2017, 2, e00191-16.

31. Amir, A.; McDonald, D.; Navas-Molina, J.A.; Debelius, J.; Morton, J.T.; Hyde, E.; Robbins-Pianka, A.; Knight, R. Correcting for Microbial Blooms in Fecal Samples during Room-Temperature Shipping. mSystems 2017, 2, e00199-16.

32. McDonald, D.; Price, M.N.; Goodrich, J.; Nawrocki, E.P.; DeSantis, T.Z.; Probst, A.; Andersen, G.L.; Knight, R.; Hugenholtz, P. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. 2012, 6, 610–618.

33. Mirarab, S.; Nguyen, N.; Warnow, T. SEPP: SATé-enabled phylogenetic placement. Pac. Symp. Biocomput. Pac. Symp. Biocomput. 2012, 2011, 247–258. https://doi.org/10.1142/9789814366496_0024.

34. Lozupone, C.; Knight, R. UniFrac: A New Phylogenetic Method for Comparing Microbial Communities. Appl. Environ. Microbiol. 2005, 71, 8228–8235.

Chapter 3. Paired microbiome and metabolome analyses associate bile acid changes with colorectal cancer progression

Abstract

In most cases of sporadic colorectal cancers (CRC), tumorigenesis is a multistep process driven by genomic alterations in concert with dietary influences. In addition, mounting evidence has implicated the gut microbiome as an effector in the development and progression of CRC. While large meta-analyses have provided mechanistic insight into disease progression in CRC patients, study heterogeneity has limited causal associations. To address this limitation, multi-omics studies on genetically controlled cohorts of mice were performed to distinguish genetic and dietary influences. Diet was identified as the major driver of microbial and metabolomic differences, with reductions in alpha diversity and widespread changes in cecal metabolites seen in HFD-fed mice. Similarly, the levels of non-classic amino acid conjugated forms of the bile acid cholic acid (AA-CAs) increased with HFD. We show that these AA-CAs signal through the nuclear receptor FXR and membrane receptor TGR5 to functionally impact intestinal stem cell growth. In addition, the poor intestinal permeability of these AA-CAs supports their localization in the gut. Moreover, two cryptic microbial strains, Ileibacterium valens and Ruminococcus gnavus, were shown to have the capacity to synthesize these AA-CAs. This multi-omics dataset from CRC mouse models supports diet-induced shifts in the microbiome and metabolome in disease progression with potential utility in directing future diagnostic and therapeutic developments.

3.1. Introduction

Colorectal cancer (CRC) is the 4th leading cause of cancer related deaths worldwide(Ferlay et al., 2019). Combined with an expected increase in the incidence in the coming decades, new diagnostic and therapeutic approaches for combating this disease are needed. Diet and lifestyle choices have been identified as risk factors for CRC, with ~50-60% of US cases attributed to modifiable risk factors (Islami et al., 2018). However, the convergence of environmental and genetic factors in the development and progression of CRC is not fully understood.

The intestinal microbiome has been suggested to mediate environmental risk factors in CRC. While specific microbes have been associated with different tumor stages, conflicting reports have led to a meta-analysis approach to map gut microbiome signatures associated with CRC (Feng et al., 2015) (Nakatsu et al., 2015; Yachida et al., 2019) (Song and Chan, 2019) (Scott et al., 2019; Wirbel et al., 2019). Such meta-analysis approaches have identified diagnostic microbial signatures, however causal associations of microorganisms with carcinogenesis have proven difficult (Wirbel *et al.*, 2019). This is attributable in part, to variations in human genetics and environmental conditions. Indeed, study heterogeneity was found to have a larger impact on the composition of the gut microbiome than CRC (Wirbel *et al.*, 2019).

Diets high in animal fat alter the microbiome, as well as lead to increases in bile acids (BAs). BAs are a diverse collection of amphipathic cholesterol derivatives that promote the intestinal absorption of lipids and fat-soluble vitamins. Synthesized in the liver, primary BAs are conjugated to glycine and taurine prior to storage in the gall bladder and subsequent secretion into the duodenum. Specific transporters in the ileum actively recycle the majority of BAs to the liver. Residual BAs transiting to the colon are modified by the microbiome including deconjugation, dehydroxylation, and dehydration to generate secondary BAs. In addition to their detergent effects, BAs function as endogenous ligands for the G-protein coupled bile acid receptor (TGR5), and several nuclear receptors including the farnesoid X receptor (FXR). FXR is considered the

master regulator of BA homeostasis, controlling the transcription of key genes regulating the synthesis and transport of BAs. Of note, BA modifications differentially affect their transport, receptor efficacy, and cytotoxicity.

Clinical studies have reported reduced microbial diversity, along with a shift from dietary carbohydrate utilization to amino acid degradation in CRC patients (Wirbel *et al.*, 2019). In addition, increased fecal levels of the secondary BAs lithocholic acid (LCA) and deoxycholic acid (DCA) relative to healthy controls have been shown (Gill and Rowland, 2002). While both preclinical and patient-based studies support a role for gut dysbiosis in CRC susceptibility/progression, the interactions between the microbiome and the host are incompletely understood. The ability to control genetic and environmental confounders in preclinical studies offers the potential for causal relationships to be identified, despite species differences in the composition of the microbiome and BA pools. Mice with a mutant allele of the APC gene ($APC^{min/+}$) develop multiple intestinal neoplasia predominantly within the ileum, however these lesions seldom progress past the adenoma stage (Powell et al., 1992). We previously showed that challenging $APC^{min/+}$ mice with a high fat diet (HFD) was sufficient to drive the progression from adenoma to adenocarcinoma (Fu et al., 2019). Using this $APC^{min/+}$ mouse model of colorectal cancer, we show here that the effects of a high fat diet on the cecum microbiome are more pronounced than those from the genetic mutation, significantly reducing microbial alpha diversity and perturbing the metabolome. The presence of microbially-conjugated BA with the capacity to drive intestinal cell proliferation in high fat diet (HFD) fed mice identifies potential drivers of disease progression

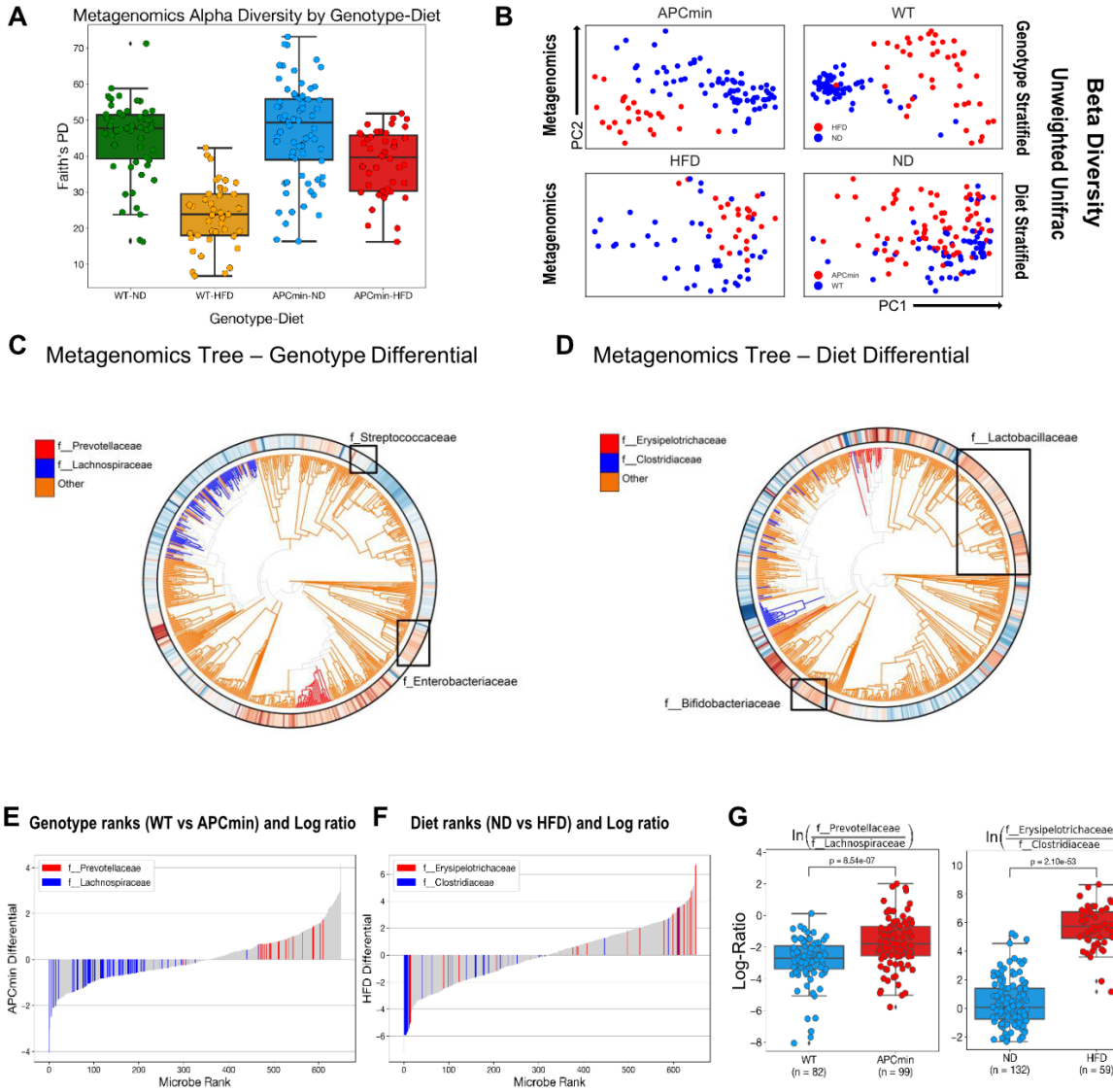
3.2. Results

To understand the effects of genetics and diet on CRC progression, we compared wild type (WT) and $APC^{min/+}$ mice maintained on a normal diet (ND) or HFD. Changes in the gut

microbiome were detected using 16S and shotgun metagenomics of cecum samples stratified by both genotype (WT compared to $APC^{min/+}$) and diet (ND compared to HFD), resulting in 4 groups (Thomas et al., 2019) (Hillmann et al., 2020; Wirbel *et al.*, 2019; Yachida *et al.*, 2019). The bacterial diversity and composition in these 4 groups were characterized by both alpha and beta-diversity (Figure 3.1A- B; S1A) (Morton et al., 2019b). Faith's phylogenetic alpha diversity of cecum microbiomes was lower in HFD-fed compared to ND-fed WT mice (Figure 3.1A) (Zhu et al., 2019). Somewhat surprisingly, microbial richness was largely unaffected in the genetic model susceptible to CRC ($APC^{min/+}$ compared to WT mice on ND) (Figure 3.1A). As seen with WT mice, HFD reduced alpha diversity in $APC^{min/+}$ mice, albeit to a lesser extent (Figure 3.1A). Similarly, we observed more profound beta-diversity differences related to diet (ND vs HFD) in both WT and $APC^{min/+}$ mice than between genotypes (Figure 3.1B, S1A). We then explored the association of specific microbial taxa with mouse genotype and/or diet through differential abundance ranking (see Methods) (Figure AB.1.S1B) (Morton *et al.*, 2019b). Of note, species of the genus *Prevotellaceae* were more associated with $APC^{min/+}$ genotype, whereas species in the order *Coriobacteriales*, family *Erysipelotrichaceae*, and genus *Lactobacillus* were more associated with HFD phenotypes (Figure 1C-D) (Zhu *et al.*, 2019). These metagenomic genetic and diet differentials also show strong associations with particular taxonomic features at the family level in the differential rank plot (Figure 1E-F) (Morton *et al.*, 2019b). In particular, *Prevotellaceae* and *Lachnospiraceae* are more strongly associated with WT and $APC^{min/+}$ mice, respectively, while *Clostridiaceae* and *Erysipelotrichaceae* are associated with ND and HFD, respectively (log-ratios in indicated comparisons, Figure 1G).

Figure 3.1: Genetics and diet reshape the gut microbiome a) Alpha-diversity of wild-type (WT) and APCmin/+ mice maintained on normal-chow diet (ND) and high fat diet (HFD). Within-sample diversity is measured by Faith's phylogenetic diversity. Metrics from shotgun metagenomics sequencing data of cecum samples are presented by genotype-diet combination. b) Unweighted Unifrac measures of beta-diversity in mice from A. Metrics from shotgun metagenomic sequencing data are stratified by genotype and diet factors and visualized using Principal Coordinate Analysis (PCoA).c-d) Ultrametric phylogenetic tree generated from shotgun metagenomics data of cecum samples in mice from A. Microbial features colored by c) Songbird genotype and d) diet differentials. Red indicates positive association while blue indicates negative association (both relative to all other features). e-f) The differential rank plot of selected microbial features separating samples by e) genetic and f) diet ()Features in red correspond to those in the numerator while those in blue correspond to features in the denominator. Features that are colored gray are not factored into the log-ratio calculations. g) Log-ratios of selected microbial families separating samples across genotype (left) and diet (right) genotype. Family selection was performed by using Qurro to inspect differentially abundant microbial groups according to Songbird differentials.

Figure 1

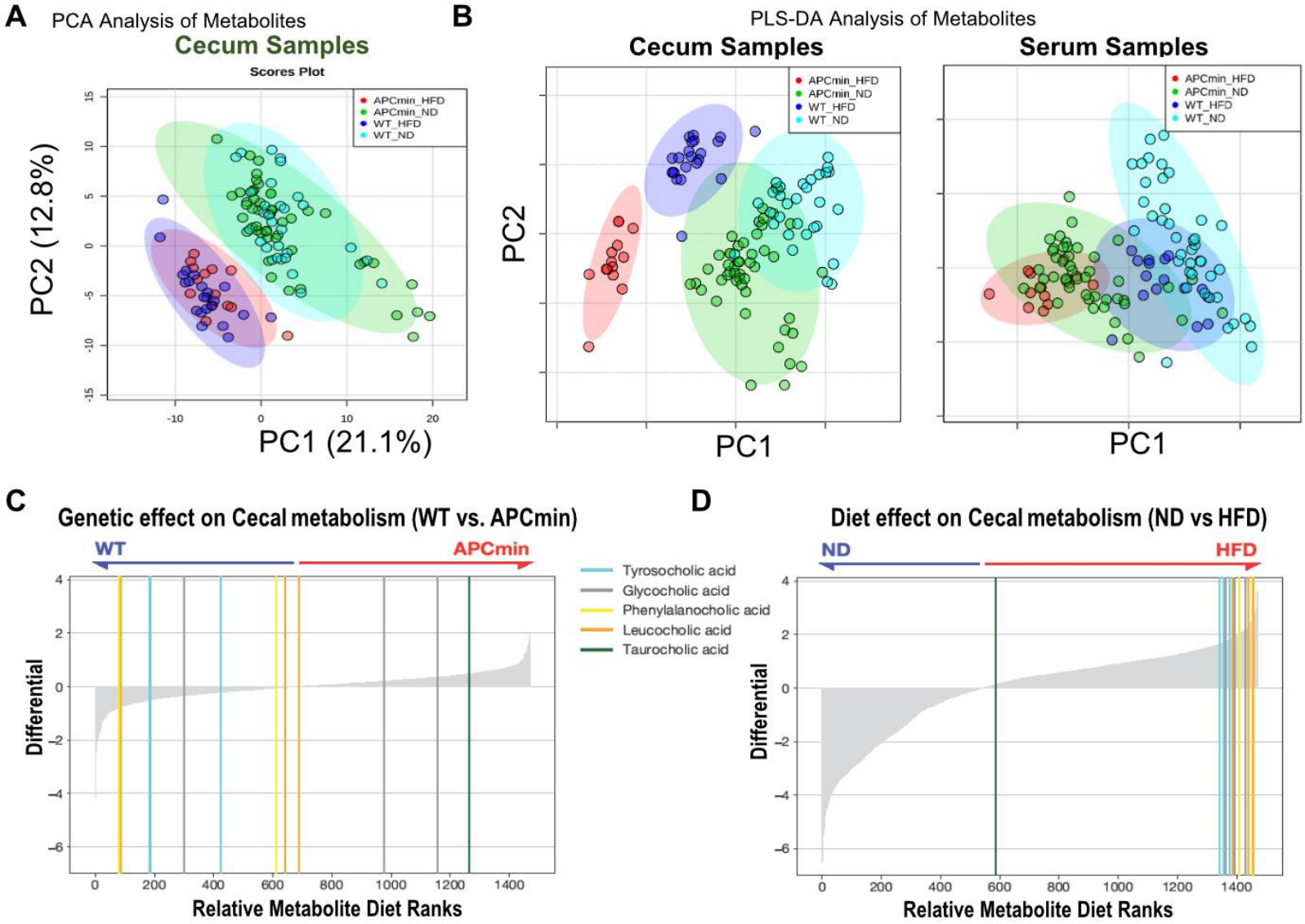

A
Metagenomics Alpha Diversity by Genotype-Diet
Faith's PD
WT-ND
WT-HFD
APCmin-ND
APCmin-HFD
Genotype-Diet
B
APCmin
WT
HFD
ND
Metagenomics
PC2
PC1
Genotype Stratified
Diet Stratified
Beta Diversity
Unweighted Unifrac
C
Metagenomics Tree – Genotype Differential
f__Prevotellaceae
f__Lachnospiraceae
Other
f_Streptococcaceae
f_Enterobacteriaceae
D
Metagenomics Tree – Diet Differential
f__Erysipelotrichaceae
f__Clostridiaceae
Other
f__Lactobacillaceae
f__Bifidobacteriaceae
E
Genotype ranks (WT vs APCmin) and Log ratio
APCmin Differential
Microbe Rank
F
Diet ranks (ND vs HFD) and Log ratio
HFD Differential
Microbe Rank
G
ln(f__Prevotellaceae / f__Lachnospiraceae)
p = 8.54e-07
Log-Ratio
WT (n = 82)
APCmin (n = 99)
ln(f__Erysipelotrichaceae / f__Clostridiaceae)
p = 2.10e-53
ND (n = 132)
HFD (n = 59)

Complementing the metagenomic analyses, paired metabolomic profiling was performed to gain additional insight into the functional consequences of the genetic and dietary changes (Koppel and Balskus, 2016). Unbiased LC-MS profiling of cecum and serum samples from both WT and $APC^{min/+}$ mice on ND and HFD was performed (Figure AB.1.S2A). Principal component analysis (PCA) of the cecum samples revealed metabolomic differences associated with diet (Figure 3.2A). To confirm this association, partial least square-discriminant analysis (PLS-DA) of cecum and serum samples showed metabolomic differences among the four sample groups, with a permutation test (100 replicates, p-value < 0.01) suggesting that the model fit is better than a chance permutation of the labels (Figure 3.2B, AB.1.S2B, C). Diet was the dominant determinant of the cecum metabolome, in agreement with the microbiome analyses, whereas less pronounced diet-induced differences were evident in the serum samples (Figure 3.2B). Over 110 metabolites were determined to be significantly dysregulated in comparing HFD and ND fed mice (p-value ≤0.05, fold change ≥1.5). Pathway enrichment analysis of the dysregulated metabolites identified 10 metabolic pathways (Figure AB.1.S4A), with the aminoacyl-tRNA biosynthesis pathway most affected. Ranking metabolites by differentials revealed relatively minor genetic effects on the cecal metabolome (largely less than 2-fold changes), with similar numbers of metabolites increased and decreased in $APC^{min/+}$ compared to WT mice (Figure 3.2C). In contrast, the magnitude of the changes was greater with diet, with HFD markedly reducing the levels of approximately 30% (decreased up to 6-fold), and increasing the concentrations of 70% of the differentially regulated metabolites (increased up to 3-fold) (Figure 3.2D). Metabolites reduced in $APC^{min/+}$ compared to WT mice included several lysophosphatidylcholine (LPC) species (Figure AB.1.S4B), in agreement with reduced LPCs reported in colorectal cancer patients (CRC) (Zhao et al., 2007), while the observed reduction in C16 acylcarnitine (ACAR 16:0) contrasts with reported increases seen in patient-derived serum (Jing et al., 2017) (Farshidfar et al., 2018). Correlation-based metabolic network analysis, where each node represents one metabolite and

the edge between two nodes represents the correlation coefficient between two metabolites (red and blue lines representing positive and negative correlations, respectively), reveals the global effects of high fat diet on the dysregulated metabolites (Figure AB.1.S4C). Exploiting the finding that structurally related molecules produce similar MS fragmentation patterns, spectral similarity scores were calculated using MS-DIAL with the embedded Bonanza spectral clustering algorithm. Subsequent network analyses facilitated the visualization of chemical similarities across the entire metabolome, wherein each node represents an ion with an associated fragmentation pattern, and the links among the nodes indicate spectral similarities (visualized in Cytoscape, Figure AB.1.S4) (Watrous et al., 2012) (Wang et al., 2016) (Forsberg et al., 2018; Huan et al., 2017; Nothias et al., 2020).

Figure 3.2: Genetics and diet affect serum and fecal metabolomes. a) Principal component analysis (PCA) of cecum metabolites from WT and APCmin/+ mice maintained on ND and HFD. b) Partial least square-discriminant analysis (PLS-DA) score plots of cecum (left) and serum metabolites (right) from mice in A, with a model p-value < 0.01. (c-d) Songbird differential rank plots of the association of metabolites with (genotype) left and diet (right). Differentials were calculated with multinomial regression and validated by comparing to a null model.

Figure 2

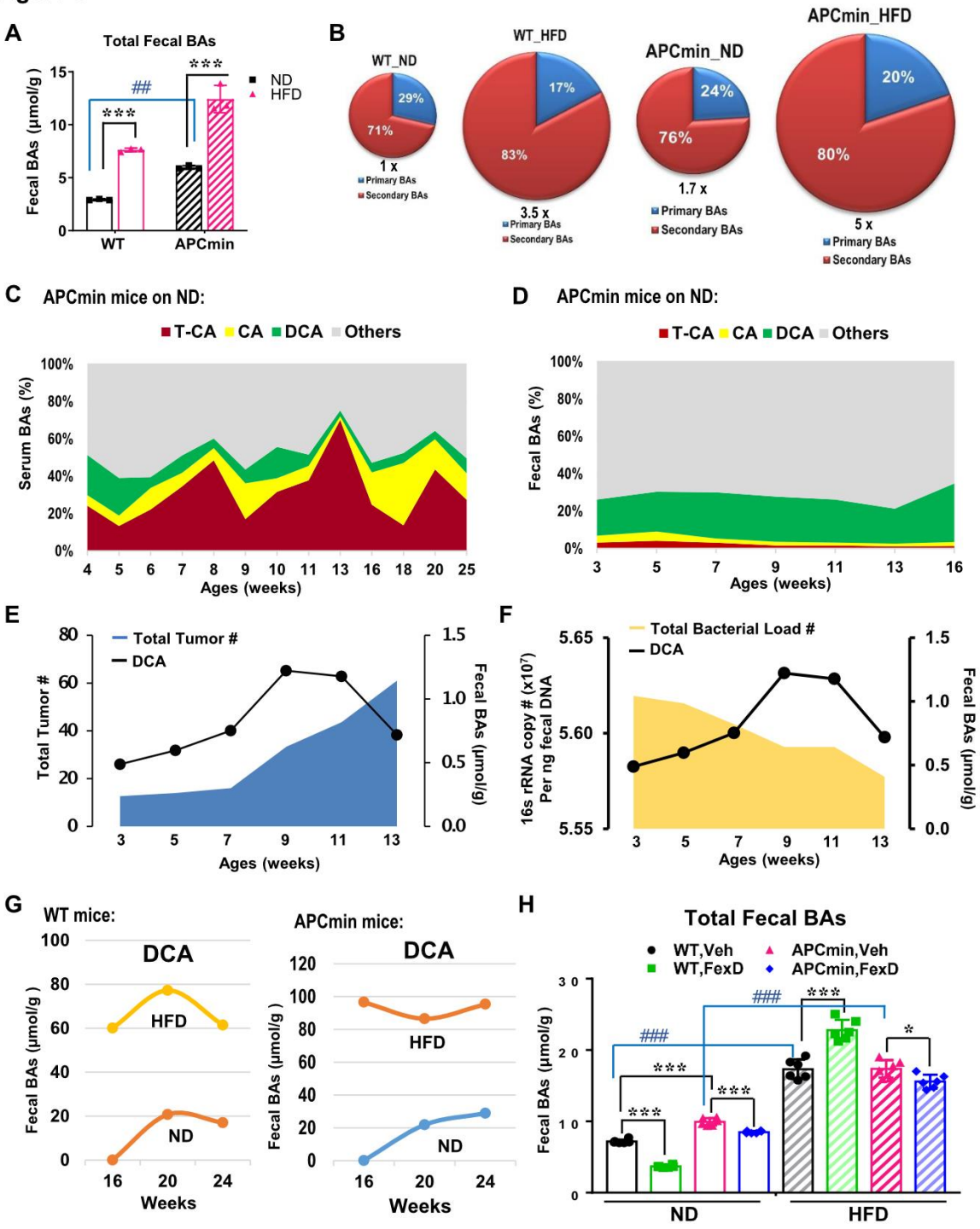

A
PCA Analysis of Metabolites
Cecum Samples
Scores Plot
APCmin_HFD
APCmin_ND
WT_HFD
WT_ND
PC2 (12.8%)
PC1 (21.1%)
B
PLS-DA Analysis of Metabolites
Cecum Samples
Serum Samples
PC2
PC1
C
Genetic effect on Cecal metabolism (WT vs. APCmin)
WT
APCmin
Tyrosocholic acid
Glycocholic acid
Phenylalanocholic acid
Leucocholic acid
Taurocholic acid
Differential
Relative Metabolite Diet Ranks
D
Diet effect on Cecal metabolism (ND vs HFD)
ND
HFD
Differential
Relative Metabolite Diet Ranks

Dietary changes are reflected in fecal bile acids

Previously, we found that HFD-induced increases in secondary BAs including DCA and TβMCA were sufficient to drive an adenoma to adenocarcinoma progression in the $APC^{min/+}$ CRC mouse model (Fu et al., 2019). To further understand the impact of dietary and genetic factors on microbially-derived secondary BAs, we measured fecal BAs in WT and $APC^{min/+}$ mice on ND or HFD. Consistent with previous findings, HFD and the APC mutation independently and cooperatively increased fecal BA content (Figure 3.3A). In addition, HFD increased the proportion of secondary BAs in both WT and $APC^{min/+}$ mice (Figure 3.3B). Given that total bacterial load and alpha diversity decreased with HFD, the proportional increases in secondary bile acids implicate compositional changes in the microbiome, rather than absolute bacterial load, in mediating these changes (Figure 3.1A).

Figure 3.3: Genetics and diet affect fecal bile acids Wild-type (WT) and APCmin/+ mice were maintained on the normal-chow diet (ND) or high fat diet (HFD) from 4 weeks of age.a) Total fecal bile acids. b) Proportions of primary and secondary bile acids in feces. (c-d) Progressive changes in bacterially-mediated conversion of tauro-cholic acid (T-CA) to cholic acid (CA) and deoxycholic acid (DCA) in c) serum and d) feces. e) Temporal changes in intestinal tumor burden and fecal DCA levels in APCmin/+ mice on ND. f) Temporal changes in bacterial load and DCA in feces from APCmin/+ mice on ND. g) Temporal changes of fecal DCA levels in WT and APCmin/+ mice on ND and HFD during tumor progression (16 to 24 weeks). h) Fecal bile acid levels in WT and APCmin/+ mice on ND and HFD treated with the FXR agonist FexD (50mg/kg/day) or vehicle for 8 weeks. n= 3-6. Data represent the mean ± SEM. For two group comparation, Student's unpaired t-test. For more than two group comparation, one-way Anova. *, # $p<0.05$; **, # # $p<0.01$; ***, # # # $p<0.005$.

Figure 3


A
Total Fecal BAs
Fecal BAs (μmol/g)
ND
HFD
WT
APCmin
B
WT_ND
WT_HFD
APCmin_ND
APCmin_HFD
Primary BAs
Secondary BAs
C
APCmin mice on ND:
T-CA CA DCA Others
Serum BAs (%)
Ages (weeks)
D
APCmin mice on ND:
Fecal BAs (%)
E
Total Tumor #
DCA
F
Total Bacterial Load #
16s rRNA copy # (x10⁷) Per ng fecal DNA
G
WT mice:
APCmin mice:
HFD
ND
Weeks
H
Total Fecal BAs
WT,Veh
WT,FexD
APCmin,Veh
APCmin,FexD

To complement these metabolomic studies, the progressive changes in total and specific fecal BA species were determined by enzymatic assay and targeted mass spectrometry, respectively in $APC^{min/+}$ mice (Figure AB.1.S5A, B). Contrasting with largely static serum levels, fecal levels of ω-muricholic acid (ωMCA) increased with age (Figures AB.1.S5C-E). A similar pattern was seen in serum and fecal deoxycholic acid (DCA) (Figures 3.3C, D), while a transient decrease in β-muricholic acid (βMCA) levels was seen coinciding with tumor initiation (~7weeks, Figure AB.1.S5F). This lack of correlation between fecal and serum levels of ωMCA and DCA is presumed to be a consequence of differential BA uptake in the colon (Degirolamo et al., 2011; Wahlstrom et al., 2016). However, at increased tumor load (~13 weeks of age), the reduction in fecal bacterial load coincided with reduced DCA and ωMCA levels (Figures 3.3E, 3.F, AB.1.S5E), consistent with the role of the gut microbiome in the generation of secondary BAs. In contrast, βMCA levels increased during the later stages of tumor progression, potentially driven by tumor-specific changes in the microbiome (Figure AB.1.S5F).

We next sort to determine the specific BA species contributing to the HFD-induced increases in fecal BAs in both WT and $APC^{min/+}$ mice (Figure 3.3A). Notably, fecal DCA and ωMCA levels increased 60-100 and 150-300 fold, respectively, in mice maintained for 16 weeks on HFD (Figure 3.3G, AB.1.S5G). $APC^{min/+}$ mice were more susceptible to the dietary challenge, with increases in fecal DCA and ωMCA levels 3-5 and 6-7 fold greater than those in WT mice, respectively (Figure 3.3G, AB.1.S5G). In contrast, the diet-induced changes in serum CA levels were relatively minor, however concentrations were an order of magnitude higher in HFD-fed $APC^{min/+}$ mice (Figure AB.1.S3H).

As the master regulator of BA homeostasis, activation of FXR can reduce serum BA levels in HFD-fed WT and $APC^{min/+}$ mice (Fu *et al.*, 2019). To explore the impact of FXR on the fecal BA pool, ND and HFD-fed mice were treated with the intestinally-biased FXR agonist FexD (50 mg/kg/day for 8 weeks; Figures AB.1.S6A, C). Intestinal FXR activation reduced total fecal BAs

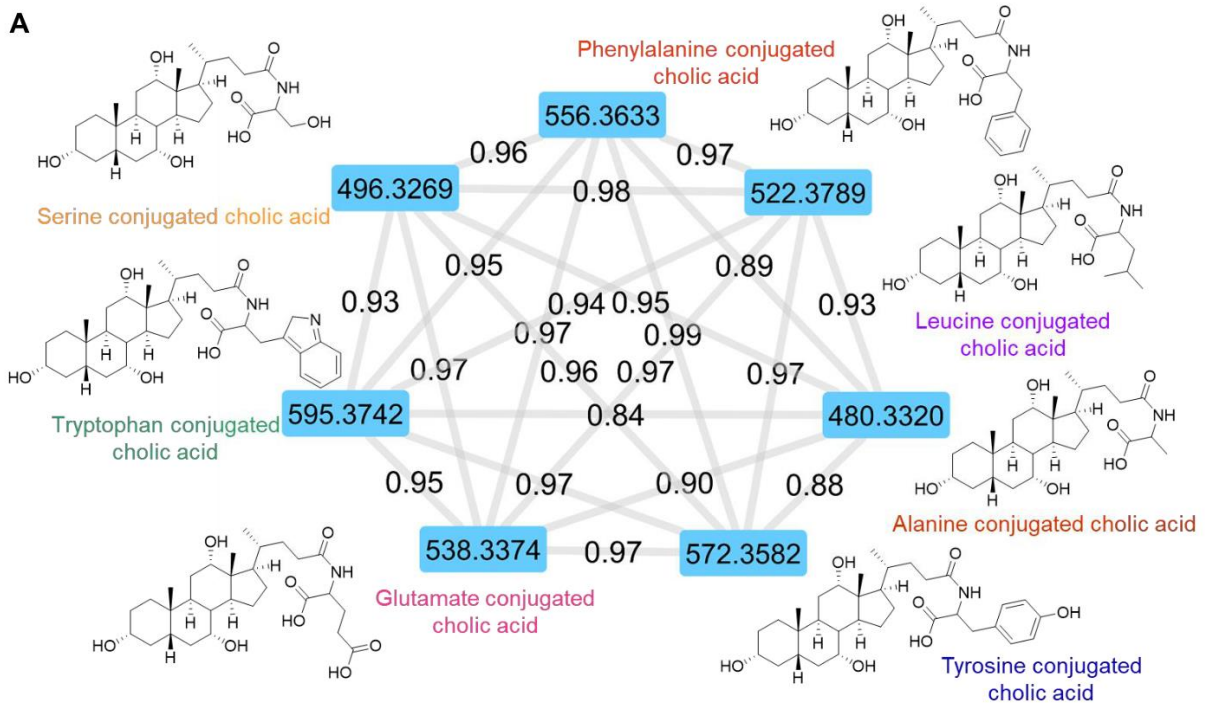
in both WT and APC$^{min/+}$ mice on ND (Figure 3.3H). In contrast, a differential effect was evident in HFD-fed mice, with FexD treatment increasing and decreasing total fecal BA levels in WT and APC$^{min/+}$ mice, respectively (Figure 3.3H). Profiling the fecal BA composition of ND-fed mice revealed model-specific FexD-mediated reductions that were largely lost in the HFD cohorts, illustrating the complexity of factors affecting BA homeostasis (Figures AB.1.S6B, D) (Friedman et al., 2018; McCarville et al., 2020; Morton et al., 2019a).

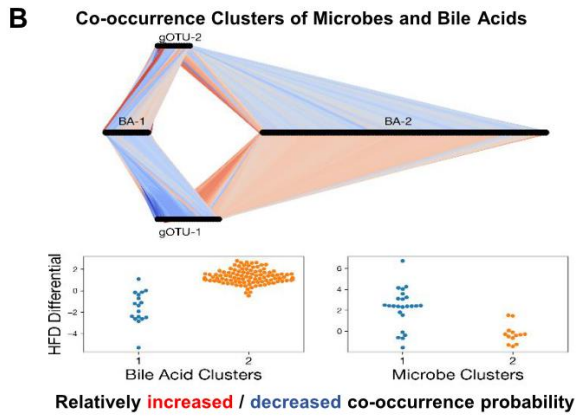
Novel conjugated bile acids associated with high fat diet

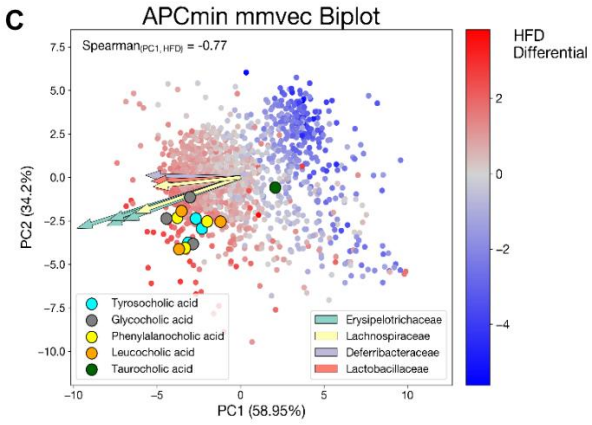
Recently, we identified 3 novel amino acid conjugated cholic acid species, and showed that these microbially-generated BAs were enriched in patients with inflammatory bowel disease (Quinn et al., 2020). Building on this study, we interrogated the cecal metabolome datasets for evidence of non-classic amino acid conjugated BAs. In agreement with our earlier study, phenylalanine (Phe), leucine (Leu), and tyrosine (Tyr) conjugated cholic acid were detected. In addition, serine (Ser), alanine (Ala), tryptophan (Trp), and glutamine (Glu) conjugated cholic acid were also identified (Figure 3.4A). The core cholic acid mass spectral fragmental pattern was evident in these non-classic conjugated CA derivatives (AA-CAs), with additional patterns consistent with the presence of the identified amino acids conjugated through an amide bond at the normal glycine/taurine conjugation site (Figure 3.4A). Moreover, the proposed structures were validated using synthesized standards with retention time and MS/MS fragmentation patterns matching on several instrument platforms, including targeted mass spectrometry. While the levels of AA-CAs varied between individual mice, HFD increased the levels of Gly-CA and Phe-CA in both WT and APC$^{min/+}$ mice, and the concentrations of Leu-CA and Ser-CA in WT mice (Figure AB.1.S6E). Consistent with earlier observations (Quinn *et al.*, 2020), these AA-CAs were detected in cecum but not in serum samples, supporting the notion that they are synthesized by the gut microbiome (Quinn *et al.*, 2020).

Figure 3.4: Non classic amino acid conjugated bile acids in cecum sample a) MS^2 spectra network analysis of the detected 7 novel bile acids. Chemical structure and molecular weight are presented. b) Mmvec microbe-metabolite co-occurrences study of tumor progression in $APC^{min/+}$ mice on ND (adenoma) and HFD (adenocarcinoma). Conditional probabilities exhibit a biclustering pattern between bile acids and gOTUS corresponding to ND and HFD. Connections between microbes and metabolites correspond to increased or decreased co-occurrence probability relative to all other microbes. Association was assigned by comparing cluster features to both metagenomic and metabolomic Songbird differentials. c) Biplot of mmvec results from $APC^{min/+}$ mice. Points represent metabolites and arrows represent most informative microbial features. Color of points corresponds to the Songbird-calculated association of each metabolite with the high-fat diet compared to the normal diet. Novel bile acids are highlighted with different colors. Spearman correlation between PC1 of the mmvec ordination and the HFD differential was 0.77.

Figure 4


A
Phenylalanine conjugated cholic acid
556.3633
0.96
0.97
496.3269
0.98
522.3789
Serine conjugated cholic acid
0.95
0.89
0.93
0.94
0.95
0.93
0.97
0.99
Leucine conjugated cholic acid
0.97
0.96
0.97
0.97
Tryptophan conjugated cholic acid
595.3742
0.84
480.3320
0.95
0.97
0.90
0.88
Alanine conjugated cholic acid
538.3374
0.97
572.3582
Glutamate conjugated cholic acid
Tyrosine conjugated cholic acid


B
Co-occurrence Clusters of Microbes and Bile Acids
gOTU-2
BA-1
BA-2
gOTU-1
HFD Differential
Bile Acid Clusters
Microbe Clusters
Relatively increased / decreased co-occurrence probability


C
APCmin mmvec Biplot
Spearman(PC1, HFD) = -0.77
PC2 (34.2%)
PC1 (58.95%)
HFD Differential
Tyrosocholic acid
Glycocholic acid
Phenylalanocholic acid
Leucocholic acid
Taurocholic acid
Erysipelotrichaceae
Lachnospiraceae
Deferribacteraceae
Lactobacillaceae

Non-classic conjugated BAs are microbially generated

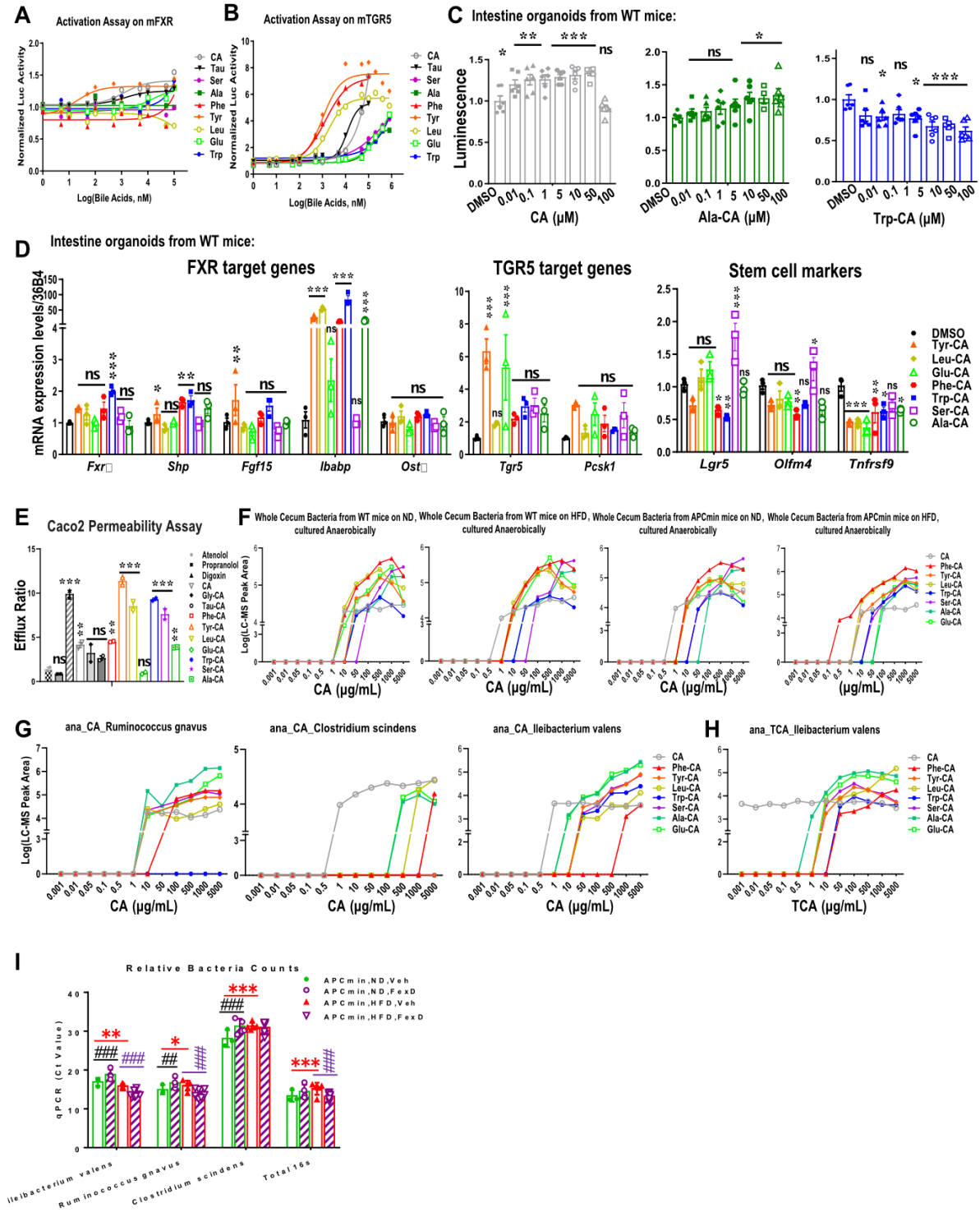
To explore causal associations between microbial content and the metabolome, the co-occurrence probabilities between microbial taxa and metabolites in $APC^{min/+}$ cecum samples were estimated (Morton *et al.*, 2019a). The conditional probability of observing a metabolite, given that a microbe was observed, was estimated using the neural network MMvec (microbe-metabolite vectors)(Morton *et al.*, 2019a), that predicts metabolite abundances from microbe sequences. Using operational taxonomic units (OTUs) to cluster microbes based on sequence similarities, associations between BAs and microbial species were determined by normalized conditional probabilities. These analyses revealed the effect of diet on the co-occurrence of clusters of microbes and specific BAs in $APC^{min/+}$ mice (Figure 3.4B) (Morton *et al.*, 2019a). Moreover, the MMvec showed clear stratification of the metabolomics data according to diet effects in $APC^{min/+}$ mice (Figure 3.4C). We then performed differential abundance on the identified metabolites to parallel the metagenomic analysis in determining which metabolic features were associated with diet and genotype. Considering that these AA-BAs also use amino acids as resources, we used the MMvec results to identify candidate producers by Spearman correlation analysis of the first MMvec principal component with HFD log-fold changes in $APC^{min/+}$ mice (Figure 3.4C). A strong correlation was observed, indicating that PC1 seems to be strongly driven by diet (Figure AB.1.S6F) (Morton *et al.*, 2019a). Detailed correlations between candidate producers and different BA categories are presented as a cluster map (Figure AB.1.S6G). Several bacterial species are highly correlated with AA-CA production (Morton *et al.*, 2019a) (Quinn *et al.*, 2020). In particular, Tyr-, Phe-, and Leu-conjugated CA are highly associated with *Erysipelotrichaceae*, *Lachnospiraceae*, and *Lactobacillaceae* (Figure 3.4C) (Henke et al., 2019; Quinn *et al.*, 2020; Yachida *et al.*, 2019).

Novel conjugated bile acids are biologically functional

We next explored how these alternative amino acid conjugations affect CA-driven physiology. Initially, signaling through FXR and TGR5 (encoded by G-protein-coupled bile acid receptor 1, GPBAR1, a membrane bound BA receptor) was evaluated using luciferase reporters containing FXR or TGR5 downstream cAMP response elements transfected into kidney HEK293 cell overexpressing human or mouse FXR or TGR5 genes (Dobin et al., 2013; Sorrentino et al., 2020). In these reporter assays, Phe, Tyr, Trp, and Glu conjugation increased CA signaling via human, and to a lesser extent mouse FXR compared to taurine and glycine, while Leu, Ser, and Ala conjugated CA functioned as weak FXR agonists or even antagonists (Figures 3.5A, AB.1.S7A). In terms of TGR5, Leu, Phe, and Tyr conjugation increased, and Glu, Trp, Ala and Ser conjugation decreased CA activation of mouse TGR5 relative to the taurine conjugate, while all alternate conjugations reduced or eliminated activation of human TGR5 (Figures 3.5B, AB.1.S7B). To associate AA-CA signaling with functional outcomes, the abilities to promote intestinal cell proliferation were compared (Fu *et al.*, 2019; Sorrentino *et al.*, 2020). Taurine conjugation reduced the ability of CA to promote growth of intestinal organoids derived from WT mice, but had less of an effect on $APC^{min/+}$ derived intestinal organoids (Figures AB.1.S7C, E). Ala conjugation similarly attenuated CA-driven proliferation, while Ser, Leu, and Glu conjugated CA largely eliminated the proliferative effects (Figures 3.5C, AB.1.S7D). Somewhat unexpectedly, Ser-CA, a bile acid previously found to associate with lymphocytic choriomeningitis virus (LCMV) infection in mice and with human Crohn's disease (Wang *et al.*, 2016) increased the expression of intestinal stem cell marker genes in WT organoids (Figure 3.5D). In contrast, CA conjugated with the aromatic amino acids (Trp, Tyr, Phe) inhibited organoid proliferation, consistent with reduced expression of stem cell marker genes in WT organoids (Figures 3.5C, D, AB.1.S7D-F). Interestingly, these AA-CAs showed varying activation of FXR and TGR5 target genes, alluding to the complexity of BA functionality (Figures 3.5D, AB.1.S7G).

Figure 3.5: Non classic conjugated BAs are bioactive and can be synthesized by specific gut microbesa) Dose-dependent activation of exogenous mouse FXR by amino acid conjugated cholic acid species. Luciferase activity in HEK293 cells expressing a luciferase reporter gene functionally linked to an FXR-responsive element (FXRE-Luc). b) Dose-dependent activation of exogenous mouse TGR5 by amino acid conjugated cholic acid species. Luciferase activity in HEK293 cells expressing a luciferase reporter gene functionally linked to a cAMP-responsive element which is downstream of TGR5.c) Dose-dependent proliferation of intestinal organoids from WT mice treated with Ala-CA (left) and Trp-CA (right), measured by luminescent cell viability assay. d) Relative expression of FXR and TGR5 target genes, and intestinal stem cell marker genes in intestinal organoids from WT mice treated with amino acid conjugated cholic acid species at 10μM. e) Cellular transport of amino acid conjugated cholic acid species, as determined by the efflux ratio in Caco2 cells. Atenolol and propranolol serve as negative and positive controls, respectively. Digoxin serves as a positive control for P-glycoprotein-mediated efflux. f) Dose-dependent generation of conjugated cholic acid species in anaerobic cultures of cecal bacteria from WT and $APC^{min/+}$ mice on ND and HFD. Cultures were supplemented with increasing concentrations of cholic acid (CA) for 48h prior to mass spectral analysis. g) Dose-dependent generation of conjugated cholic acid species in anaerobic cultures of *Ruminococcus gnavus*, *Clostridium scindens,* and *Ileibacterium Valens.* Cultures were supplemented with increasing concentrations of cholic acid (CA) for 48h prior to mass spectral analysis. h) Dose-dependent generation of conjugated cholic acid species in anaerobic cultures of *Ileibacterium Valens.* Cultures were supplemented with increasing concentrations of taurocholic acid (T-CA) for 48h prior to mass spectral analysis. i) Changes in *Ileibacterium Valens*, *Ruminococcus gnavus*, and *Clostridium scindens* levels in ND and HFD-fed $APC^{min/+}$ mice treated with FexD or vehicle for 8 weeks, determined by qPCR. Data represent the mean ± SEM. *p<0.05; **p<0.01; ***p<0.005. Student's unpaired t-test.

Figure 5


A
Activation Assay on mFXR
Normalized Luc Activity
Log(Bile Acids, nM)
CA
Tau
Ser
Ala
Phe
Tyr
Leu
Glu
Trp
B
Activation Assay on mTGR5
C
Intestine organoids from WT mice:
Luminescence
DMSO
CA (μM)
Ala-CA (μM)
Trp-CA (μM)
D
Intestine organoids from WT mice:
FXR target genes
TGR5 target genes
Stem cell markers
mRNA expression levels/36B4
Fxr
Shp
Fgf15
Ibabp
Ost
Tgr5
Pcsk1
Lgr5
Olfm4
Tnfrsf9
DMSO
Tyr-CA
Leu-CA
Glu-CA
Phe-CA
Trp-CA
Ser-CA
Ala-CA
E
Caco2 Permeability Assay
Efflux Ratio
Atenolol
Propranolol
Digoxin
CA
Gly-CA
Tau-CA
Phe-CA
Tyr-CA
Leu-CA
Glu-CA
Trp-CA
Ser-CA
Ala-CA
F
Whole Cecum Bacteria from WT mice on ND, cultured Anaerobically
Whole Cecum Bacteria from WT mice on HFD, cultured Anaerobically
Whole Cecum Bacteria from APCmin mice on ND, cultured Anaerobically
Whole Cecum Bacteria from APCmin mice on HFD, cultured Anaerobically
Log(LC-MS Peak Area)
CA (μg/mL)
(μg/mL)
G
ana_CA_Ruminococcus gnavus
ana_CA_Clostridium scindens
ana_CA_Ileibacterium valens
H
ana_TCA_Ileibacterium valens
TCA (μg/mL)
I
Relative Bacteria Counts
APCmin,ND,Veh
APCmin,ND,FexD
APCmin,HFD,Veh
APCmin,HFD,FexD
qPCR (Ct Value)
Ileibacterium valens
Ruminococcus gnavus
Clostridium scindens
Total16s

Wnt signaling is an important driver of intestinal stem cell growth, with nuclear β-catenin activity largely mediated by the downstream TCF/LEF (T-cell factor/lymphoid enhancer-binding factor) pathway. Surprisingly, each of the AA-CAs promoted Wnt signaling, albeit at supraphysiological concentrations, as measured by TCF/LEF activity in the colon cancer cell line, HT29 cells. (Figures AB.1.S7H-I).

While 90% of classic BAs are recycled to the liver, we previously reported that Phe-CA, Tyr-CA, and Ala-CA were not detected in mouse or human portal or peripheral blood (Quinn *et al.*, 2020). To predict the potential for these AA-CAs to be recycled, Caco 2 permeability assays were performed with Atenolol and Propranolol as low and high permeability controls, and Digoxin as a substrate for transporter mediated uptake. With the exception of Glu-CA, the alternate AA-CAs displayed high efflux ratios (>2). However, the markedly reduced permeabilities compared to Gly-CA suggests that these alternately conjugated BAs are transported at much lower levels or not at all into the bloodstream (Figures 3.5E, AB.1.S7J).

To support the notion that these AA-BAs are microbially derived, total cecal bacteria were cultured in the presence of increasing concentrations of exogenous cholic acid. Each of the AA-CAs was detected in anaerobic cultures, with increased levels seen in cecal bacterial collected from HFD mice (Figure 3.5F). Aerobic cultures were similarly able to generate the AA-CAs, albeit at lower levels (Figure AB.1.S8A). In addition, anaerobic cultures were able to utilize Tauro-CA as a substrate, consistent with the presence of bile salt hydrolases (BSHs) in these cecum cultures (Figure AB.1.S8B). However, the majority of cultures did not efficiently conjugate CDCA or DCA (Figures AB.1.S8C-D). To validate the predicted associations of AA-CAs with specific bacterial species (Figure 3.4C), individual cultures of *Ruminococcus gnavus*, *Clostridium scindens*, *Ileibacterium valens*, *Lactobacillus reuteri,* and *Lactobacillus acidophilus* were incubated with increasing concentrations of CA. Interestingly, preferences for conjugating the

different amino acids were seen with individual bacteria, (Figures 3.5G, AB.1.S8E) (Wong and Yu, 2019) (Buffie et al., 2015; Henke *et al.*, 2019). Moreover, activation of FXR increased bacterial loads of *Ileibacterium valens* and *Ruminococcus gnavus* in HFD fed $APC^{min/+}$ mice (Figure 3.5I). These data highlight the complexity of the microbiome-metabolome interaction, as well as implicate specific strains such as *Ileibacterium valens* in CRC (Cox et al., 2017; Kadosh et al., 2020; Yachida *et al.*, 2019).

3.4. Discussion

This study demonstrates the combinatorial effects of genetic and dietary risk factors on the gut microbiome and metabolome. Our results show that both genetic and dietary risk factors contribute to the alterations in the serum and cecum metabolome profiles. Interestingly, the effects of diet are more pronounced in cecum than in serum. Distinct from the effects on serum metabolites, we find that diet is the major determinant of cecum metabolites and gut microbiome species.

As early dietary sensors and genetic effectors, bile acids have emerged as pleiotropic signaling molecules mediating intestinal tumorigenesis and inflammation (Fiorucci and Distrutti, 2015; Wahlstrom *et al.*, 2016). Recent technological advances have led to the characterization of more than 170 bile acids (Hoffmann et al., 2022; Petras et al., 2021; Wang *et al.*, 2016), of which more than 60 have been directly observed in human fecal samples (Quinn *et al.*, 2020). Here we characterize 7 non-classic amino acid-conjugated BAs enriched in HFD fed mice, consistent with previous association studies (Quinn *et al.*, 2020) (Hoffmann *et al.*, 2022; Morton *et al.*, 2019a; Wang *et al.*, 2016). These microbially-modified cholic acid derivatives appear restricted to the gut, distinguishing them from host-conjugated BAs. We show that non-classic amino acid conjugation selectively modulates cholic acid signaling via FXR and TGR5, as well as its ability to promote Wnt signaling and intestinal stem cell proliferation; key steps in CRC initiation and progression

(Fu *et al.*, 2019; Sorrentino *et al.*, 2020). Microbial diversity analysis across both genotype and diet demonstrates strong microbial association with tumorigenesis in both alpha and beta diversity. We characterize specific microbial taxa associated with both genetic and dietary effects. Furthermore, the strong taxonomic and phylogenetic association of identified microbial features points to conserved evolutionary signals strongly coupled to diet response and, to a lesser extent, genotype. Our multi-omics analyses also show that diet is the strongest driver of microbe-metabolite interactions, especially so in the identified BAs. Notably, MMvec is an unsupervised neural network, indicating a high degree of confidence in our results as the multi-omics results concord with our supervised differential abundance analysis of microbes and metabolites. Moreover, we identify potential gut microbes capable of conjugating cholic acid including *Ileibacterium valens*. *Ileibacterium valens* has been recently implicated in microbial-induced obesity and intestinal inflammation through its production of Il17 cytokines and antimicrobial peptides (Cox *et al.*, 2017). Our findings show an enrichment of *Ileibacterium valens* strains in adenocarcinoma mouse models, suggesting that this species may promote tumorigenesis (Yachida *et al.*, 2019).

In general, the modulation of gut microbiota and bile acid profiles holds promise as a novel therapeutic approach for the treatment of gastrointestinal cancers and represents the next frontier for gastrointestinal cancer research.

3.5. Materials and Methods Animals

WT C57BL/6J (Cat # 000664) and APCmin/+ (Cat # 002020) were purchased from Jackson Laboratory. All animal experiments were performed in the specific pathogen-free facilities at the Salk Institute following the Institutional Animal Care and Use Committee's guidelines. WT and APCmin/+ mice were maintained on a normal chow diet (ND) or placed on a high-fat diet (HFD, Harlan Teklad, 60% of calories from fat) from 4 weeks of age. For early intervention

experiments, Fexaramine D (FexD, 50mg/kg in corn oil) or vehicle was orally gavaged daily from 8 weeks of age for APCmin/+ mice on ND, or 6 weeks for APCmin/+ mice on HFD (Fu *et al.*, 2019).

Isolation and Generation of Intestinal organoid

Intestines were washed in ice-cold PBS (Mg2+/Ca2+) (Corning, cat # 21-031-CM), containing 2% BSA (Gemini Bio-products, cat #900-208) and 2% antibiotic-antimycotic (Gibco, cat #15240-062). Crypts and villi were exposed by dicing the intestines into small pieces (1-2 cm long), followed by extensive washes to remove contaminants. Then, a gentle cell dissociation reagent (Stem cell technologies, cat #7174) was used according to the manufacturer's instructions. Briefly, intestinal pieces were incubated on a gently rotating platform for 15 minutes. After that, the gentle cell dissociation reagent was removed and the intestines were washed 3 times with a PBS wash buffer with vigorous pipetting. The first and second fractions that usually contain loose pieces of mesenchyme and villi were not used. Fractions three and four containing the intestinal crypts were collected and pooled. Isolated crypts were filtered through a 70µm nylon cell strainer (Falcon, cat #352350). Crypts were counted, then embedded in Matrigel (Corning, growth factor reduced, cat #354230), and cultured in Intesticult organoid growth medium (Stem cell technologies, cat #6005). For mouse colon organoids, additional Wnt3a (300ng/µl, R&D, cat #5036-WN-010) was added.

Intestinal organoids used in this study were generated from WT mice, APCmin/+ mice, Lgr5-EGFP-IRES-CreERT2 mice, FXRKO mice, FXRfl/fl mice.

Bile Acids Total Amount and Composition Measurement

Metabolites such as Bile Acids were measured in mouse serum, cecum, and fecal samples by Total bile acid assay kit (Diazyme laboratories, cat #DZ042A-K). Serum samples were

diluted 1:5 with a blank buffer, and calculations performed using standard controls included in the kit. For fecal samples, total bile acids and total fat were extracted from 500mg feces.

Composition profiling of the total Bile Acids pool is measured using targeted Liquid chromatography-mass spectrometry (LC-MS). Authentic bile acid standards were purchased from Sigma, except glycolithocholic acid (GLCA), murideoxycholic acid (MDCA), HDCA, α-HCA, β-MCA, α-MCA ω-MCA, and Tauro-β-muricholic acid (T-βMCA) which were purchased from Steraloids (Newport, RI), taurocholic acid (TCA) from Calbiochem (San Diego, CA), and the deuterated bile acid standards cholic-2,2,4,4- *d*4 acid, chenodeoxycholic-2,2,4,4-*d*4 acid, and lithocholic-2,2,4,4-*d*4 acid from C/D/N Isotopes (Quebec, Canada). Mouse serum (20µl) was protein precipitated with 80µl of ice-cold acetonitrile containing 3.28ng of deuterated cholic acid (2, 2, 4, 4-d4 cholic acid) as an internal standard, vortexed 1min and centrifuged at 10,000 rpm for 10 min at 4°C. The supernatant was evaporated under vacuum at room temperature and reconstituted in assay mobile phase and transferred to a 96-well plate for analysis. A Nextera UPLC (SHIMADZU, Kyoto, Japan) system used in combination with a Q-TRAP 5500 Mass Spectrometer (AB SCIEX, Toronto, Canada) with Analyst Software 1.6.2 (Kakiyama et al., 2014).

Chromatographic separations were performed with an ACQUITY (WATERS, Milford, MA) UPLC BEH C18 column (1.7microns, 2.1x100mm). The temperatures of the column and autosampler were 65 degrees and 12 degrees, respectively. The sample injection was 1µL. The mobile phase consisted of 10% Acetonitrile and 10% Methanol in water containing 0.1% Formic Acid (Mobile Phase A) and 10% Methanol in Acetonitrile 0.1% Formic Acid (Mobile Phase B) delivered as a gradient: 0-5-min Mobile Phase B held at 22%; 5-12-min Mobile Phase B increased linearly to 60%, 12-15min Mobile Phase B increased linearly to 80% and 15-19min Mobile Phase B constant at 80% at a flow rate of 0.5ml/min. The mass spectrometer was operated in negative electrospray mode working in the multiple reaction mode (MRM). Operating parameters were Curtain gas 30psi; Ion spray voltage 4500 V; Temperature 550C; Ion Source

Gas 1 60psi; Ion Source Gas2 65psi. Transition MRMs, declustering potential, entrance potentials, and collision cell exit potentials were optimized using the Analyst software. Dwell times were 25msec.

Cell Lines, Cell viability assay, and Cell Luciferase assay

The human intestinal cancer cell lines HCT116, Caco2, and HT29 were acquired from ATCC and cultured according to the supplier's instructions. FexD and novel AA-BAs (in house production) were dissolved in DMSO for *in vitro* experiments. CellTiter-Glo Luminescent Cell Viability Assay Kit (Promega, Cat #G7572) was used to assay cell viability after drug treatment. For luciferase assay, FXRE-Luc plasmids (FXR responsive element), human and mouse version of FXR expression plasmids were transfected into each cell line, then different drugs were added, and luciferase activities were measured by Dual-Luciferase Reporter kit (Promega cat #PRE1910). Activation of TGR5 signaling was measured by Cignal cAMP response elements Reporter (Luc) Kit (Qiagen, CCS- 001L). Wnt signaling reporter assay by Cignal TCF/LEF Reporter (Luc) Kit (Qiagen, CCS- 018L) was used.

Organoid Studies

Organoids were treated with drugs either on day 2 or day 3 after plating to capture the early growth phase. Images of organoid morphology changes after drug treatment were taken with an Olympus IX51 microscope. CellTiter-Glo Luminescent 3D Cell Viability Assay Kit (Promega, Cat #G9683) was used to check the cell viability after drug treatment. Organoids were directly lysed using TRIzol reagent (Ambion, cat #15596026), followed by a brief sonication (PowerLyzertm 24 MO Bio Laboratories Inc). RNeasy Mini Kit (Qiagen, cat #74106) was used for RNA extraction.

Gene Expression Analysis

Total RNA isolated from mouse intestine was perfused with RNAlater for 24h at 4C and then tissues were homogenized in TRIzol reagent (Ambion, cat #15596026) with beads using PowerLyzertm 24 (Mo Bio Laboratories Inc), then extracted by using RNeasymini kit (Qiagen, cat #74106) as per the manufacturer's instructions. Total RNA isolated from mouse liver and intestinal segments were directly homogenized in TRIzol. cDNA was synthesized from 1µg of DNase-treated total RNA using Bio-Rad iScript Reverse Transcription supermix (#1708841) and mRNA levels were quantified by quantitative PCRwith Advanced Universal SyBr Green Supermix (Bio-Rad, cat #725271). All samples wererun in technical triplicates and relative mRNA levels were calculated by using the standardcurve methodology and normalized to 36B4. All primers are listed in Supplementary Table. AB.2.S1.

RNA-seq library generation, High-throughput sequencing, and analysis

RNA quality was confirmed using the Agilent 2100 Bioanalyzer and RNA-seq libraries were prepared from three biological replicates for each experimental condition and sequenced on the Illumina HiSeq 2500 using barcoded multiplexing and a 100-bp read length. Image analysis and base calling were done with Illumina CASAVA-1.8.2. The quality of the reads was assessed with fastqc. Reads were mapped against the reference genome and transcript annotation (GRCm38.p6) using STAR (Dobin *et al.*, 2013). RSEM was utilized to quantify gene expression from BAM files. Differentially expressed genes (n = 3) were determined using rsem-generate-data-matrix and rsem-run-ebseq commands (Li and Dewey, 2011). For GSEA, normalized expression of gene matrix from RSEM results was used with previously reported gene signatures. GSEA was performed with the default setting (Subramanian et al., 2005). To generate heatmaps, z-scores were calculated from the matrix of normalized expression in each row using R.

Fecal and Serum sample preparation for untargeted LC-MS

Fecal swabs were mixed with 500 µL ice-cold methanol. The mixture was vortexed for 2 min followed by 20 min sonication in an ice-cold water bath. The mixture was then left in a -20 freezer overnight for complete protein precipitation. As for the serum sample, 100 µL of mouse serum was first mixed with 400 µL ice-cold methanol. The mixture was then vortexed for 2 s and then left in the -20 freezer overnight for complete protein precipitation. The clear supernatant, which contains metabolites, was separated from the precipitated protein by centrifugation at 14,000 rpm for 15min and dried using speed vac. Finally, the dried metabolite solution was then reconstituted in 100 µL 1:1 ACN: H_2O for LC-MS analysis.

Data processing and interpretation

Metabolomics data was processed in MS-DIAL (Tsugawa et al., 2015) using default peak picking, alignment parameter settings. Due to the low abundance of amino acid conjugated bile acids in untargeted metabolomics results, their MS signals were further manually checked for better quantitative precision and accuracy (Yu et al., 2021). Uni-variate, Multi-variate statistical analyses and pathway enrichment analysis were performed in MetaboAnalyst (https://www.metaboanalyst.ca) (Chong et al., 2018; Guo and Huan, 2020).

LC-MS-based Targeted Analysis for Seven Amino Acid Conjugated Bile AcidsStandards Preparation

Seven targeted amino acid conjugated bile acids standard solutions were prepared (see Table). 1.9 mg T1, 1.1 mg T2, 1.5 mg T3, 1.2 mg T4, 1.5 mg T5, 1.2 mg T6, and 1.3 mg T7 were all dissolved in 1 mL solvent (ACN:H2O=1:1, v:v), and then diluted 10 times in the same solvent. The prepared solutions were transported into glass vials for LC-MS analysis.

Table 3.1: Targeted amino acid conjugated bile acids

Abb.	Name
T1	Tryptophan conjugated cholic acid
T2	Serine conjugated cholic acid
T3	Glutamate conjugated cholic acid
T4	Alanine conjugated cholic acid
T5	Phenylalanine conjugated cholic acid
T6	Tyrosine conjugated cholic acid
T7	Leucine conjugated cholic acid

LC-MS Analysis

The LC-MS analysis was performed on Bruker Impact II™ UHR-QqTOF (Ultra-High Resolution Qq-Time-Of-Flight) mass spectrometer coupled with the Agilent 1290 Infinity™ II LC system. 2 µL seven standard solutions and 10 µL culture media sample solutions were injected in sequence onto a Waters ACQUITY UPLC BEH C18 Column (130Å, 1.7 µm, 1.0 mm X 100 mm). 2 µL NaFA was injected for internal mass calibration. The mobile phase A was H2O (0.1% Formic acid); mobile phase B was ACN (0.1% Formic acid). The chromatographic gradient was run at a flow rate of 0.150 mL/min as follows: 0-8 min: linear gradient from 95% to 75% A; 8-14 min: linear gradient from 75% to 30% A; 14-20 min: linear gradient from 30% to 5% A; 20-23 min: hold at 5% A; 23-23.01 min:linear gradient from 5% to 95% A; 23.01-30 min: hold at 95% A. The mass spectrometerwas operated in Auto MS/MS and positive mode. The ionization source capillary voltagewas set to 4.5 kV. The nebulizer gas pressure was set to 1.6 bar. The dry gas temperaturewas set to 220 °C. The collision energy for MS/MS was set to 7 eV. The data acquisitionwas performed in a range of 50-1200 m/z at a frequency of 8 Hz. Raw LC-MS data are publicly available on MetaboLights (www.ebi.ac.uk/metabolights/MTBLS5765).

Data Interpretation of targeted bile acid analysis

Bruker Data Analysis was used to calibrate the raw MS spectra and extract retention time, m/z, and intensity of the seven standard metabolites from their chromatograms. The extracted information was subsequently used as a reference to analyze the culture of media samples. We used the Bruker software TargetAnalysis to identify and relatively quantify the seven metabolites in all the culture media samples(Guo and Huan, 2020). The retention time, m/z, formula, and name of the seven metabolites wereregistered in the searching database of the software. The key searching parameter was set as follows: retention tolerance was 0.2-0.8 min; mass accuracy tolerance was 5-10 mDa; mSigma tolerance was 50-200. The chromatogram of each culture media sample was calibrated by internally injected NaFA (250 mM) before targeting the seven metabolites based upon retention time, m/z, and formula. The peak height and area of the corresponding identified metabolite were displayed on the result panel. In addition, the MS/MS spectra from the raw chromatogram were also manually validated for reassured identification.

Microbiome Analysis

QIIME 2 was used to calculate diversity metrics for both 16S rRNA gene amplicon (16S) and metagenomics data. Faith's phylogenetic diversity was calculated for each diet-genotype combination and compared with t-tests. For 16S data, the deblurred SEPP insertion tree was used with Greengenes 13_8 reference phylogeny. Shotgun metagenomic sequencing data were aligned to the Web of Life database of 10,575 reference bacterial and archaeal genomesusing the SHOGUN v1.0.7 pipeline (Zhu *et al.*, 2019) in the Bowtie2 alignment mode (Hillmann *et al.*, 2020). Non-unique alignments (i.e., where one read was simultaneously aligned to multiple reference genomes) were excluded. The frequencies of reads assigned to individual reference genomes were calculated. A feature table with columns as reference genomes (OGUs) (Zhu et al., 2022) and

rows as samples wasconstructed for downstream analysis. Unweighted UniFrac was used for both data types to compute beta diversity within and among groups.

Songbird was used to compute feature differential ranks for both diet and genotype (Morton *et al.*, 2019a; Morton *et al.*, 2019b). Differential ranks and log-ratios were visualized and calculated using Qurro(Wu et al., 2015). Statistical comparisons of sample log-ratios were performed using t-tests. The following parameters were used for the Songbird model: epochs=5000, batch size=30, differentialprior=0.5, learning rate=0.0005. The following formula was used:

diet * genotype + host_age + sex + treatment_of_drug

The regression model was compared to a "null" model with no covariates to ensure there was no overfitting. The phylogenetic tree of metagenomics features was created using Empress (https://journals.asm.org/doi/10.1128/mSystems.01216-20) Feature differentials were clipped to be centered around 0 and passed into Empress as feature metadata files.

Microbiome-Metabolome Association Analysis

Co-occurrence probabilities between microbes and metabolites were calculated using mmvec. The following MMvec parameters were used: input prior=1.0, output prior=0.5, batch size=300 (Morton *et al.*, 2019a). The input metagenomics feature table was subset to include only $APC^{min/+}$ mice samples. For the clustering analysis, the conditional ranks table was first subset to include only bile acids. This table was then Z-scored across all microbes and filtered to exclude any microbes that were highly co-occurrent (> 2.5 SD) with fewer than1% of bile acids. The network package in Python was used to create a bipartite graph ofthe resulting features into HFD/ND-associated bile acids and gOTUs.

Validation of AA-BAs synthesized by cecum whole bacteria cultured with bile acid substrates

Whole bacteria strains were harvested from the cecum of WT and $APC^{min/+}$ mice on both ND and HFD. Briefly, 3-5 mice in each group were sacrificed and the cecum pouch was opened in an anaerobic chamber. The cecum contents of the same group were pooled and washed with pre-reduced anaerobic transport media (ATM) (#AS-911, Anaerobe System Inc). The pooled bacteria pellets were divided and 1/5 were cultured on one plate of Yeast Casitone Fatty Acids Agar with Carbohydrates either without or with blood (YCFAC or YCFAC-B) plates (#AS-675, #AS-677, Anaerobe System Inc) in an anaerobic chamber or with oxygen. After 48-72 hrs culturing at 37 degrees, whole bacteria were harvested and pooled from 5 plates using ATM media and combined to represent one group of mice. 100 µL of combined cecal suspension was transferred to YCFAC broth with a gradient (from 1ng/ml to 5mg/ml) Cholic acid (CA) as a substrate. After 48-72 hrs culturing in 37 degrees, metabolites were extracted from the supernatant and subjected to novel AA-BAs detection by LC-MS/MS analysis (description of sample preparation seen below).

Validation of AA-BAs synthesized by single microbes cultured with bile acid

Bacterial pellets from *Ruminococcus gnavus* strain VPI C7-9 (#29149) and *Clostridium scindens* strain VPI 13733 (#35704), were purchased from ATCC and rehydrated in 0.5ml ATCC 260 broth medium (Tryptic Soy Broth (BD 211825) 30g, Sheep Blood 50ml, DI water 950ml) under anaerobic conditions. 100ul of resuspended culture was then plated on ATCC 260 Medium (Tryptic Soy Agar (BD 236950) 40g with 5% Sheep Blood (defibrinated) 50 ml, DI Water 950 ml). The remaining rehydrated bacterial culture was transferred to 5ml ATCC 260 broth medium. Lactobacillus acidophilus strain VPI 11091 (#9224) and Lactobacillus reuteri strain IDCC3701 (#BAA-2837), were purchased from ATCC and rehydrated in 0.5ml ATCC 416 MRS broth medium. 100µl of bacteria culture was then plated on ATCC 416 MRS agar. Ileibacterium valens strain NYU-BL-A3 (#TSD- 63) were purchased from ATCC and rehydrated in 0.5ml MTGE broth medium (anaerobe Systems). 100ul of bacteria culture was then plated on Brucella agar

supplemented with 5% Sheep blood, Vitamin K, and Hemin (anaerobe Systems). All cultures were incubated in an anaerobic atmosphere containing a gas mix of 5% hydrogen and 95% nitrogen at 37°C for 24-48 hours. All bacteria cultures were treated with 12 concentrations of cholic acid (gradient final concentrations ranging from 1ng/ml to 5mg/ml) in each specific culturing medium and cultured in total anaerobic conditions. After culturing for 48hrs, supernatants were collected for untargeted metabolomics. To extract metabolites, 200 µL supernatants were first mixed with 600 µL LC-MS grade ice-cold methanol. The solution was placed in a -20°C freezer for 2 hours to denature and precipitate the proteins. Further centrifugation (14000 rpm, 4°C, 15 min) removed the precipitated proteins and the supernatant was carefully transferred to a new vial. The solvent was evaporated in a Speedvac at 4°C. The dried sample was reconstituted in 200 µL solvent (ACN:H2O=1:3,v:v). The reconstituted sample was centrifuged (14000 rpm, 4°C, 15 min) again to removeany insoluble particles. The final solution was transferred into the LC glass insert for untargeted LC-MS/MS analysis in data-dependent acquisition mode on a Bruker Impact II Ultra-High Resolution Qq-Time-Of-Flight Mass Spectrometer (UHR-QqTOF-MS) coupled with an Agilent 1290 Infinity II Ultra-High-Performance Liquid Chromatography (UHPLC) system.

PCR amplification of 16s rRNA

The presence of specific bacteria 16S rRNA genes in sample materials was first determined using a nested-PCR approach. Briefly, a community's 16S rRNA genes were amplified using universal bacterial primers (Supplementary Table AB.2.S1) and 20 to 30 ng of community DNA as template. Following amplification, 2 µl of PCR product was analyzed by agarose gel electrophoresis to verify that 16S rRNA genes were amplified from the community DNA. Then, 2 µl of 1:2 and 1:50 dilutions of the 16S rRNA gene amplicons were used as templates in a second round of PCR with species-specific bacteria 16S rRNA gene-specific primer pair.

3.6. Acknowledgements

We thank L. Ong and C. Brondos for administrative assistance. This work was funded by grants from the NIH (DK057978, HL105278 and HL088093), National Cancer Institute (CA014195), the Leona M. and Harry B. Helmsley Charitable Trust (2017PG-MED001), SWCRF Investigator Award and Ipsen/Biomeasure to R.M.E; National Health and Medical Research Council of Australia Project grants (1087297) to C.L. and M.D.; UCSDPostdoc Microbiome Center Seed Pilot Grant to T.F. T.F. is supported by Hewitt MedicalFoundation Fellowship, a Salk Alumni Fellowship and Crohn's & Colitis Foundation (CCFA) Visiting IBD Research Fellowship. R.M.E. and M.D. are supported, in part, by a Stand Up To Cancer-Cancer Research UK-Lustgarten Foundation Pancreatic Cancer Dream Team Research Grant (Grant Number: SU2C-AACR-DT-20-16). Stand Up To Cancer is a program of the Entertainment Industry Foundation. Research grants are administered by the American Association for Cancer Research, the scientific partner of SU2C. R.M.E is an investigator of the Howard Hughes Medical Institute and March of Dimes Chair in Molecular and Developmental Biology at the Salk Institute.

Chapter 3 has been submitted for publication of the materials as it may appear in *Cell Reports*, “Paired microbiome and metabolome analyses associate bile acid changes with colorectal cancer progression” Ting Fu,, Tao Huan, Gibraan Rahman, Hui Zhi, Zhenjiang Xu, Tae Gyu Oh, Jian Guo, Sally Coulter, Anupriya Tripathi, Cameron Martino, Justin L McCarville, Qiyun Zhu, Fritz Cayabyab, Mingxiao He, Shipei Xing, Ruth T. Yu, Annette Atkins, Christopher Liddle, Janelle Ayres, Manuela Raffatellu, Pieter C. Dorrestein, Michael Downes, Rob Knight3, and Ronald M. Evans. The dissertation author is the co-first author of this paper in conjuction with Dr. Ting Fu and Dr. Tao Huan.

3.7. Author Contributions

T.F., M.D., R.K. and R.M.E. designed and supervised the research. T.F., T.H., G.R. performed majority of the experiments and analyzed results, with technical assistance from H.Z., Z.X., T.G.O., J.G., A.T., F.C., and M.H. C.L. and S.C. performed bile acid analyses, T.G.O, C.L., and R.T.Y. analyzed mice genomic data. T.H., P.C.D., J.G., A.T. and S.X. conducted untargeted metabolites profiling in serum and cecum samples. M.R.and H.Z. assist anaerobic bacterial culture assay. R.K., G.R., Z.X., C.M., Q.Z. and T.R. performed gut microbiome 16S rRNA gene amplicon and shotgun metagenomics sequencing, and metabolite-microbiome association studies. J.M., and J.A. provided microbial culture facilities and scientific input. T.F., T.H., G.R., A.R.A., M.D., R.K. and R.M.E. prepared the manuscript.

3.8. References

1. Buffie, C.G., Bucci, V., Stein, R.R., McKenney, P.T., Ling, L., Gobourne, A., No, D., Liu, H., Kinnebrew, M., Viale, A., et al. (2015). Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile. Nature *517*, 205-208. 10.1038/nature13828.

2. Chong, J., Soufan, O., Li, C., Caraus, I., Li, S., Bourque, G., Wishart, D.S., and Xia, J. (2018). MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. Nucleic Acids Res *46*, W486-W494. 10.1093/nar/gky310.

3. Cox, L.M., Sohn, J., Tyrrell, K.L., Citron, D.M., Lawson, P.A., Patel, N.B., Iizumi, T., Perez-Perez, G.I., Goldstein, E.J.C., and Blaser, M.J. (2017). Description of two novel members of the family Erysipelotrichaceae: Ileibacterium valens gen. nov., sp. nov. and Dubosiella newyorkensis, gen. nov., sp. nov., from the murine intestine, and emendation to the description of Faecalibaculum rodentium. Int J Syst Evol Microbiol *67*, 1247-1254. 10.1099/ijsem.0.001793.

4. Degirolamo, C., Modica, S., Palasciano, G., and Moschetta, A. (2011). Bile acids and colon cancer: Solving the puzzle with nuclear receptors. Trends Mol Med *17*, 564-572. 10.1016/j.molmed.2011.05.010.

5. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21. 10.1093/bioinformatics/bts635.

6. Farshidfar, F., Kopciuk, K.A., Hilsden, R., McGregor, S.E., Mazurak, V.C., Buie, W.D., MacLean, A., Vogel, H.J., and Bathe, O.F. (2018). A quantitative multimodal metabolomic assay for colorectal cancer. BMC Cancer *18*, 26. 10.1186/s12885-017-3923-z.

7. Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., Zhang, D., Xia, H., Xu, X., Jie, Z., et al. (2015). Gut microbiome development along the colorectal adenoma-carcinoma sequence. Nat Commun *6*, 6528. 10.1038/ncomms7528.

8. Ferlay, J., Colombet, M., Soerjomataram, I., Mathers, C., Parkin, D.M., Pineros, M., Znaor, A., and Bray, F. (2019). Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. Int J Cancer *144*, 1941-1953. 10.1002/ijc.31937.

9. Fiorucci, S., and Distrutti, E. (2015). Bile Acid-Activated Receptors, Intestinal Microbiota, and the Treatment of Metabolic Disorders. Trends Mol Med *21*, 702-714. 10.1016/j.molmed.2015.09.001.

10. Forsberg, E.M., Huan, T., Rinehart, D., Benton, H.P., Warth, B., Hilmers, B., and Siuzdak, G. (2018). Data processing, multi-omic pathway mapping, and metabolite activity analysis using XCMS Online. Nat Protoc *13*, 633-651. 10.1038/nprot.2017.151.

11. Friedman, E.S., Li, Y., Shen, T.D., Jiang, J., Chau, L., Adorini, L., Babakhani, F., Edwards, J., Shapiro, D., Zhao, C., et al. (2018). FXR-Dependent Modulation of the Human Small Intestinal Microbiome by the Bile Acid Derivative Obeticholic Acid. Gastroenterology *155*, 1741-1752 e1745. 10.1053/j.gastro.2018.08.022.

12. Fu, T., Coulter, S., Yoshihara, E., Oh, T.G., Fang, S., Cayabyab, F., Zhu, Q., Zhang, T., Leblanc, M., Liu, S., et al. (2019). FXR Regulates Intestinal Cancer Stem Cell Proliferation. Cell *176*, 1098-1112 e1018. 10.1016/j.cell.2019.01.036.

13. Gill, C.I., and Rowland, I.R. (2002). Diet and cancer: assessing the risk. Br J Nutr *88 Suppl 1*, S73-87. 10.1079/BJN2002632.

14. Guo, J., and Huan, T. (2020). Comparison of Full-Scan, Data-Dependent, and Data-Independent Acquisition Modes in Liquid Chromatography-Mass Spectrometry Based Untargeted Metabolomics. Anal Chem *92*, 8072-8080. 10.1021/acs.analchem.9b05135.

15. Henke, M.T., Kenny, D.J., Cassilly, C.D., Vlamakis, H., Xavier, R.J., and Clardy, J. (2019). Ruminococcus gnavus, a member of the human gut microbiome associated with Crohn's disease, produces an inflammatory polysaccharide. Proc Natl Acad Sci U S A *116*, 12672-12677. 10.1073/pnas.1904099116.

16. Hillmann, B., Al-Ghalith, G.A., Shields-Cutler, R.R., Zhu, Q., Knight, R., and Knights, D. (2020). SHOGUN: a modular, accurate and scalable framework for microbiome quantification. Bioinformatics *36*, 4088-4090. 10.1093/bioinformatics/btaa277.

17. Hoffmann, M.A., Nothias, L.F., Ludwig, M., Fleischauer, M., Gentry, E.C., Witting, M., Dorrestein, P.C., Duhrkop, K., and Bocker, S. (2022). High-confidence structural annotation of metabolites absent from spectral libraries. Nat Biotechnol *40*, 411-421. 10.1038/s41587-021-01045-9.

18. Huan, T., Forsberg, E.M., Rinehart, D., Johnson, C.H., Ivanisevic, J., Benton, H.P., Fang, M., Aisporna, A., Hilmers, B., Poole, F.L., et al. (2017). Systems biology guided by XCMS Online metabolomics. Nat Methods *14*, 461-462. 10.1038/nmeth.4260.

19. Islami, F., Goding Sauer, A., Miller, K.D., Siegel, R.L., Fedewa, S.A., Jacobs, E.J., McCullough, M.L., Patel, A.V., Ma, J., Soerjomataram, I., et al. (2018). Proportion and number of cancer cases and deaths attributable to potentially modifiable risk factors in the United States. CA Cancer J Clin *68*, 31-54. 10.3322/caac.21440.

20. Jing, Y., Wu, X., Gao, P., Fang, Z., Wu, J., Wang, Q., Li, C., Zhu, Z., and Cao, Y. (2017). Rapid differentiating colorectal cancer and colorectal polyp using dried blood spot mass spectrometry metabolomic approach. IUBMB Life *69*, 347-354. 10.1002/iub.1617.

21. Kadosh, E., Snir-Alkalay, I., Venkatachalam, A., May, S., Lasry, A., Elyada, E., Zinger, A., Shaham, M., Vaalani, G., Mernberger, M., et al. (2020). The gut microbiome switches mutant p53 from tumour-suppressive to oncogenic. Nature *586*, 133-138. 10.1038/s41586-020-2541-0.

22. Kakiyama, G., Muto, A., Takei, H., Nittono, H., Murai, T., Kurosawa, T., Hofmann, A.F., Pandak, W.M., and Bajaj, J.S. (2014). A simple and accurate HPLC method for fecal bile acid profile in healthy and cirrhotic subjects: validation by GC-MS and LC-MS. J Lipid Res *55*, 978-990. 10.1194/jlr.D047506.

23. Koppel, N., and Balskus, E.P. (2016). Exploring and Understanding the Biochemical Diversity of the Human Microbiota. Cell Chem Biol *23*, 18-30. 10.1016/j.chembiol.2015.12.008.

24. Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323. 10.1186/1471-2105-12-323.

25. McCarville, J.L., Chen, G.Y., Cuevas, V.D., Troha, K., and Ayres, J.S. (2020). Microbiota Metabolites in Health and Disease. Annu Rev Immunol *38*, 147-170. 10.1146/annurev-immunol-071219-125715.

26. Morton, J.T., Aksenov, A.A., Nothias, L.F., Foulds, J.R., Quinn, R.A., Badri, M.H., Swenson, T.L., Van Goethem, M.W., Northen, T.R., Vazquez-Baeza, Y., et al. (2019a). Learning representations of microbe-metabolite interactions. Nat Methods *16*, 1306-1314. 10.1038/s41592-019-0616-3.

27. Morton, J.T., Marotz, C., Washburne, A., Silverman, J., Zaramela, L.S., Edlund, A., Zengler, K., and Knight, R. (2019b). Establishing microbial composition measurement standards with reference frames. Nat Commun *10*, 2719. 10.1038/s41467-019-10656-5.

28. Nakatsu, G., Li, X., Zhou, H., Sheng, J., Wong, S.H., Wu, W.K., Ng, S.C., Tsoi, H., Dong, Y., Zhang, N., et al. (2015). Gut mucosal microbiome across stages of colorectal carcinogenesis. Nat Commun *6*, 8727. 10.1038/ncomms9727.

29. Nothias, L.F., Petras, D., Schmid, R., Duhrkop, K., Rainer, J., Sarvepalli, A., Protsyuk, I., Ernst, M., Tsugawa, H., Fleischauer, M., et al. (2020). Feature-based molecular networking in the GNPS analysis environment. Nat Methods *17*, 905-908. 10.1038/s41592-020-0933-6.

30. Petras, D., Caraballo-Rodriguez, A.M., Jarmusch, A.K., Molina-Santiago, C., Gauglitz, J.M., Gentry, E.C., Belda-Ferre, P., Romero, D., Tsunoda, S.M., Dorrestein, P.C., and Wang, M. (2021). Chemical Proportionality within Molecular Networks. Anal Chem *93*, 12833-12839. 10.1021/acs.analchem.1c01520.

31. Powell, S.M., Zilz, N., Beazer-Barclay, Y., Bryan, T.M., Hamilton, S.R., Thibodeau, S.N., Vogelstein, B., and Kinzler, K.W. (1992). APC mutations occur early during colorectal tumorigenesis. Nature *359*, 235-237. 10.1038/359235a0.

32. Quinn, R.A., Melnik, A.V., Vrbanac, A., Fu, T., Patras, K.A., Christy, M.P., Bodai, Z., Belda-Ferre, P., Tripathi, A., Chung, L.K., et al. (2020). Global chemical effects of the microbiome include new bile-acid conjugations. Nature *579*, 123-129. 10.1038/s41586-020-2047-9.

33. Scott, A.J., Alexander, J.L., Merrifield, C.A., Cunningham, D., Jobin, C., Brown, R., Alverdy, J., O'Keefe, S.J., Gaskins, H.R., Teare, J., et al. (2019). International Cancer Microbiome Consortium consensus statement on the role of the human microbiome in carcinogenesis. Gut *68*, 1624-1632. 10.1136/gutjnl-2019-318556.

34. Song, M., and Chan, A.T. (2019). Environmental Factors, Gut Microbiota, and Colorectal Cancer Prevention. Clin Gastroenterol Hepatol *17*, 275-289. 10.1016/j.cgh.2018.07.012.

35. Sorrentino, G., Perino, A., Yildiz, E., El Alam, G., Bou Sleiman, M., Gioiello, A., Pellicciari, R., and Schoonjans, K. (2020). Bile Acids Signal via TGR5 to Activate Intestinal Stem Cells and Epithelial Regeneration. Gastroenterology *159*, 956-968 e958. 10.1053/j.gastro.2020.05.067.

36. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A *102*, 15545-15550. 10.1073/pnas.0506580102.

37. Thomas, A.M., Manghi, P., Asnicar, F., Pasolli, E., Armanini, F., Zolfo, M., Beghini, F., Manara, S., Karcher, N., Pozzi, C., et al. (2019). Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nat Med *25*, 667-678. 10.1038/s41591-019-0405-7.

38. Tripathi, A., Vazquez-Baeza, Y., Gauglitz, J.M., Wang, M., Duhrkop, K., Nothias-Esposito, M., Acharya, D.D., Ernst, M., van der Hooft, J.J.J., Zhu, Q., et al. (2020). Chemically informed analyses of metabolomics mass spectrometry data with Qemistree. Nat Chem Biol. 10.1038/s41589-020-00677-3.

39. Tsugawa, H., Cajka, T., Kind, T., Ma, Y., Higgins, B., Ikeda, K., Kanazawa, M., VanderGheynst, J., Fiehn, O., and Arita, M. (2015). MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. Nat Methods *12*, 523-526. 10.1038/nmeth.3393.

40. Wahlstrom, A., Sayin, S.I., Marschall, H.U., and Backhed, F. (2016). Intestinal Crosstalk between Bile Acids and Microbiota and Its Impact on Host Metabolism. Cell Metab *24*, 41-50. 10.1016/j.cmet.2016.05.005.

41. Wang, M., Carver, J.J., Phelan, V.V., Sanchez, L.M., Garg, N., Peng, Y., Nguyen, D.D., Watrous, J., Kapono, C.A., Luzzatto-Knaan, T., et al. (2016). Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol *34*, 828-837. 10.1038/nbt.3597.

42. Watrous, J., Roach, P., Alexandrov, T., Heath, B.S., Yang, J.Y., Kersten, R.D., van der Voort, M., Pogliano, K., Gross, H., Raaijmakers, J.M., et al. (2012). Mass spectral molecular networking of living microbial colonies. Proc Natl Acad Sci U S A *109*, E1743-1752. 10.1073/pnas.1203689109.

43. Wirbel, J., Pyl, P.T., Kartal, E., Zych, K., Kashani, A., Milanese, A., Fleck, J.S., Voigt, A.Y., Palleja, A., Ponnudurai, R., et al. (2019). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat Med *25*, 679-689. 10.1038/s41591-019-0406-6.

44. Wong, S.H., and Yu, J. (2019). Gut microbiota in colorectal cancer: mechanisms of action and clinical applications. Nat Rev Gastroenterol Hepatol *16*, 690-704. 10.1038/s41575-019-0209-8.

45. Wu, M., McNulty, N.P., Rodionov, D.A., Khoroshkin, M.S., Griffin, N.W., Cheng, J., Latreille, P., Kerstetter, R.A., Terrapon, N., Henrissat, B., et al. (2015). Genetic determinants of in vivo fitness and diet responsiveness in multiple human gut Bacteroides. Science *350*, aac5992. 10.1126/science.aac5992.

46. Yachida, S., Mizutani, S., Shiroma, H., Shiba, S., Nakajima, T., Sakamoto, T., Watanabe, H., Masuda, K., Nishimoto, Y., Kubo, M., et al. (2019). Metagenomic and metabolomic

analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nat Med *25*, 968-976. 10.1038/s41591-019-0458-7.

47. Yu, H., Chen, Y., and Huan, T. (2021). Computational Variation: An Underinvestigated Quantitative Variability Caused by Automated Data Processing in Untargeted Metabolomics. Anal Chem. 10.1021/acs.analchem.0c03381.

48. Zhao, Z., Xiao, Y., Elson, P., Tan, H., Plummer, S.J., Berk, M., Aung, P.P., Lavery, I.C., Achkar, J.P., Li, L., et al. (2007). Plasma lysophosphatidylcholine levels: potential biomarkers for colorectal cancer. J Clin Oncol *25*, 2696-2701. 10.1200/JCO.2006.08.5571.

49. Zhu, Q., Huang, S., Gonzalez, A., McGrath, I., McDonald, D., Haiminen, N., Armstrong, G., Vazquez-Baeza, Y., Yu, J., Kuczynski, J., et al. (2022). Phylogeny-Aware Analysis of Metagenome Community Ecology Based on Matched Reference Genomes while Bypassing Taxonomy. mSystems *7*, e0016722. 10.1128/msystems.00167-22.

50. Zhu, Q., Mai, U., Pfeiffer, W., Janssen, S., Asnicar, F., Sanders, J.G., Belda-Ferre, P., Al-Ghalith, G.A., Kopylova, E., McDonald, D., et al. (2019). Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. Nat Commun *10*, 5477. 10.1038/s41467-019-13443-4.

Chapter 4. BIRDMAn: A Bayesian differential abundance framework that enables robust inference of host-microbe associations

Abstract

Quantifying the differential abundance (DA) of specific taxa among experimental groups in microbiome studies is challenging due to data characteristics (e.g., compositionality, sparsity) and specific study designs (e.g., repeated measures, meta-analysis, cross-over). Here we present BIRDMAn (**B**ayesian **I**nferential **R**egression for **D**ifferential **M**icrobiome **An**alysis), a flexible DA method that can account for microbiome data characteristics and diverse experimental designs. Simulations show that BIRDMAn models are robust to uneven sequencing depth and provide a >20-fold improvement in statistical power over existing methods. We then use BIRDMAn to identify antibiotic-mediated perturbations undetected by other DA methods due to subject-level heterogeneity. Finally, we demonstrate how BIRDMAn can construct state-of-the-art cancer-type classifiers using The Cancer Genome Atlas (TCGA) dataset, with substantial accuracy improvements over random forests and existing DA tools across multiple sequencing centers. Collectively, BIRDMAn extracts more informative biological signals while accounting for study-specific experimental conditions than existing approaches.

4.1 Introduction

Advances in sequencing technology and computational methods have enabled researchers to experimentally characterize microbiomes across wide ranges of biological conditions, including psychiatric diseases[1,2], cancer[3,4], and COVID-19[5,6]. However, as the understanding of microbial effects on human health and disease has increased, the experimental

questions, hypotheses, and concomitant statistics have grown in complexity, with study designs now commonly involving longitudinal analyses[7–9], experimental interventions[10–12], and meta-analyses[7]. Although such approaches can provide mechanistic insights into the microbiome's effect(s) on the host, their conclusions are often limited by the ability to perform valid statistical analyses that are sufficiently flexible to account for the added experimental complexity.

One common but critical challenge in these contexts is when population-level heterogeneity (such as subject-to-subject variation) is confounded by technical variability. For example, samples originating from the same sequencing center will tend to be more similar to each other than those sequenced from different centers[13]. The confounding factors that may explain these differences make it difficult to determine consistent microbial biomarkers associated with biological variables or conditions of interest[8]—an effect compounded by other microbiome data difficulties, such as high sparsity, high-dimensionality, and compositionality. Moreover, statistical tools that can properly assess and account for strong structural effects while still indicating which microbes truly vary between biological conditions are limited to date[15].

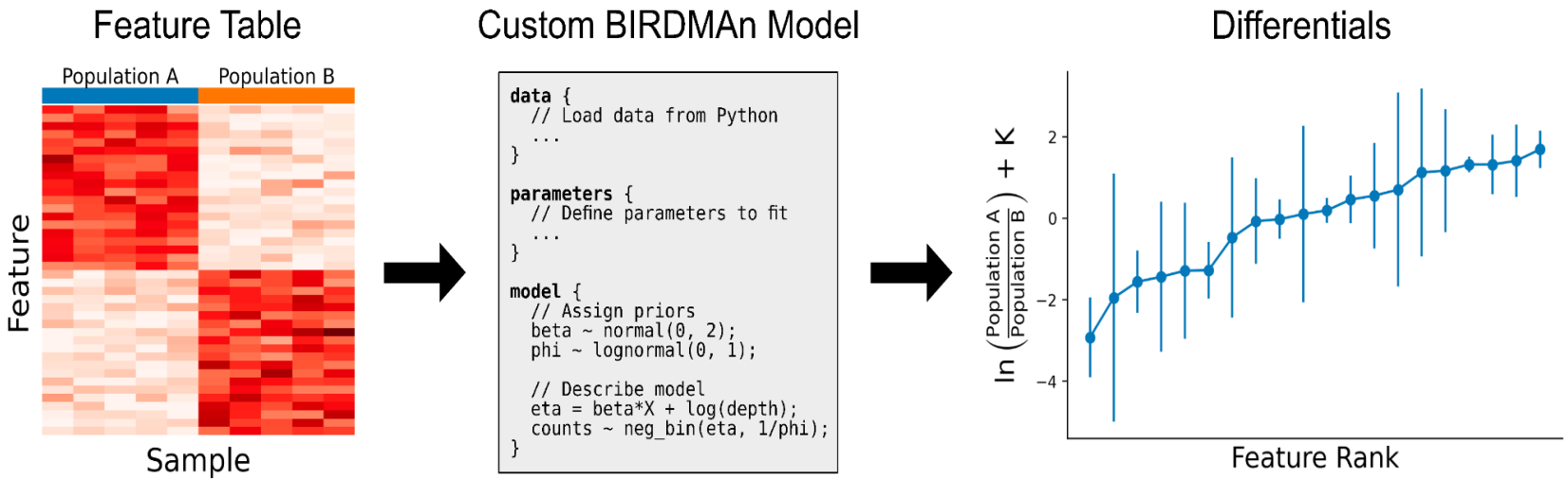
Making matters more difficult, disagreement exists about how to benchmark differential abundance (DA) tools and methods. Previous efforts have commonly focused on comparing the results of hypothesis testing while accounting for the multiplicity of features through false-discovery-rate (FDR) correction[15–17]. Studies have demonstrated that tools designed for differential abundance often report contradictory results with different microbial abundances among biologically distinct sampling groups[19].

Addressing these challenges requires a more robust statistical framework for benchmarking differential abundance methods and would benefit from flexible DA modeling approaches. Thus, we developed BIRDMAn (**B**ayesian **I**nferential **R**egression for **D**ifferential **M**icrobiome **An**alysis), a flexible computational framework for hierarchical Bayesian modeling of

microbiome data that simultaneously accounts for its high sparsity, high-dimensionality, and compositionality.

The Bayesian approach to statistical modeling provides unique advantages compared to frequentist solutions, such as the inclusion of prior information, uncertainty estimation of parameters, native hierarchical modeling, and edge case smoothing (e.g., estimating log fold changes when a feature is only present in one group). Implemented within the Stan programming language (commonly used for designing probabilistic models), BIRDMAn flexibly enables parameter estimation of all biological variables and non-biological covariates. These advantages allow us to demonstrate how explicitly modeling population-level effects in probabilistic BIRDMAn models increases the amount of true biological signal recovered compared to existing tools on both simulated and real-world datasets. Moreover, the BIRDMAn workflow significantly lowers the barrier of entry for differential abundance methods development and implementation. Additionally, to address reproducibility issues of prior DA tool benchmarking, we present a novel approach that employs techniques from compositional data analysis, making the comparison of tools more interpretable and statistically valid.

Figure 4.1: Overview of BIRDMAn workflow for customizable differential abundance analysis. A table of counts by features is modeled using Bayesian probabilistic programming, resulting in credible intervals of the estimated parameter posterior distributions. The statistical model can be customized using the Stan probabilistic programming language and fit using the BIRDMAn Python interface.


Feature Table
Population A
Population B
Feature
Sample
Custom BIRDMAn Model
data {
// Load data from Python
...
}
parameters {
// Define parameters to fit
...
}
model {
// Assign priors
beta ~ normal(0, 2);
phi ~ lognormal(0, 1);
// Describe model
eta = beta*X + log(depth);
counts ~ neg_bin(eta, 1/phi);
}
Differentials
ln(Population A / Population B) + K
2
0
−2
−4
Feature Rank

4.2 Results

BIRDMAn is implemented as a Python interface to the Stan probabilistic programming language, which utilizes Hamiltonian Monte Carlo sampling, one of the state-of-the-art approaches for Bayesian uncertainty estimation[20]. Users can employ pre-configured model designs or flexibly customize inputs to account for their specific experimental design and biological questions; BIRDMAn then fits and processes these models (Fig 4.1). The results of these analyses are the posterior distributions of the defined parameters of interest, such as log-fold changes and their uncertainty given the data (see Methods).

To showcase the statistical properties of BIRDMAn models, we first leverage simulations to evaluate the accuracy of estimating differential uncertainty in the context of realistic biological scenarios. Then, we apply BIRDMAn models on real-world data, demonstrating superiority for resolving subject-level heterogeneity in an antibiotics experiment, as well as alleviating sequencing center-specific effects in a cancer genomics dataset, each while capturing biologically-informative signals.

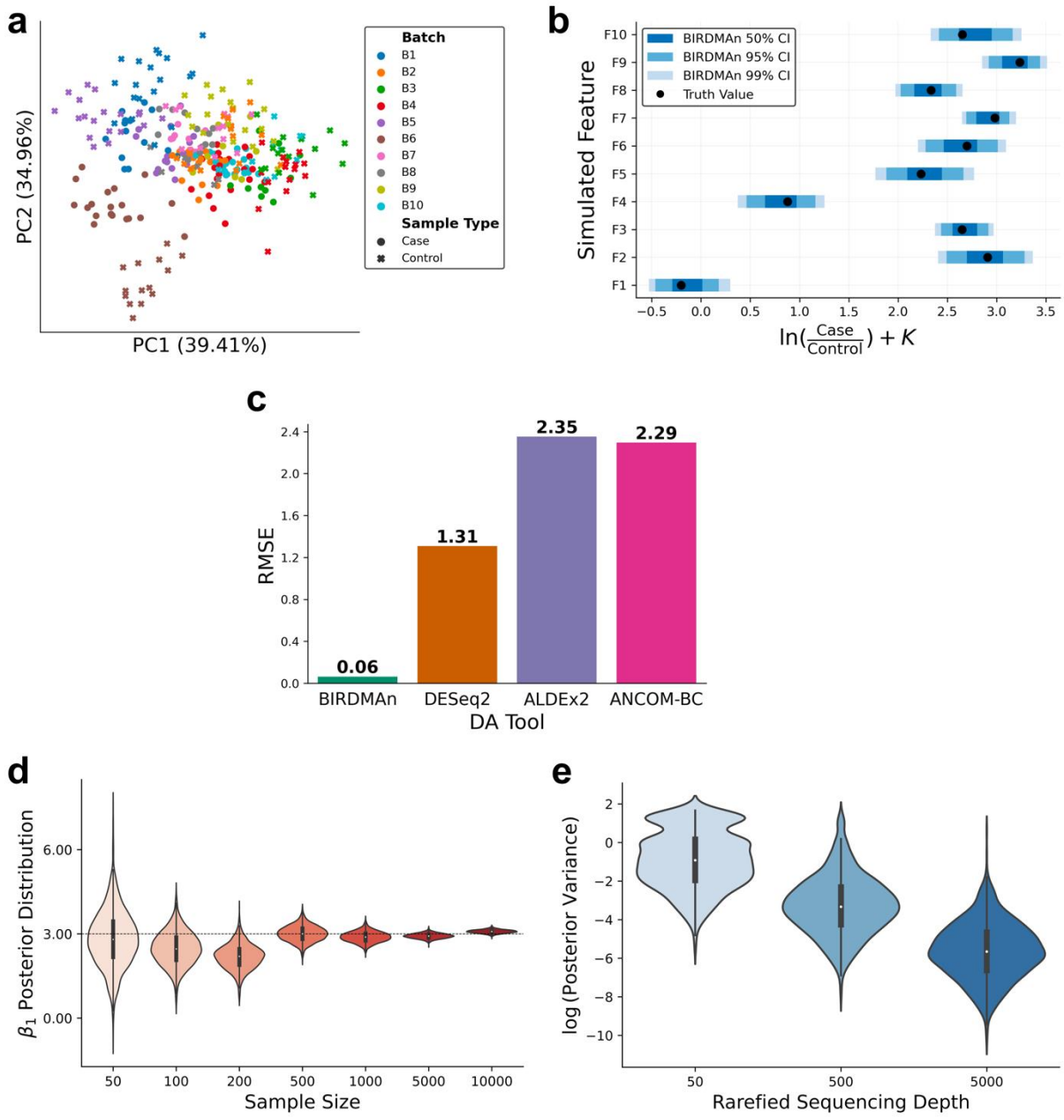
Simulations demonstrate BIRDMAn model accuracy and precision

A common difficulty in benchmarking differential abundance methods is the lack of ground truth. We typically do not know which microbial taxa are truly increasing or decreasing across experimental conditions. To gain insights into the robustness of BIRDMAn models, we performed a data-driven simulation of a case-control microbiome dataset with one binary covariate, large batch effects (10 features, 10 batches, and 300 samples), data overdispersion, and known differentials associated with case status (see Methods) (Fig 4.2a). We then used BIRDMAn to estimate the model parameters for each feature and compared the Bayesian posterior estimates with the true value, finding that BIRDMAn models recovered the ground truth differentials with high accuracy and precision (Fig 4.2b) while outperforming other tools in terms of root mean

square error (RMSE) (Fig 4.2c). This highlights how BIRDMAn model customization permits more accurate estimations of differentials.

One advantage of Bayesian models is that they can leverage posterior estimates to summarize the uncertainty of these differentials, taking into account the sample size and the sequencing depth. As expected, we show that when BIRDMAn models are fitted on larger sample sizes, the uncertainty decreases, highlighting how incorporating more data, and avoiding rarefaction, enables a more accurate estimation of the differentials (Fig 4.2d). Furthermore, we show that decreasing the sequencing depth also increases the uncertainty, highlighting how rarefaction could degrade parameter estimates' precisions in BIRDMAn models (Fig 4.2e). Since BIRDMAn can handle variable sequencing depths, there is no need to perform rarefaction before model fitting, which is desirable when analyzing microbiome datasets[21].

Figure 4.2: Benchmarking differential abundance methods on simulated data (a) Robust Aitchison principal components plot of the simulated data, showing the large separation by batch effect. Simulations of 10 batches (B1 to B10) of microbiome results, each containing 10 features (F1 to F10), where each feature has a true differential abundance between cases and controls that is the same for each batch, and also a random per batch bias. (b) Recovery of the true simulated log ratio between cases and controls for each feature (black dots), with credible intervals on average centered on the true log ratio (blue bars). (c) Superior performance of BIRDMAn over other differential abundance methods in minimizing the RMSE of the difference between the estimated mean posterior log ratio between cases and controls, revealing a >20-fold improvement in RMSE over the nearest competitor, DESeq2. (d) Estimated distributions of log-fold changes from Bayesian analysis tighten as the number of samples increases. Dashed line represents the true simulated value for each simulation. (e) Rarefaction simulation performed using multinomial count generative models (1000 features) at three different sequencing depths shows that the variance of the posterior distribution decreases as depth increases.


a
PC2 (34.96%)
PC1 (39.41%)
Batch
B1
B2
B3
B4
B5
B6
B7
B8
B9
B10
Sample Type
Case
Control
b
Simulated Feature
BIRDMAn 50% CI
BIRDMAn 95% CI
BIRDMAn 99% CI
Truth Value
F10
F9
F8
F7
F6
F5
F4
F3
F2
F1
−0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
ln(Case/Control) + K
c
RMSE
2.4
2.0
1.6
1.2
0.8
0.4
0.0
0.06
1.31
2.35
2.29
BIRDMAn
DESeq2
ALDEx2
ANCOM-BC
DA Tool
d
β1 Posterior Distribution
6.00
3.00
0.00
50 100 200 500 1000 5000 10000
Sample Size
e
log (Posterior Variance)
2
0
−2
−4
−6
−8
−10
50 500 5000
Rarefied Sequencing Depth

BIRDMAn models capture biological signals missed by other methods during dual-course longitudinal antibiotics

Another challenge for DA methods is to compare multiple samples from the same subject longitudinally (repeated measures) since concomitant host-specific variation can obscure phenotypically-associated microbial changes. Methods designed for longitudinal data[22–26] cannot easily account for modeling perturbations and struggle with scaling to high dimensions. To demonstrate the use of BIRDMAn on repeated measure study designs, we evaluated a published longitudinal study of two courses of the antibiotic ciprofloxacin (Cp) (3 subjects, 7 timepoints)[27]. Notably, this study originally concluded that inter-subject variability drove the response to antibiotics by examining beta-diversities, which do not account for auto-correlation effects of repeated measures[28] (Fig 4.3a). Other studies have also highlighted the importance of properly accounting for the microbial community composition prior to antibiotics when assessing varying responses[29,30], which requires accurate temporal modeling.

Given BIRDMAn's flexibility, we constructed a customized DA model that leverages Linear Mixed Effects models, accounting for repeated measurements from subjects while computing temporal differences (see Methods). This model design then enabled the exploration of common microbial community changes associated with antibiotic perturbation, which the originally published methods could not identify. With the computed log-fold changes over time (Supp Fig AC.1.S1a), we investigated how consistent antibiotic induced shifts were across subjects. For each temporal difference, we took the top and bottom 40 OTUs to calculate sample log-ratios, which were used to predict antibiotics intake[31]. From these log-ratios, we observed strong, statistically significant temporal shifts associated with each successive time interval (Supp Fig AC.1.S1b).

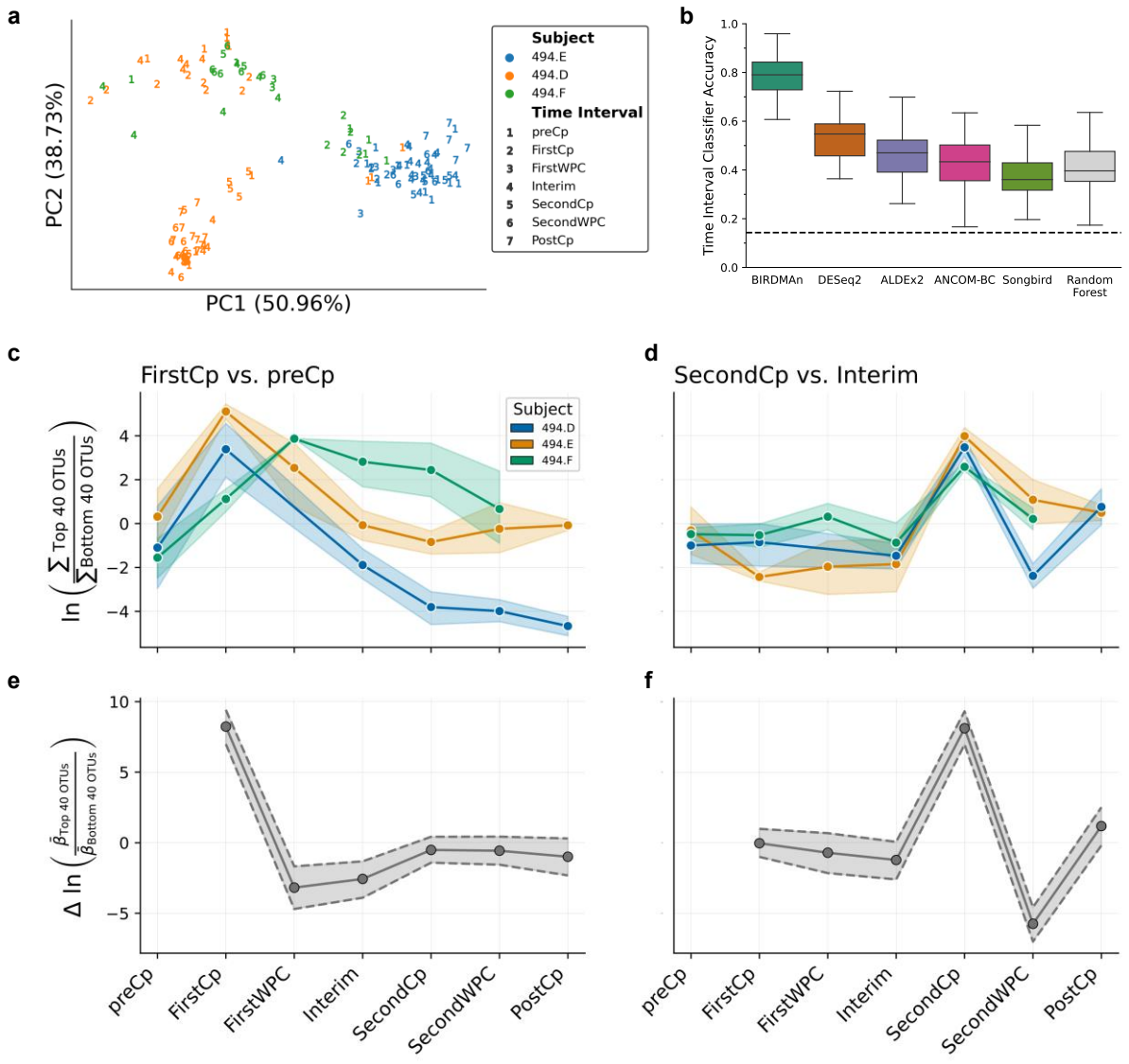
To determine if existing tools could have identified these timepoint-specific perturbations, we also developed a multinomial logistic regression classifier based on the BIRDMAn results to predict the corresponding time interval. We then compared our prediction performances against classifiers built using ALDEx2[32], ANCOM-BC[33], Songbird[31], and DESeq2[34] results on the same samples, as well as a classifier built on the center log-ratio transformed table (see Methods). Remarkably, BIRDMAn-informed classifiers were able to accurately differentiate between the different treatment groups (accuracy > 0.65) (Supp Fig AC.1.S1c) and showed substantially better prediction accuracy compared to all other methods (Fig 4.3b). We also verified that this superior performance held across varying numbers of OTUs used in log-ratio calculation (Supp Fig AC.1.S1d). Ultimately, these findings show how BIRDMAn can identify clear-cut biological changes that were missed or obscured by other approaches, highlighting its ability to confirm expected biological hypotheses.

We used the sample log-ratios associated with the First and Second Cp applications and plotted the dynamics over time (Fig 4.3c, d). Accordingly, we plotted the corresponding derivative log-fold changes computed from BIRDMAn (Fig 4.3e, f) and see that our trajectories match between the sample log-ratios and the estimated log-fold changes, indicating that our model was able to successfully capture the overall signal independent of subject.

The antibiotic used in the original work, Cp, is known to primarily target (though not exclusively) gram negative bacteria[35,36]. We thus hypothesized that the differential abundance results should reflect the longitudinal dynamics of gram negative bacterial abundance. In the top and bottom 40 most changed taxa after FirstCp, 17.5% of the numerator taxa were gram negative, whereas 27.5% of the denominator were gram negative (Supp Fig AC.1.S2e). Given the Cp antibiotic mechanism, it is likely that gram negative taxa in the denominator decreased which caused the increased log-ratio[37,38] (Fig. 4.2c). We see that there is a sharp decrease in this log-ratio at FirstWPC, which could be attributed to gut homeostasis[37,38]. However, we see a weaker

pattern in the top/bottom 40 microbes after SecondCp, where 2.5% of the numerator taxa were gram negative and 10% of the denominator taxa were gram negative. In contrast to the FirstCp, the microbes most affected by SecondCp quickly returned to their original abundances. Furthermore, we see that the microbes most altered by FirstCp were not affected by SecondCp. Altogether this hints at newly acquired antimicrobial resistant genes after the application of FirstCp.

Figure 4.3: Differential abundance analysis on dual course antibiotics dataset. (a) Robust Aitchison principal components plot of full dataset shows samples cluster primarily by host subject. (b) Balanced accuracy of multinomial classification of time point by tool. Differential abundance classifiers were constructed using logistic regression with the log-ratios of the top 40 and bottom 40 OTUs associated with each timepoint as predictors. Repeated k-fold cross-validation was performed with 5 splits and 10 repeats. The mean classifier error is at least twice as great with all other differential abundance tools as with BIRDMAn. Dashed line represents random guessing performance among the seven timepoints. (c, d) Dynamics of sample log-ratios of (c) first Cp course and (d) second Cp course colored by subject. (e, f) Dynamics of BIRDMAn-estimated log-fold changes associated with (e) FirstCp effect with preCp as reference and (f) SecondCp effect with Interim as reference. Shaded intervals represent the 90% credible interval of the estimated posterior distributions.



BIRDMAn models mitigate batch effects in cancer microbiome data

To investigate how generalizable BIRDMAn models are with respect to population heterogeneity, we conducted a meta-analysis using cancer microbiome data derived from The Cancer Genome Atlas (TCGA). This dataset is known to have large structural batch effects[4], where the samples were processed at multiple centers across North America, resulting in an artificial separation of cancer microbiomes by sequencing center if not otherwise accounted for (Fig 4.4a, Supp Fig AC.1.S2a)[4,39]. These effects can make it difficult to determine microbial biomarkers associated with tumors rather than artifacts of technical variation, but correcting for this could enable downstream host-microbial cancer analyses. We thus tested how well BIRDMAn models could extract biological signals from this dataset while accounting for technical batch effects modeled as random effects. We additionally modeled each microbial feature's abundance using this approach to determine the specificity of these microbes for each cancer type (see Methods and Code).

Since cancer types are known to have distinct microbiomes[4,40], we first confirmed that BIRDMAn models could extract cancer type-specific differences despite the technical variation observed in this study. From our log-ratio classification benchmarks, we observe that our custom BIRDMAn model can detect a substantially stronger differential signature between the cancer types compared to ALDEx2, ANCOM-BC, DESeq2, Songbird, and Random Forests (Fig 4.4b; note the axis log-scaling) after controlling for the batch effects due to the sequencing center (Supp Fig AC.1.S2c).

To determine the generalizability of our results, we then constructed a leave-one-center-out cross-validation benchmark using logistic regression on the BIRDMAn-computed log-ratios. Four cancer types with at least three represented data submitting centers (head and neck cancer [HNSC], bladder cancer [BLCA], thyroid cancer [THCA], and cervical cancer [CESC]) were included in this benchmark. The receiver operating characteristic (ROC) curves demonstrated

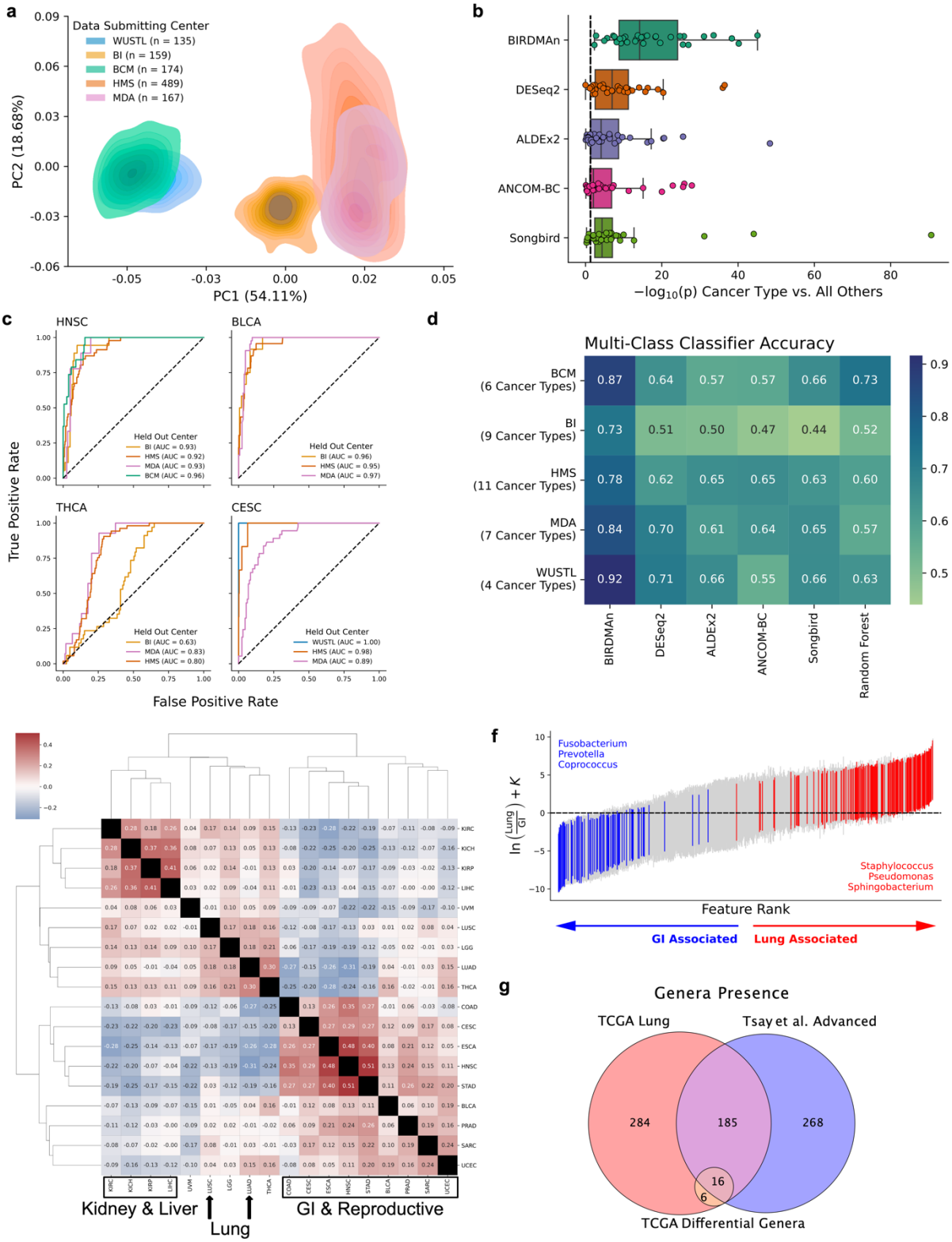
strong classification performance (Fig 4.4c), indicating that BIRDMAn captures generalizable microbial signals across multiple sequencing centers. Generalizability can be a major challenge in microbiome studies[3], where classifiers become overfitted for individual cohorts. We observe this with other DA tools (ALDEx2, DESeq2, ANCOM-BC, Songbird) and even Random Forests (Supp Fig AC.1.S2d), where most tools struggle to achieve an area under the ROC curves (AUROC) of >0.8. BIRDMAn is competitive with these tools, achieving an AUROC >0.9 in HNSC, BLCA, and CESC cancers while achieving the highest predictive accuracy in BLCA and CESC cancers. The high classifier accuracy leaving out each individual center demonstrates that no one center's data strongly affects the classifier accuracy, with the exception of BI for THCA.

To investigate the heterogeneity across different cancer types, we next computed Kendall correlations of BIRDMAn-estimated microbial log-fold changes across all pairs of cancer types. This analysis revealed similarities among cancer types that we would expect, including strong similarities between kidney cancer subtypes (KIRC, KICH, KIRP), lung cancer subtypes (LUAD, LUSC), and gastrointestinal (GI) cancers (COAD, ESCA, HNSC, STAD), Additionally, the BIRDMAn-informed data suggested some novel associations, such as the similarity between kidney cancers and liver cancer (LIHC). When clustering the individual microbes' differentials (Supp Fig AC.1.S2b), we also observed that numerous GI-specific microbes differentiated GI cancers from other cancer types.

When focusing on comparing GI cancers to lung cancers, we found that the resulting BIRDMAn log-fold changes accurately reflected known biology surrounding the niches in which these microbiomes are commonly found. Specifically, *Fusobacterium*[41], *Prevotella*[42], and *Coproccus*[43] are genera commonly found in the GI tract; conversely, *Pseudomonas*[44], *Staphyloccus*[45], and *Sphingobacterium*[46] genera include opportunistic pathogens that are commonly found in lung infections (Fig 4.4f). We cross-referenced our results against the Tsay *et al*. cohort that utilized 16S rRNA sequencing to investigate lung cancer. Out of the 469 genera in

the TCGA lung issues, we observed that 39% of these microbes were also observed in the Tsay *et al*. cohort, despite known previous discordant findings comparing 16S rRNA sequencing and whole genome sequencing[47,48]. Furthermore, when we focus on the top 100 microbes that are detected to be associated with lung cancer, 70% of the represented genera were observed in both the TCGA and Tsay *et al*. datasets. Altogether, this shows how BIRDMAn models can provide biologically-informative results while properly accounting for and mitigating strong structural batch effects that currently confound other DA approaches.

Figure 4.4: Differential abundance analysis on whole-genome sequenced cancer microbiome dataset. (a) Whole-genome sequenced cancer microbiome data from TCGA shows strong batch effects by sequencing center (colored by center; see Supp Fig AC.1.S2a for per cancer type plots). Samples are summarized by the 2D kernel density estimate for each center. (b) T-test p-values comparing log-ratios of each cancer type vs. all others within each center. Dashed line represents p=0.05. All differential abundance methods show significant differences with log-ratios to separate the microbes in each individual cancer type from those found in all other cancer types, but BIRDMAn outperforms other methods in highlighting this difference. (c) ROC curves for leave-one-center-out cross-validation for four cancer types where at least 3 centers sequenced that cancer type (BRCA was not included as it was used as reference). Classifiers were built to predict one-vs-rest for that cancer type. BI = Broad Institute of MIT and Harvard; BCM = Baylor College of Medicine; HMS = Harvard Medical School; MDA = MD Anderson Institute for Applied Cancer Science; WUSTL = Washington University School of Medicine. (d) Multinomial (mean) classification accuracy of classifiers to predict cancer type given the log-ratios computed from the top and bottom 200 taxa associated with each cancer type. Random Forests classifier, which is frequently used in this field but is not based on differential abundance, was included as a comparison for this class of methods. Classifications were performed within each center to remove batch effects from predictions. BIRDMAn outperforms all other methods, including Random Forests, for all tumor types. (e) Clustermap of Kendall tau correlation coefficients of pairwise cancer type differentials (breast cancer as reference). (f) Comparison of lung-associated genera with GI-associated genera. Highlighted genera are known to be associated with either lung or GI microbiome and show strong directionality in the BIRDMAn results. (g) Venn diagram of genera present in TCGA lung samples and genera present in advanced stage lung cancer from work published by Tsay et al. Additionally, the 22 genera represented in the top 100 features associated with TCGA lung cancer cancers are included. A majority of these genera (16/22) are present in both datasets.


a
Data Submitting Center
WUSTL (n = 135)
BI (n = 159)
BCM (n = 174)
HMS (n = 489)
MDA (n = 167)
PC2 (18.68%)
PC1 (54.11%)
b
BIRDMAn
DESeq2
ALDEx2
ANCOM-BC
Songbird
−log10(p) Cancer Type vs. All Others
c
HNSC
BLCA
THCA
CESC
Held Out Center
True Positive Rate
False Positive Rate
d
Multi-Class Classifier Accuracy
BCM (6 Cancer Types)
BI (9 Cancer Types)
HMS (11 Cancer Types)
MDA (7 Cancer Types)
WUSTL (4 Cancer Types)
BIRDMAn
DESeq2
ALDEx2
ANCOM-BC
Songbird
Random Forest
e
Kidney & Liver
Lung
GI & Reproductive
f
Fusobacterium
Prevotella
Coprococcus
Staphylococcus
Pseudomonas
Sphingobacterium
Feature Rank
GI Associated
Lung Associated
g
Genera Presence
TCGA Lung
Tsay et al. Advanced
284
185
268
16
6
TCGA Differential Genera

4.3. Methods

Performing Bayesian inference with Stan

Parameter estimation was performed using Bayesian inference. Our approach utilizes Bayes' Rule where θ represents the parameter space and D represents our collected data:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}.$$

Because the evidence term, $P(D)$, is simply a normalizing constant, we can rewrite Bayes' Rule as follows, substituting terms with their common nomenclature:

$$\text{Posterior} \sim \text{Likelihood} \cdot \text{Prior}$$

Thus, our objective with Bayesian inference is to obtain the posterior distribution by modeling the likelihood function of our data as well as our prior knowledge of the parameters. Absent a model formulation involving conjugate priors, we cannot compute the posterior distribution analytically. Instead, we use Stan to draw samples from the posterior distribution using the No-U-Turn Hamiltonian Monte Carlo sampler[20]. A series of Markov chains are initialized and allowed to "warm-up" in their exploration of the parameter posterior distributions. Once the defined number of warm-up iterations has concluded, a set number of samples are drawn from each of the chains. Multiple chains are run to ensure that model convergence occurs.

We implement Bayesian inference using the CmdStanPy interface in Python, calling the C++ Stan toolchain for efficient sampling. The warm-up iterations are discarded by default and the sampling iterations are saved for each Markov chain.

Negative binomial model parameterization

We fit counts of each microbe in a dataset according to a negative binomial distribution as an approximation of multinomial logistic regression[55]. Due to overdispersion, standard count models such as Poisson are inappropriate for sequencing data[21]. We note that the negative binomial model can be considered an extension to the Poisson model with additional variance components[56].

The negative binomial models used in this work are described by parameters for both mean and overdispersion. This is in contrast to traditional parameters in negative binomial models described by the probability of success and the number of failures before an instance of a success. The former model, often referred to as the "alternative parameterization," is more amenable to generalized linear modeling through hierarchical models as the mean can be modeled directly.

The basic format of the alternative parameterization negative binomial model is described below where n corresponds to the count, ϕ the overdispersion, and μ the mean count.

$$\mathrm{NB}\left(n \mid \mu, \phi\right) = \binom{n + \phi - 1}{n} \left(\frac{\mu}{\mu + \phi}\right)^{n} \left(\frac{\phi}{\mu + \phi}\right)^{\phi}$$

We use a log-link function, $\mu = \exp\left(\eta\right)$ to model the mean where the log mean count, η, can be represented by linear terms. To account for variable sequencing depth among samples, we include log sequencing depth as an offset term in our models.

BIRDMAn framework

We developed BIRDMAn as a framework for highly-customizable Bayesian differential abundance modeling. BIRDMAn abstracts much of the Bayesian workflow away for usage with microbiome data. An object-oriented approach allows users to subclass basic models for their

custom implementations. BIRDMAn includes, by default, a Negative Binomial model implementation. This can be used without writing any new Stan code or subclassing any BIRDMAn objects.

BIRDMAn models take BIOM tables[57] as input containing the sample and observation IDs. Sample metadata can be provided as Pandas DataFrames. We provide a method, create_regression, with which users can provide an R-style formula to automatically create the design matrix using the patsy Python package. Another method, specify_model, allows the specification of the desired parameters and dimensions to return. This method is used by create_inference to convert CmdStanPy output to ArviZ[58] InferenceData objects.

There are two base classes included with BIRDMAn termed the TableModel and the SingleFeatureModel. The TableModel allows fitting an entire dataset at once, while the SingleFeatureModel allows for fitting individual features. The SingleFeatureModel is advantageous as it allows for highly parallelized workflows. Because there are often hundreds or thousands of features in a microbiome dataset, we note that using multiple CPUs to run many features at once is often more efficient than fitting the entire table. We provide a convenience class, ModelIterator, to iterate through the features in a given table. This class also allows for dividing the table into chunks. This allows users to customize the number of features to fit at once depending on their computational resources.

Simulations

All simulations were performed through the fixed_param option in CmdStanPy. Ground-truth parameters were provided into a negative binomial generative model to simulate data from mean and dispersion parameters.

For the data-driven simulation, we randomly drew values for batch offset, batch dispersion, and base dispersion parameters. These parameters were fed into a model with $\beta_0 = N(-8, 1)$,

$\beta_1 = N(2, 1)$. Log sampling depth was simulated from a Poisson-Lognormal distribution with λ drawn from $N(5000, 0.2)$. We simulated 300 samples comprising 10 total batches with 10 total features.

For the variable sample size simulations, we simulated feature counts for 500 samples with $\beta_0 = 8$, $\beta_1 = 3$, and $\frac{1}{\phi} = 10$. Log sequencing depths were simulated using a Poisson-Lognormal model with λ drawn from $N(50000, 0.5)$ where depth varied.

To simulate variable rarefaction depth, we first drew ground truth intercept and beta values from $N(-8, 1)$ and $N(2, 1)$ respectively for 1000 features. These values were used to generate counts for 300 samples through the multinomial distribution. We used the multinomial distribution to enforce the same sampling depth for all samples, simulating rarefaction.

Antibiotics case study

16S data was downloaded from Qiita study 494; we used 16S OTUs picked against the GreenGenes_13.8[59] reference database at 97% sequence similarity. OTU picking was performed with SortMeRNA[60] with Qiita default parameter values. Features present in fewer than 10 samples were filtered. We also removed samples with a total sequencing depth less than 1000.

To account for the longitudinal nature of this design, we used backwards difference encoding such that each time point was compared to the one immediately before it. We implemented the subject identifiers as a random effect with both random intercepts and random slopes. The posterior draws were centered around the mean. Ranking of OTUs by differentials for log-ratio feature selection was done using the posterior means.

We performed t-tests comparing the log-ratios between groups of samples at different timepoints. The alternative hypothesis was chosen such that samples from the later time point would have higher log-ratios than those from the initial timepoint due to the anticipated effect of Cp on microbial populations.

We then implemented multinomial logistic regression, random forest classification, and repeated k-fold cross-validation through scikit-learn for our machine learning approach. Because DESeq2 & Songbird supports contrasts, we computed the same contrasts as BIRDMAn for parity. With ALDEx2 and ANCOM-BC, we computed the differentials associated with each timepoint using preCp as reference. For the random forest classifier, we used the CLR-transformed feature table (with a pseudocount of 1) entries as the predictors. All models were also provided one-hot-encoded vectors for subject identifiers. Performance was measured using balanced accuracy. For multinomial logistic regression we used the lbfgs solver with 1000 max iterations. For the random forest classifier we used a set random seed and 100 estimators. We used repeated stratified k-fold cross validation with 5 splits and 10 repeats and a random seed. All other parameters not mentioned were set to the scikit-learn defaults.

Posterior draws for timepoint-contrast differentials were analyzed with (1) FirstCp-associated features with preCp-associated features as reference and (2) SecondCp-associated features with Interim-associated features as reference. In this way, the posterior distribution reflects how each Cp course affects the selected bacterial features over time.

For determining the Gram status of each OTU, we used the BugBase[61] web interface. We took the set intersection of Gram positive and Gram negative features with the features associated with both FirstCp and SecondCp to determine the Gram status breakdown of both numerator and denominator features.

TCGA case study

The bacterial TCGA tables were obtained from those processed in Narunsky-Haziza et al.[62] and Poore et al.[4] All TCGA sequence data were accessed via the Cancer Genomics Cloud[63] (CGC) as sponsored by SevenBridges (https://cgc.sbgenomics.com) after obtaining data access from the TCGA Data Access Committee through dbGaP (https://

dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login). On Qiita[64], TCGA WGS host-depleted and quality-controlled fastq files were used to generate a metagenomic table by direct genome alignments based on Woltka v0.1.1[65] against the RefSeq[66] release 200 (built as of May 14, 2020). The resulting tables can be found on Qiita under study ID 13722, of which we filtered to only analyze the bacteria and then were subsequently decontaminated through decontam[67] (https://github.com/benjjneb/decontam) (version 1.14.0) following the protocol described in Poore et al.[4]

After initial table generation, we removed samples from data submitting centers with very few samples. We also filtered our data to only include samples from white, African-American, and Asian races. Additionally, we only included samples from patients who were alive at the time of sample procurement and retained only one sample per subject. To filter out lowly prevalent features, we removed features present in fewer than 50 total samples. To remove samples with low sequencing depth, we set a threshold of 500 reads. Finally, we included only cancer types with at least 20 instances in the dataset for statistical power.

We then built statistical models to model the differential associated with each cancer type. Because TCGA did not include “normal” samples from healthy individuals, we used breast cancer (BRCA) tumor samples as reference. Both race[68] and gender were also included as covariates. Data submitting center was incorporated as a random effect (both random intercepts and random slopes).

Posterior means were computed for each feature’s association with each individual cancer type. For each cancer type, we ranked the differentials and used the top and bottom 200 features associated with that cancer type to compute log-ratios per sample. These log-ratios were used as predictor variables in our machine learning models.

Because not every cancer type was represented in each center, we performed multi-class classification within centers. For each center, we fit a model to predict cancer type from our log-

ratios. This procedure was performed with 5 repeats of stratified 2-fold cross-validation. We repeated this machine learning process for cancer type differentials from DESeq2, ALDEx2, Songbird, and ANCOM-BC. For comparison, we fit a random forest classifier on the CLR-transformed feature table to predict cancer type as well.

The leave-one-center-out models were fit using binomial logistic regression with balanced class weights. For each cancer type, we fit a model on all but one center and used that model to predict cancer type for the held-out center. We also used the same random forest classifier as previously described for comparison.

Analysis & visualization software

Analysis of the results in this work were primarily performed through Python (v3.8.13). Pandas[69] (v1.1.5) and NumPy[70] (v1.22.3) were used for general data analysis. SciPy[71] (v1.7.3) was used for computing statistical tests. For interfacing with multidimensional arrays we used xarray[72] (v0.20.1) and ArviZ[58] (0.12.1). Machine learning models were fit and cross-validated using scikit-learn[73] (v1.0.2). Python figures were generated using seaborn[74] (v0.11.2) and Matplotlib[75] (v3.5.1) as well as Matplotlib-venn (v0.11.7). We used biom-format[57] (2.1.12) and scikit-bio (v0.5.6) for statistical analysis of microbiome data structures.

R analysis was performed using the tidyverse[76] packages dplyr (v1.0.9), stringr (v1.4.0), and ggplot2 (v3.3.6). Phylogenetic visualization was performed using treeio[77] (v1.18.0) and ggtree[78] (v3.2.0). BIOM tables were read using the biomformat R package (v1.22.0).

4.4 Code and data availability

All data used were downloaded from publicly available Qiita studies. The scripts and Stan models used to analyze these data as well as Jupyter notebooks for the visualizations are

available at https://github.com/knightlab-analyses/birdman-analyses-final. The BIRDMAn software package is available at https://github.com/biocore/BIRDMAn and the documentation is available at https://birdman.readthedocs.io/. All analyses in this work were performed using BIRDMAn v0.1.0.

4.5. Acknowledgements

We thank the members of the Knight Lab and Morton Lab for feedback and bug reporting for the BIRDMAn software. We thank the developers of ArviZ and CmdStanPy for responding to and addressing issues on GitHub as well as the users on the Stan forums for answering our questions.

This work was supported in part by NIH U19AG063744, NIH 1DP1AT010885, and NIH U24CA248454. J.T.M. was funded by the intramural research program of the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD).

Chapter 4 has been submitted for publication of the material as it may appear in *Nature Microbiology*, "BIRDMAn: A Bayesian differential abundance framework that enables robust inference of host-microbe associations" Gibraan Rahman, James T. Morton, Cameron Martino, Gregory D. Sepich-Poore, Celeste Allaband, Caitlin Guccione, Yang Chen, Daniel Hakim, Mehrbod Estaki, and Rob Knight. The dissertation author was the primary investigator and first author of this paper.

Author information

G.R., J.T.M., and R.K. conceived the idea for the study. G.R. & J.T.M. developed the BIRDMAn software package. G.R., J.T.M., C.G., G.D.S-P., & C.M. contributed to the case study and simulation analysis. C.A., J.T.M., C.M., & R.K. helped to define the scope of the analyses. G.R. & Y.C. contributed to the documentation for BIRDMAn. M.E., Y.C., D.H., & C.M. gave critical

feedback on the usage and documentation of the software. All authors helped write and review the manuscript.

Conflicts of interest

G.D.S.-P. and R.K. are inventors on a US patent application (PCT/US2019/059647) submitted by The Regents of the University of California and licensed by Micronoma; that application covers methods of diagnosing and treating cancer using multi-domain microbial biomarkers in blood and cancer tissues. G.D.S.-P. and R.K. are founders of and report stock interest in Micronoma. G.D.S.-P. has filed several additional US patent applications on cancer bacteriome and mycobiome diagnostics that are owned by The Regents of the University of California or Micronoma. R.K. additionally is a member of the scientific advisory board for GenCirq, holds an equity interest in GenCirq, and can receive reimbursements for expenses up to US $5,000 per year.

4.4. References

1. Sochocka, M., Donskow-Łysoniewska, K., Diniz, B. S., Kurpas, D., Brzozowska, E. & Leszek, J. The Gut Microbiome Alterations and Inflammation-Driven Pathogenesis of Alzheimer's Disease—a Critical Review. *Mol. Neurobiol.* **56,** 1841–1851 (2019).

2. Fouquier, J., Moreno Huizar, N., Donnelly, J., Glickman, C., Kang, D.-W., Maldonado, J., Jones, R. A., Johnson, K., Adams, J. B., Krajmalnik-Brown, R. & Lozupone, C. The Gut Microbiome in Autism: Study-Site Effects and Longitudinal Analysis of Behavior Change. *mSystems* **6,** e00848-20 (2021).

3. Wirbel, J., Pyl, P. T., Kartal, E., Zych, K., Kashani, A., Milanese, A., Fleck, J. S., Voigt, A. Y., Palleja, A., Ponnudurai, R., Sunagawa, S., Coelho, L. P., Schrotz-King, P., Vogtmann, E., Habermann, N., Niméus, E., Thomas, A. M., Manghi, P., Gandini, S., Serrano, D., Mizutani, S., Shiroma, H., Shiba, S., Shibata, T., Yachida, S., Yamada, T., Waldron, L., Naccarati, A., Segata, N., Sinha, R., Ulrich, C. M., Brenner, H., Arumugam, M., Bork, P. & Zeller, G. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* **25,** 679–689 (2019).

4. Poore, G. D., Kopylova, E., Zhu, Q., Carpenter, C., Fraraccio, S., Wandro, S., Kosciolek, T., Janssen, S., Metcalf, J., Song, S. J., Kanbar, J., Miller-Montgomery, S., Heaton, R., Mckay, R., Patel, S. P., Swafford, A. D. & Knight, R. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* **579,** 567–574 (2020).

5. Villapol, S. Gastrointestinal symptoms associated with COVID-19: impact on the gut microbiome. *Transl. Res.* **226,** 57–69 (2020).

6. Zuo, T., Zhan, H., Zhang, F., Liu, Q., Tso, E. Y. K., Lui, G. C. Y., Chen, N., Li, A., Lu, W., Chan, F. K. L., Chan, P. K. S. & Ng, S. C. Alterations in Fecal Fungal Microbiome of Patients With COVID-19 During Time of Hospitalization until Discharge. *Gastroenterology* **159,** 1302-1310.e5 (2020).

7. Poyet, M., Groussin, M., Gibbons, S. M., Avila-Pacheco, J., Jiang, X., Kearney, S. M., Perrotta, A. R., Berdy, B., Zhao, S., Lieberman, T. D., Swanson, P. K., Smith, M., Roesemann, S., Alexander, J. E., Rich, S. A., Livny, J., Vlamakis, H., Clish, C., Bullock, K., Deik, A., Scott, J., Pierce, K. A., Xavier, R. J. & Alm, E. J. A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nat. Med.* **25,** 1442–1452 (2019).

8. Kostic, A. D., Gevers, D., Siljander, H., Vatanen, T., Hyötyläinen, T., Hämäläinen, A.-M., Peet, A., Tillmann, V., Pöhö, P., Mattila, I., Lähdesmäki, H., Franzosa, E. A., Vaarala, O., de Goffau, M., Harmsen, H., Ilonen, J., Virtanen, S. M., Clish, C. B., Orešič, M., Huttenhower, C., Knip, M., DIABIMMUNE Study Group & Xavier, R. J. The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell Host Microbe* **17,** 260–273 (2015).

9. Proctor, L. M., Creasy, H. H., Fettweis, J. M., Lloyd-Price, J., Mahurkar, A., Zhou, W., Buck, G. A., Snyder, M. P., Strauss, J. F., Weinstock, G. M., White, O., Huttenhower, C., & The Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project. *Nature* **569,** 641–648 (2019).

10. Gopalakrishnan, V., Spencer, C. N., Nezi, L., Reuben, A., Andrews, M. C., Karpinets, T. V., Prieto, P. A., Vicente, D., Hoffman, K., Wei, S. C., Cogdill, A. P., Zhao, L., Hudgens, C. W., Hutchinson, D. S., Manzo, T., Macedo, M. P. de, Cotechini, T., Kumar, T., Chen, W. S., Reddy, S. M., Sloane, R. S., Galloway-Pena, J., Jiang, H., Chen, P. L., Shpall, E. J., Rezvani, K., Alousi, A. M., Chemaly, R. F., Shelburne, S., Vence, L. M., Okhuysen, P. C., Jensen, V. B., Swennes, A. G., McAllister, F., Sanchez, E. M. R., Zhang, Y., Chatelier, E. L., Zitvogel, L., Pons, N., Austin-Breneman, J. L., Haydu, L. E., Burton, E. M., Gardner, J. M., Sirmans, E., Hu, J., Lazar, A. J., Tsujikawa, T., Diab, A., Tawbi, H., Glitza, I. C., Hwu, W. J., Patel, S. P., Woodman, S. E., Amaria, R. N., Davies, M. A., Gershenwald, J. E., Hwu, P., Lee, J. E., Zhang, J., Coussens, L. M., Cooper, Z. A., Futreal, P. A., Daniel, C. R., Ajami, N. J., Petrosino, J. F., Tetzlaff, M. T., Sharma, P., Allison, J. P., Jenq, R. R. & Wargo, J. A. Gut microbiome modulates response to anti–PD-1 immunotherapy in melanoma patients. *Science* **359,** 97–103 (2018).

11. Spencer, C. N., McQuade, J. L., Gopalakrishnan, V., McCulloch, J. A., Vetizou, M., Cogdill, A. P., Khan, M. A. W., Zhang, X., White, M. G., Peterson, C. B., Wong, M. C., Morad, G., Rodgers, T., Badger, J. H., Helmink, B. A., Andrews, M. C., Rodrigues, R. R., Morgun, A., Kim, Y. S., Roszik, J., Hoffman, K. L., Zheng, J., Zhou, Y., Medik, Y. B., Kahn, L. M., Johnson, S., Hudgens, C. W., Wani, K., Gaudreau, P.-O., Harris, A. L., Jamal, M. A., Baruch, E. N., Perez-Guijarro, E., Day, C.-P., Merlino, G., Pazdrak, B., Lochmann, B. S., Szczepaniak-Sloane, R. A., Arora, R., Anderson, J., Zobniw, C. M., Posada, E., Sirmans, E., Simon, J., Haydu, L. E., Burton, E. M., Wang, L., Dang, M., Clise-Dwyer, K., Schneider, S., Chapman, T., Anang, N.-A. A. S., Duncan, S., Toker, J., Malke, J. C., Glitza, I. C., Amaria, R. N., Tawbi, H. A., Diab, A., Wong, M. K., Patel, S. P., Woodman, S. E., Davies, M. A., Ross, M. I., Gershenwald, J. E., Lee, J. E., Hwu, P., Jensen, V., Samuels, Y., Straussman, R., Ajami, N. J., Nelson, K. C., Nezi, L., Petrosino, J. F., Futreal, P. A., Lazar, A. J., Hu, J., Jenq, R. R., Tetzlaff, M. T., Yan, Y., Garrett, W. S., Huttenhower, C., Sharma, P., Watowich, S. S., Allison, J. P., Cohen, L., Trinchieri, G., Daniel, C. R. & Wargo, J. A. Dietary fiber and probiotics influence the gut microbiome and melanoma immunotherapy response. *Science* **374,** 1632–1640 (2021).

12. Lee, K. A., Thomas, A. M., Bolte, L. A., Björk, J. R., de Ruijter, L. K., Armanini, F., Asnicar, F., Blanco-Miguez, A., Board, R., Calbet-Llopart, N., Derosa, L., Dhomen, N., Brooks, K., Harland, M., Harries, M., Leeming, E. R., Lorigan, P., Manghi, P., Marais, R., Newton-Bishop, J., Nezi, L., Pinto, F., Potrony, M., Puig, S., Serra-Bellver, P., Shaw, H. M., Tamburini, S., Valpione, S., Vijay, A., Waldron, L., Zitvogel, L., Zolfo, M., de Vries, E. G. E., Nathan, P., Fehrmann, R. S. N., Bataille, V., Hospers, G. A. P., Spector, T. D., Weersma, R. K. & Segata, N. Cross-cohort gut microbiome associations with immune checkpoint inhibitor response in advanced melanoma. *Nat. Med.* **28,** 535–544 (2022).

13. Hiergeist, A., Reischl, U. & Gessner, A. Multicenter quality assessment of 16S ribosomal DNA-sequencing for microbiome analyses reveals high inter-center variability. *Int. J. Med. Microbiol.* **306,** 334–342 (2016).

14. Wang, Y. & LêCao, K.-A. Managing batch effects in microbiome data. *Brief. Bioinform.* **21,** 1954–1970 (2020).

15. Chen, W., Zhang, S., Williams, J., Ju, B., Shaner, B., Easton, J., Wu, G. & Chen, X. A comparison of methods accounting for batch effects in differential expression analysis of UMI count based single cell RNA sequencing. *Comput. Struct. Biotechnol. J.* **18,** 861–873 (2020).

16. Vandeputte, D., Kathagen, G., D'hoe, K., Vieira-Silva, S., Valles-Colomer, M., Sabino, J., Wang, J., Tito, R. Y., De Commer, L., Darzi, Y., Vermeire, S., Falony, G. & Raes, J. Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551,** 507–511 (2017).

17. Kumar, M. S., Slud, E. V., Okrah, K., Hicks, S. C., Hannenhalli, S. & Corrada Bravo, H. Analysis and correction of compositional bias in sparse sequencing count data. *BMC Genomics* **19,** 799 (2018).

18. Nixon, M. P., Letourneau, J., David, L. A., Mukherjee, S. & Silverman, J. D. A Statistical Analysis of Compositional Surveys. Preprint at https://doi.org/10.48550/arXiv.2201.03616 (2022)

19. Nearing, J. T., Douglas, G. M., Hayes, M., MacDonald, J., Desai, D., Allward, N., Jones, C. M. A., Wright, R., Dhanani, A., Comeau, A. M. & Langille, M. G. I. *Microbiome differential abundance methods produce disturbingly different results across 38 datasets*. 2021.05.10.443486 (2021). doi:10.1101/2021.05.10.443486

20. Hoffman, M. D. & Gelman, A. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. Preprint at https://doi.org/10.48550/arXiv.1111.4246 (2011)

21. McMurdie, P. J. & Holmes, S. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLOS Comput. Biol.* **10,** e1003531 (2014).

22. Äijö, T., Müller, C. L. & Bonneau, R. Temporal probabilistic modeling of bacterial compositions derived from 16S rRNA sequencing. *Bioinforma. Oxf. Engl.* **34,** 372–380 (2018).

23. Silverman, J. D., Durand, H. K., Bloom, R. J., Mukherjee, S. & David, L. A. Dynamic linear models guide design and analysis of microbiota studies within artificial human guts. *Microbiome* **6,** 202 (2018).

24. Joseph, T. A., Shenhav, L., Xavier, J. B., Halperin, E. & Pe'er, I. Compositional Lotka-Volterra describes microbial dynamics in the simplex. *PLOS Comput. Biol.* **16,** e1007917 (2020).

25. Joseph, T. A., Pasarkar, A. P. & Pe'er, I. Efficient and Accurate Inference of Mixed Microbial Population Trajectories from Longitudinal Count Data. *Cell Syst.* **10,** 463-469.e6 (2020).

26. Shenhav, L., Furman, O., Briscoe, L., Thompson, M., Silverman, J. D., Mizrahi, I. & Halperin, E. Modeling the temporal dynamics of the gut microbial community in adults and infants. *PLOS Comput. Biol.* **15,** e1006960 (2019).

27. Dethlefsen, L. & Relman, D. A. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proc. Natl. Acad. Sci.* **108,** 4554–4561 (2011).

28. Martino, C., Shenhav, L., Marotz, C. A., Armstrong, G., McDonald, D., Vázquez-Baeza, Y., Morton, J. T., Jiang, L., Dominguez-Bello, M. G., Swafford, A. D., Halperin, E. & Knight,

R. Context-aware dimensionality reduction deconvolutes gut microbial community dynamics. *Nat. Biotechnol.* **39,** 165–168 (2021).

29. Gibbons, S. M. Keystone taxa indispensable for microbiome recovery. *Nat. Microbiol.* **5,** 1067–1068 (2020).

30. Chng, K. R., Ghosh, T. S., Tan, Y. H., Nandi, T., Lee, I. R., Ng, A. H. Q., Li, C., Ravikrishnan, A., Lim, K. M., Lye, D., Barkham, T., Raman, K., Chen, S. L., Chai, L., Young, B., Gan, Y.-H. & Nagarajan, N. Metagenome-wide association analysis identifies microbial determinants of post-antibiotic ecological recovery in the gut. *Nat. Ecol. Evol.* **4,** 1256–1267 (2020).

31. Morton, J. T., Marotz, C., Washburne, A., Silverman, J., Zaramela, L. S., Edlund, A., Zengler, K. & Knight, R. Establishing microbial composition measurement standards with reference frames. *Nat. Commun.* **10,** 2719 (2019).

32. Fernandes, A. D., Reid, J. N., Macklaim, J. M., McMurrough, T. A., Edgell, D. R. & Gloor, G. B. 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 (2014).

33. Lin, H. & Peddada, S. D. Analysis of compositions of microbiomes with bias correction. *Nat. Commun.* **11,** 3514 (2020).

34. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).

35. Oliphant, C. M. & Green, G. M. Quinolones: A Comprehensive Review. *Am. Fam. Physician* **65,** 455–465 (2002).

36. Card, R. M., Mafura, M., Hunt, T., Kirchner, M., Weile, J., Rashid, M.-U., Weintraub, A., Nord, C. E. & Anjum, M. F. Impact of Ciprofloxacin and Clindamycin Administration on Gram-Negative Bacteria Isolated from Healthy Volunteers and Characterization of the Resistance Genes They Harbor. *Antimicrob. Agents Chemother.* **59,** 4410–4416 (2015).

37. Peterson, C. T., Sharma, V., Elmén, L. & Peterson, S. N. Immune homeostasis, dysbiosis and therapeutic modulation of the gut microbiota. *Clin. Exp. Immunol.* **179,** 363–377 (2015).

38. Ramirez, J., Guarner, F., Bustos Fernandez, L., Maruy, A., Sdepanian, V. L. & Cohen, H. Antibiotics as Major Disruptors of Gut Microbiota. *Front. Cell. Infect. Microbiol.* **10,** (2020).

39. Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C. & Stuart, J. M. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45,** 1113–1120 (2013).

40. Nejman, D., Livyatan, I., Fuks, G., Gavert, N., Zwang, Y., Geller, L. T., Rotter-Maskowitz, A., Weiser, R., Mallel, G., Gigi, E., Meltser, A., Douglas, G. M., Kamer, I., Gopalakrishnan, V., Dadosh, T., Levin-Zaidman, S., Avnet, S., Atlan, T., Cooper, Z. A., Arora, R., Cogdill, A. P., Khan, M. A. W., Ologun, G., Bussi, Y., Weinberger, A., Lotan-Pompan, M., Golani, O., Perry, G., Rokah, M., Bahar-Shany, K., Rozeman, E. A., Blank, C. U., Ronai, A.,

Shaoul, R., Amit, A., Dorfman, T., Kremer, R., Cohen, Z. R., Harnof, S., Siegal, T., Yehuda-Shnaidman, E., Gal-Yam, E. N., Shapira, H., Baldini, N., Langille, M. G. I., Ben-Nun, A., Kaufman, B., Nissan, A., Golan, T., Dadiani, M., Levanon, K., Bar, J., Yust-Katz, S., Barshack, I., Peeper, D. S., Raz, D. J., Segal, E., Wargo, J. A., Sandbank, J., Shental, N. & Straussman, R. The human tumor microbiome is composed of tumor type-specific intracellular bacteria. *Science* **368,** 973–980 (2020).

41. Bullman, S., Pedamallu, C. S., Sicinska, E., Clancy, T. E., Zhang, X., Cai, D., Neuberg, D., Huang, K., Guevara, F., Nelson, T., Chipashvili, O., Hagan, T., Walker, M., Ramachandran, A., Diosdado, B., Serna, G., Mulet, N., Landolfi, S., Ramon y Cajal, S., Fasani, R., Aguirre, A. J., Ng, K., Élez, E., Ogino, S., Tabernero, J., Fuchs, C. S., Hahn, W. C., Nuciforo, P. & Meyerson, M. Analysis of Fusobacterium persistence and antibiotic response in colorectal cancer. *Science* **358,** 1443–1448 (2017).

42. Lo, C.-H., Wu, D.-C., Jao, S.-W., Wu, C.-C., Lin, C.-Y., Chuang, C.-H., Lin, Y.-B., Chen, C.-H., Chen, Y.-T., Chen, J.-H., Hsiao, K.-H., Chen, Y.-J., Chen, Y.-T., Wang, J.-Y. & Li, L.-H. Enrichment of Prevotella intermedia in human colorectal cancer and its additive effects with Fusobacterium nucleatum on the malignant transformation of colorectal adenomas. *J. Biomed. Sci.* **29,** 88 (2022).

43. Flemer, B., Lynch, D. B., Brown, J. M. R., Jeffery, I. B., Ryan, F. J., Claesson, M. J., O'Riordain, M., Shanahan, F. & O'Toole, P. W. Tumour-associated and non-tumour-associated microbiota in colorectal cancer. *Gut* **66,** 633–643 (2017).

44. Nunley, D. R., Grgurich, W., Iacono, A. T., Yousem, S., Ohori, N. P., Keenan, R. J. & Dauber, J. H. Allograft Colonization and Infections With Pseudomonas in Cystic Fibrosis Lung Transplant Recipients. *Chest* **113,** 1235–1243 (1998).

45. Laroumagne, S., Lepage, B., Hermant, C., Plat, G., Phelippeau, M., Bigay-Game, L., Lozano, S., Guibert, N., Segonds, C., Mallard, V., Augustin, N., Didier, A. & Mazieres, J. Bronchial colonisation in patients with lung cancer: a prospective study. *Eur. Respir. J.* **42,** 220–229 (2013).

46. Lambiase, A., Rossano, F., Del Pezzo, M., Raia, V., Sepe, A., de Gregorio, F. & Catania, M. R. Sphingobacterium respiratory tract infection in patients with cystic fibrosis. *BMC Res. Notes* **2,** 262 (2009).

47. Brumfield, K. D., Huq, A., Colwell, R. R., Olds, J. L. & Leddy, M. B. Microbial resolution of whole genome shotgun and 16S amplicon metagenomic sequencing using publicly available NEON data. *PLOS ONE* **15,** e0228899 (2020).

48. Durazzi, F., Sala, C., Castellani, G., Manfreda, G., Remondini, D. & De Cesare, A. Comparison between 16S rRNA and shotgun sequencing data for the taxonomic characterization of the gut microbiota. *Sci. Rep.* **11,** 3030 (2021).

49. Hawinkel, S., Mattiello, F., Bijnens, L. & Thas, O. A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Brief. Bioinform.* **20,** 210–221 (2019).

50. Yang, L. & Chen, J. A comprehensive evaluation of microbial differential abundance analysis methods: current status and potential solutions. *Microbiome* **10,** 130 (2022).

51. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome Datasets Are Compositional: And This Is Not Optional. *Front. Microbiol.* **8,** (2017).

52. Williamson, B. D., Hughes, J. P. & Willis, A. D. A multiview model for relative and absolute microbial abundances. *Biometrics* **78,** 1181–1194 (2022).

53. Hawinkel, S., Rayner, J. C. W., Bijnens, L. & Thas, O. Sequence count data are poorly fit by the negative binomial distribution. *PLOS ONE* **15,** e0224909 (2020).

54. Townes, F. W. Review of Probability Distributions for Modeling Count Data. *ArXiv200104343 Stat* (2020). at <http://arxiv.org/abs/2001.04343>

55. Taddy, M. Distributed multinomial regression. *Ann. Appl. Stat.* **9,** (2015).

56. Lindén, A. & Mäntyniemi, S. Using the negative binomial distribution to model overdispersion in ecological count data. *Ecology* **92,** 1414–1421 (2011).

57. McDonald, D., Clemente, J. C., Kuczynski, J., Rideout, J. R., Stombaugh, J., Wendel, D., Wilke, A., Huse, S., Hufnagle, J., Meyer, F., Knight, R. & Caporaso, J. G. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *GigaScience* **1,** 7 (2012).

58. Kumar, R., Carroll, C., Hartikainen, A. & Martin, O. ArviZ a unified library for exploratory analysis of Bayesian models in Python. *J. Open Source Softw.* **4,** 1143 (2019).

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

60. Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28,** 3211–3217 (2012).

61. Ward, T., Larson, J., Meulemans, J., Hillmann, B., Lynch, J., Sidiropoulos, D., Spear, J. R., Caporaso, G., Blekhman, R., Knight, R., Fink, R. & Knights, D. BugBase predicts organism-level microbiome phenotypes. 133462 Preprint at https://doi.org/10.1101/133462 (2017)

62. Narunsky-Haziza, L., Sepich-Poore, G. D., Livyatan, I., Asraf, O., Martino, C., Nejman, D., Gavert, N., Stajich, J. E., Amit, G., González, A., Wandro, S., Perry, G., Ariel, R., Meltser, A., Shaffer, J. P., Zhu, Q., Balint-Lahat, N., Barshack, I., Dadiani, M., Gal-Yam, E. N., Patel, S. P., Bashan, A., Swafford, A. D., Pilpel, Y., Knight, R. & Straussman, R. Pan-cancer analyses reveal cancer-type-specific fungal ecologies and bacteriome interactions. *Cell* **185,** 3789-3806.e17 (2022).

63. Lau, J. W., Lehnert, E., Sethi, A., Malhotra, R., Kaushik, G., Onder, Z., Groves-Kirkby, N., Mihajlovic, A., DiGiovanna, J., Srdic, M., Bajcic, D., Radenkovic, J., Mladenovic, V., Krstanovic, D., Arsenijevic, V., Klisic, D., Mitrovic, M., Bogicevic, I., Kural, D., Davis-Dusenbery, B., & Seven Bridges CGC Team. The Cancer Genomics Cloud: Collaborative, Reproducible, and Democratized-A New Paradigm in Large-Scale Computational Research. *Cancer Res.* **77,** e3–e6 (2017).

64. Gonzalez, A., Navas-Molina, J. A., Kosciolek, T., McDonald, D., Vázquez-Baeza, Y., Ackermann, G., DeReus, J., Janssen, S., Swafford, A. D., Orchanian, S. B., Sanders, J. G., Shorenstein, J., Holste, H., Petrus, S., Robbins-Pianka, A., Brislawn, C. J., Wang, M., Rideout, J. R., Bolyen, E., Dillon, M., Caporaso, J. G., Dorrestein, P. C. & Knight, R. Qiita: rapid, web-enabled microbiome meta-analysis. *Nat. Methods* **15,** 796–798 (2018).

65. Zhu, Q., Huang, S., Gonzalez, A., McGrath, I., McDonald, D., Haiminen, N., Armstrong, G., Vázquez-Baeza, Y., Yu, J., Kuczynski, J., Sepich-Poore, G. D., Swafford, A. D., Das, P., Shaffer, J. P., Lejzerowicz, F., Belda-Ferre, P., Havulinna, A. S., Méric, G., Niiranen, T., Lahti, L., Salomaa, V., Kim, H.-C., Jain, M., Inouye, M., Gilbert, J. A. & Knight, R. OGUs enable effective, phylogeny-aware analysis of even shallow metagenome community structures. 2021.04.04.438427 Preprint at https://doi.org/10.1101/2021.04.04.438427 (2021)

66. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35,** D61–D65 (2007).

67. Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A. & Callahan, B. J. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* **6,** 226 (2018).

68. Luo, M., Liu, Y., Hermida, L. C., Gertz, E. M., Zhang, Z., Li, Q., Diao, L., Ruppin, E. & Han, L. Race is a key determinant of the human intratumor microbiome. *Cancer Cell* **40,** 901–902 (2022).

69. McKinney, W. Data Structures for Statistical Computing in Python. 6 (2010).

70. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C. & Oliphant, T. E. Array programming with NumPy. *Nature* **585,** 357–362 (2020).

71. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F. & van Mulbregt, P. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17,** 261–272 (2020).

72. Hoyer, S. & Hamman, J. xarray: N-D labeled Arrays and Datasets in Python. *J. Open Res. Softw.* **5,** 10 (2017).

73. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12,** 2825–2830 (2011).

74. Waskom, M. L. seaborn: statistical data visualization. *J. Open Source Softw.* **6,** 3021 (2021).

75. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9,** 90–95 (2007).

76. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K. & Yutani, H. Welcome to the Tidyverse. *J. Open Source Softw.* **4,** 1686 (2019).

77. Wang, L.-G., Lam, T. T.-Y., Xu, S., Dai, Z., Zhou, L., Feng, T., Guo, P., Dunn, C. W., Jones, B. R., Bradley, T., Zhu, H., Guan, Y., Jiang, Y. & Yu, G. Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. *Mol. Biol. Evol.* **37,** 599–603 (2020).

78. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8,** 28–36 (2017).

Appendix A. Supplemental Material for Chapter 1 Applications and comparison of dimensionality reduction methods for microbiome data

AA.1. Supplemental Tables

Table AA.1.S1: Common characteristics of strategies for dimensionality reduction address different aspects of the data.

Term	Definition
Compositionally aware	Transforms data to account for non-independence of features in sequence count data.
Pseudo-counts or imputation	Requires no/minimal zeroes in the feature table due to numerical issues (such as logarithm transform being undefined on zeroes).
Able to incorporate phylogeny	Method is calculated with awareness of how each sampled microbial community is evolutionarily represented relative to other samples.
Operates on beta-diversity dissimilarities	Dimensionality reduction step is performed on pairwise dissimilarities (arbitrary metric) between samples, rather than the feature table itself.
Linear	Lower dimensional coordinates are computed via linear transform of features.
Repeated measures	Subjects are sampled multiple times. Commonly sampled longitudinally.
Feature relationships are interpretable	The method indicates the relevance of input microbial features with regard to its output coordinates.
Supervised component	Method takes explanatory sample variables as an additional input

Table AA.1.S2: Dimensionality reduction methods each have their own characteristics. x indicates that the characteristic applies to the method. Examples of software capable of performing each method are included in the last column.

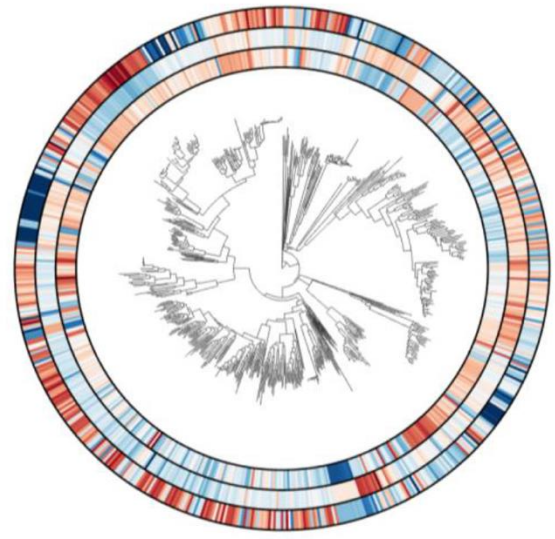
	Compositionally aware	Avoids pseud-counts or imputation	Able to incorporate phylogeny	Operates on beta-diversity dissimilarities	Linear	Repeated Measures	Feature relationships are interpretable	Supervised component	Software
PCoA		x	x	x	x				QIIME 2, CRAN phyloseq, mothur
PCA		x			x		x		scikit-learn, R built-in, mothur
UMAP		x	x	x					umap-learn, CRAN umap, QIIME 2
t-SNE		x	x	x					scikit-learn, CRAN tsne
nMDS		x	x	x					scikit-learn, CRAN vegan, mothur, CRAN phyloseq
CCA					x		x	x	scikit-bio, CRAN vegan, CRAN phyloseq
PLS-DA					x		x	x	CRAN mixOmics
Aitchison PCA	x				x		x		scikit-bio, QIIME 2
RPCA	x	x			x		x		~~DEICODE,~~ gemelli, QIIME 2, vegan
CTF	x	x			x	x	x		Gemelli, QIIME 2

Appendix B. Supplemental Material for Chapter 3 Paired microbiome and metabolome analyses associate bile acid changes with colorectal cancer progression

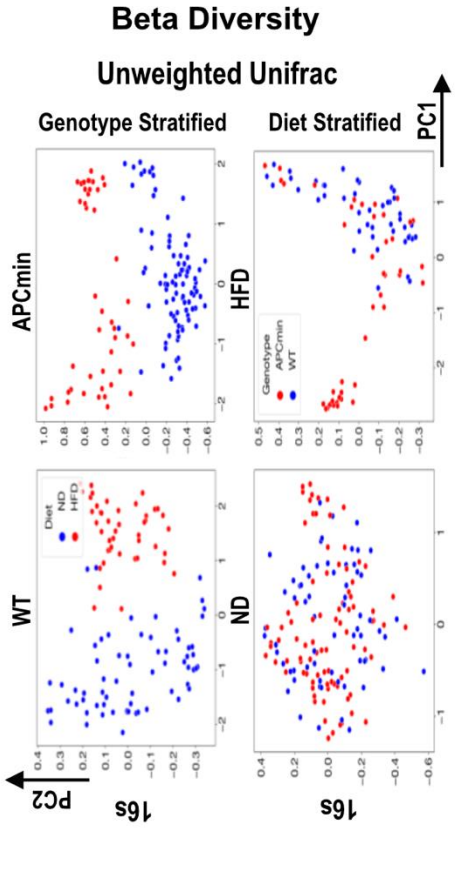
AB.1. Supplemental Figures

Figure AB.1.S1: Genetics and diet reshape the gut microbiome (A) Unweighted Unifrac measures of beta-diversity of wild-type (WT) and APCmin/+ mice maintained on normal-chow diet (ND) and high fat diet (HFD). Metrics from 16s rRNA sequencing data are stratified by genotype and diet factors and visualized using Principal Coordinate Analysis (PCoA). (B) Phylogenetic tree of metagenomics microbes with rings representing Songbird differentials. Inner, middle, and outer rings represent HFD association, APCmin/+ association, and association with the interaction term HFD, respectively.

Sup Figure 1



B



A

Figure AB.1.S2: Global metabolite changes associated with dietary and genetic risk factors
(A) Schematic workflow of the metabolomics study. (B) Principal component analysis (PCA) of serum metabolites from WT and APCmin/+ mice maintained on ND and HFD. (C) Permutation results indicate the validity of PLS-DA analysis. A 100-time permutation test was conducted, and the PLS-DA model has a model p-value < 0.01, suggesting that the model is valid.

Sup Figure 2

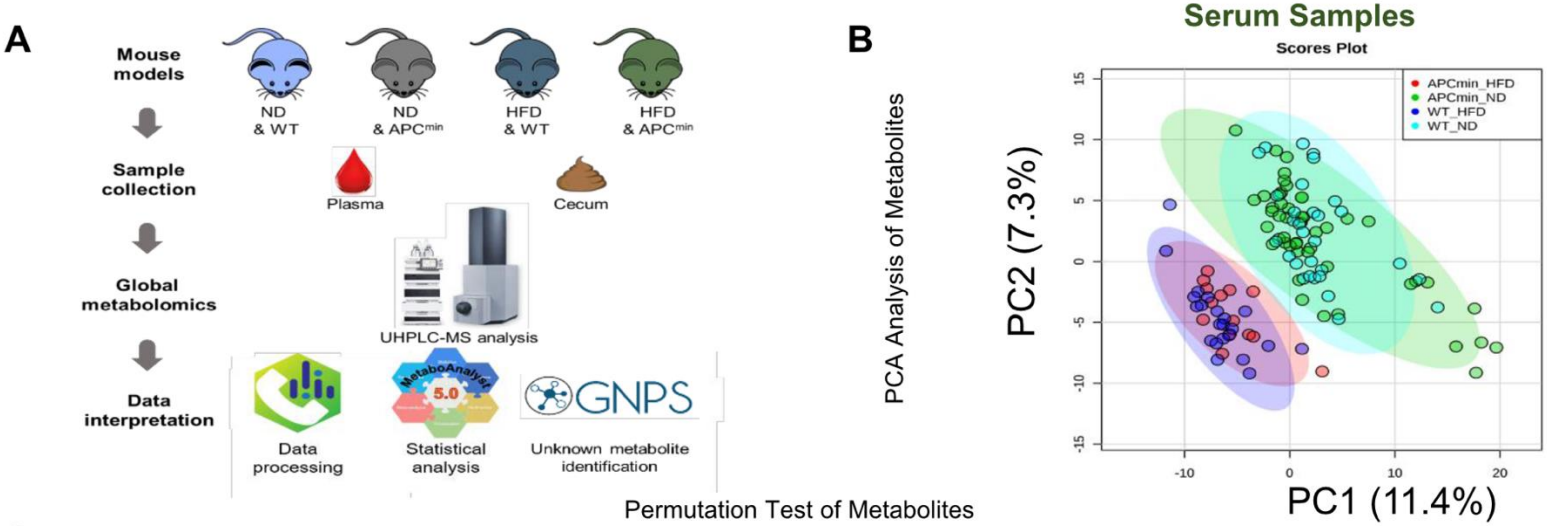
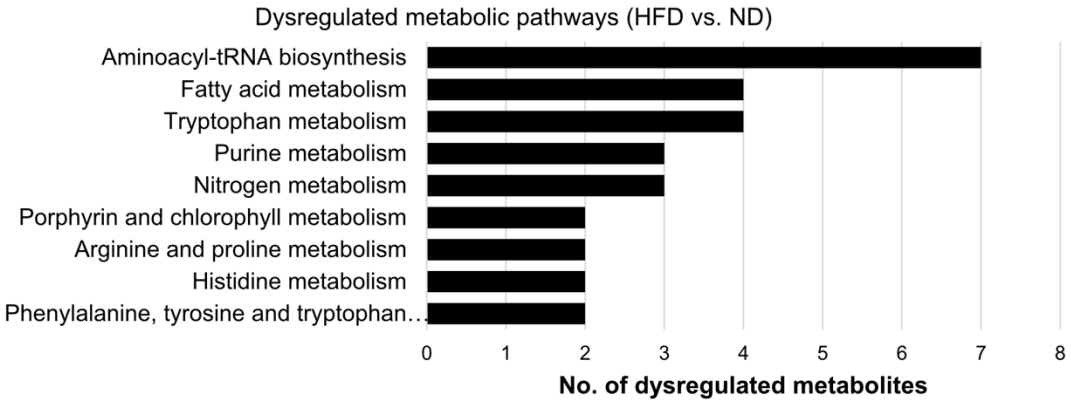

A
Mouse models
ND & WT
ND & APCmin
HFD & WT
HFD & APCmin
Sample collection
Plasma
Cecum
Global metabolomics
UHPLC-MS analysis
Data interpretation
Data processing
Statistical analysis
Unknown metabolite identification
B
Serum Samples
Scores Plot
PCA Analysis of Metabolites
APCmin_HFD
APCmin_ND
WT_HFD
WT_ND
PC2 (7.3%)
PC1 (11.4%)
Permutation Test of Metabolites
C
Cecum Samples
Serum Samples
Frequency
Observed statistic
p < 0.01 (0/100)
Permutation test statistics

Figure AB.1.S3: Differential network analysis of metabolomes(A) Over 110 metabolites were significantly dysregulated in WT maintained on ND or HFD. Pathway enrichment analysis showed that 10 metabolic pathways were significantly dysregulated with pathway p-value smaller than 0.01. (B) Box plots of significantly altered metabolites in the comparison of WT and APCmin/+ mice on ND.(C) Significantly changes in metabolites show correlation patterns with dietary factors (HFD vs. ND). Correlation-based metabolic network analysis was visualized using the Metscape plugin available in Cytoscape. Each node represents one metabolite and the edge between two nodes represents the differentiated correlation coefficient ($p < 0.05$) between two metabolites (HFD vs. ND). Nodes in the solid circle are KEGG metabolites. The edge represents the differentiated correlation coefficients.

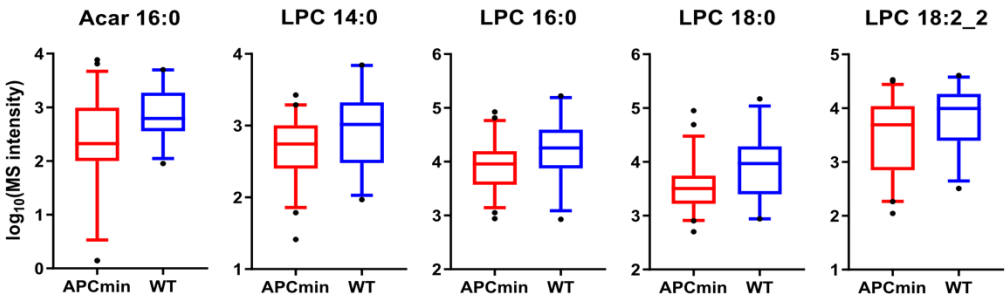
Sup Figure 3

A



A total of 110 metabolites have fold change ≥ 1.5 and p-value ≤ 0.05

B



Acar 16:0	LPC 14:0	LPC 16:0	LPC 18:0	LPC 18:2_2
FC: 0.60	FC: 0.42	FC: 0.43	FC: 0.38	FC: 0.60
p-value: 0.002	p-value: 0.04	p-value: 0.008	p-value: 0.003	p-value: 0.02

C **Diet effects on cecum metabolites (HFD vs ND on WT and APCmin mice)**

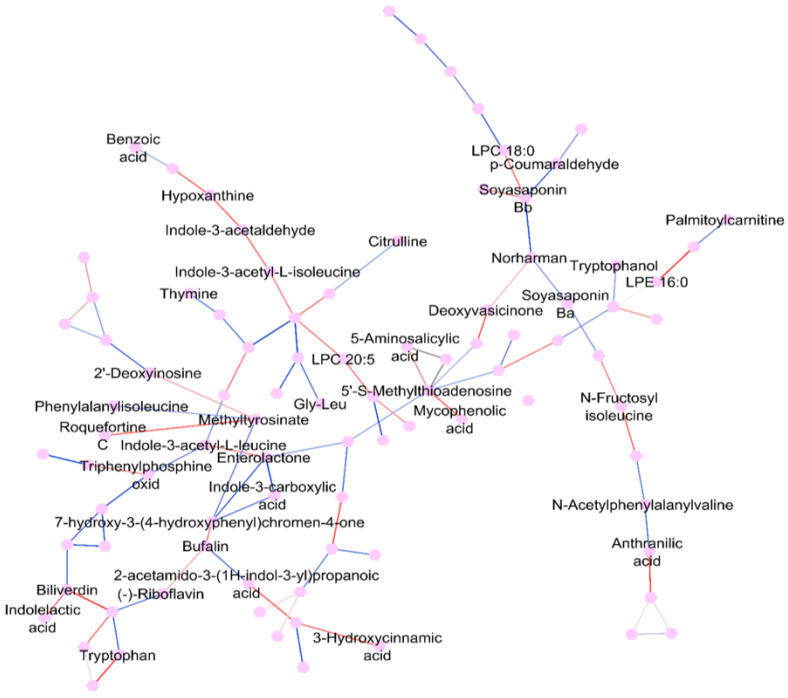
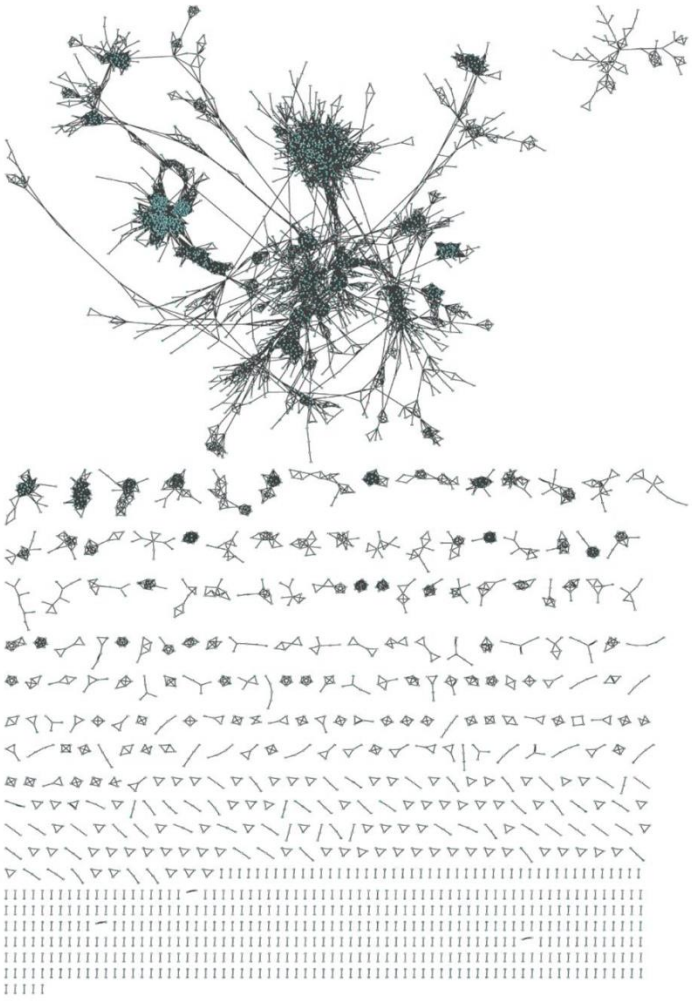


Figure AB.1.S4: Discovery of novel AA-BAs. (A) Molecular network analysis of MS2 spectral similarities in the cecum metabolomics dataset.

Sup Figure 4

A Molecular network analysis for spectral similarity



Diet effects on cecum metabolites (HFD vs ND on WT and APCmin mice)

Figure AB.1.S5 Correlation of fecal bile acid levels with adenocarcinoma progression in APC$^{min/+}$ mice (A) Progressive changes in total fecal BA levels in PCmin/+ mice. (B) Pie charts showing compositional changes in fecal BAs during intestinal tumor initiation in APCmin/+ mice. (C-D) Progressive changes in bacterially-mediated conversion of tauro-beta-muricholic acid (T-βMCA) to beta-muricholic acid (βMCA) and omega muricholic acid (ωMCA)) in serum (C) and feces (D). (E-F) Temporal changes in fecal bacterial load and ωMCA (E), and βMCA (F) levels in APCmin/+ mice on ND. (G-H) Temporal changes of fecal ωMCA (G) and CA (H) levels in WT and APCmin/+ mice upon ND and HFD during tumor progression.

Sup Figure 5

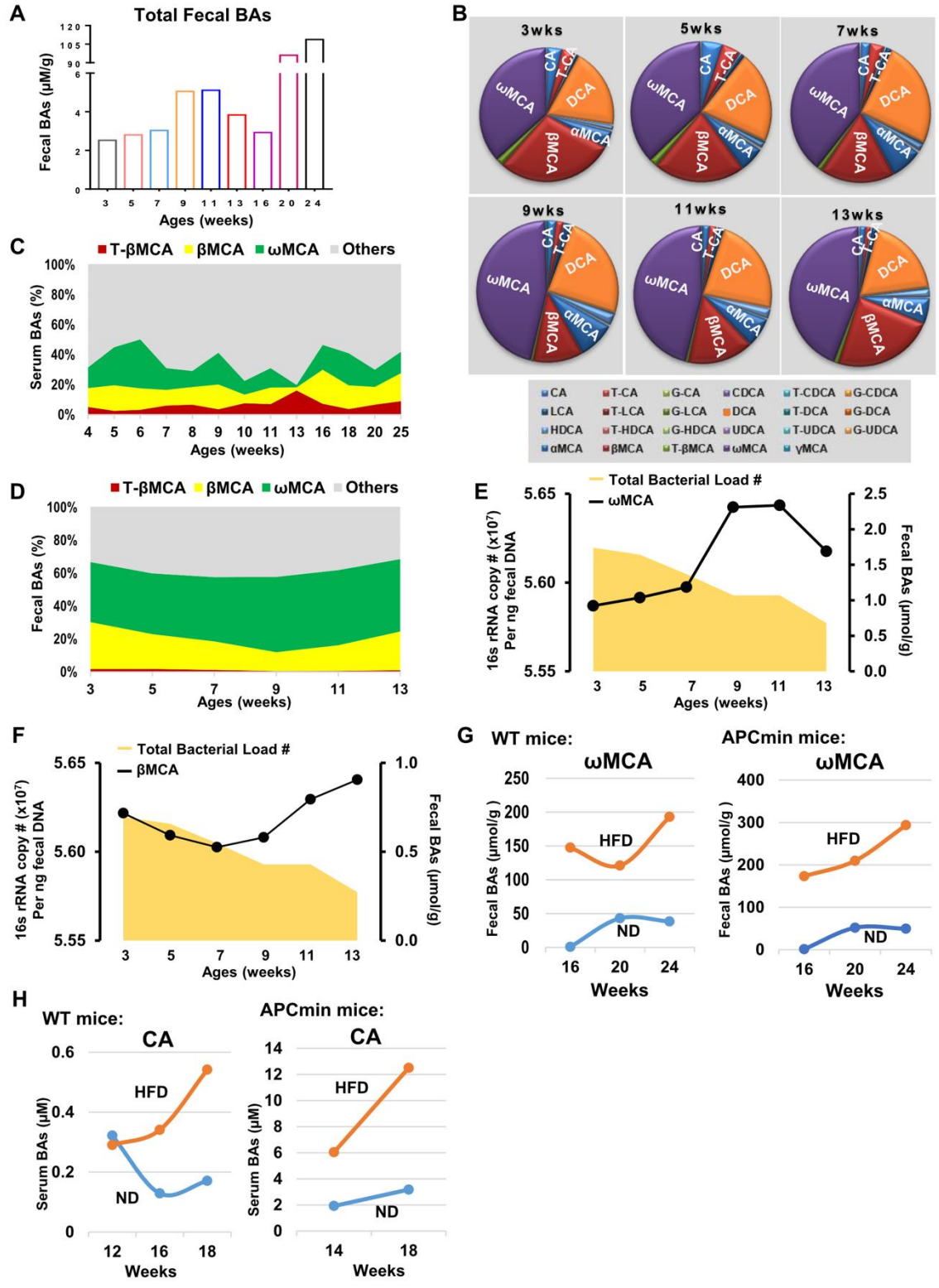

A
Total Fecal BAs
Fecal BAs (μM/g)
Ages (weeks)
B
3wks
5wks
7wks
9wks
11wks
13wks
CA
T-CA
DCA
ωMCA
βMCA
αMCA
CA T-CA G-CA CDCA T-CDCA G-CDCA
LCA T-LCA G-LCA DCA T-DCA G-DCA
HDCA T-HDCA G-HDCA UDCA T-UDCA G-UDCA
αMCA βMCA T-βMCA ωMCA γMCA
C
T-βMCA βMCA ωMCA Others
Serum BAs (%)
Ages (weeks)
D
T-βMCA βMCA ωMCA Others
Fecal BAs (%)
Ages (weeks)
E
Total Bacterial Load #
ωMCA
16s rRNA copy # (x10⁷) Per ng fecal DNA
Fecal BAs (μmol/g)
Ages (weeks)
F
Total Bacterial Load #
βMCA
16s rRNA copy # (x10⁷) Per ng fecal DNA
Fecal BAs (μmol/g)
Ages (weeks)
G
WT mice:
APCmin mice:
ωMCA
Fecal BAs (μmol/g)
HFD
ND
Weeks
H
WT mice:
APCmin mice:
CA
Serum BAs (μM)
HFD
ND
Weeks

Figure AB.1.S6: Fecal BAs are influenced by diet and genetics (A) Experimental schemes of vehicle and FexD treatment (50mg/kg/day) of WT and APCmin/+ mice on ND (also described in Figure 3H). (B) Specific fecal bile acid levels in 16-week-old ND-fed WT and APCmin/+ mice after indicated treatments. (C) Experimental schemes of vehicle and FexD treatment (50mg/kg/day) of WT and APCmin/+ mice on HFD (also described in Figure 3H). (D) Specific fecal bile acid levels in 14-week-old HFD-fed mice WT and APCmin/+ mice after indicated treatments. (E) Fecal levels of conjugated cholic acid in WT and APCmin/+ mice on ND and HFD. (F) Scatterplot of metabolites separated by Spearman correlation analysis of Mmvec PC1 results and Songbird results by diet differentials in APCmin/+ mice. Bile acids are colored in blue while amino acids are colored in red. Metabolites that belong to neither of these groups are colored gray. Density plots are shown on the margins colored by metabolite type.(G) Cluster map of Mmvec microbe-metabolite conditional probabilities focusing on bile acids. Log-conditional probabilities were Z-scored across microbes. Margins are annotated by assigned cluster (x-axis is microbes, y-axis is bile acids).

Sup Figure 6

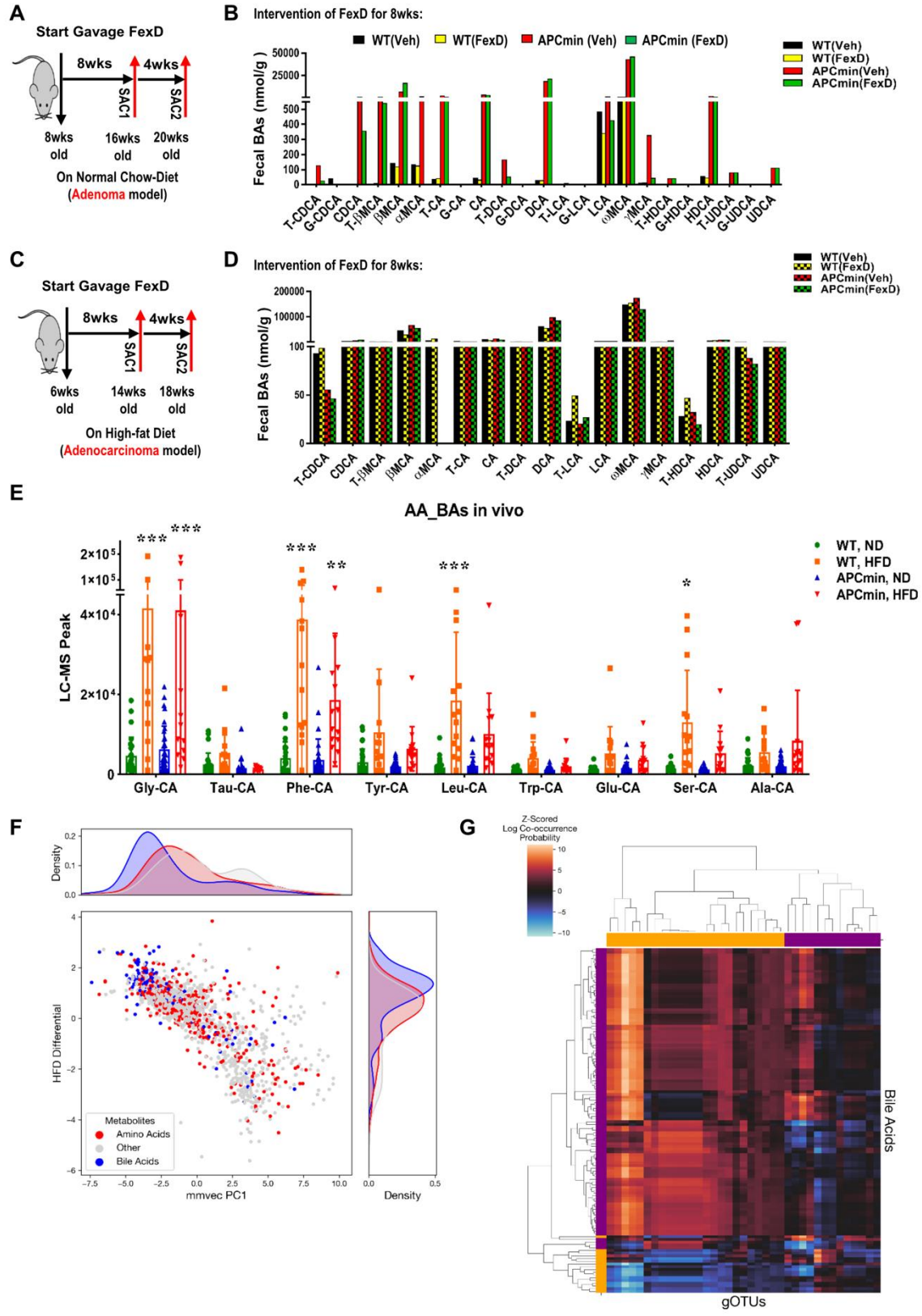

A
Start Gavage FexD
8wks
4wks
SAC1
SAC2
8wks old
16wks old
20wks old
On Normal Chow-Diet
(Adenoma model)
B
Intervention of FexD for 8wks:
WT(Veh)
WT(FexD)
APCmin (Veh)
APCmin (FexD)
Fecal BAs (nmol/g)
T-CDCA
G-CDCA
CDCA
T-βMCA
βMCA
αMCA
T-CA
G-CA
CA
T-DCA
G-DCA
DCA
T-LCA
G-LCA
LCA
ωMCA
γMCA
T-HDCA
G-HDCA
HDCA
T-UDCA
G-UDCA
UDCA
C
Start Gavage FexD
6wks old
14wks old
18wks old
On High-fat Diet
(Adenocarcinoma model)
D
E
AA_BAs in vivo
WT, ND
WT, HFD
APCmin, ND
APCmin, HFD
LC-MS Peak
Gly-CA
Tau-CA
Phe-CA
Tyr-CA
Leu-CA
Trp-CA
Glu-CA
Ser-CA
Ala-CA
F
Density
HFD Differential
Metabolites
Amino Acids
Other
Bile Acids
mmvec PC1
G
Z-Scored
Log Co-occurrence
Probability
Bile Acids
gOTUs

Figure AB.1.S7: Functionality and bacterial origins of non-classic amino acid-conjugated cholic acid(A) Dose-dependent activation of exogenous human FXR by amino acid conjugated cholic acid species. Luciferase activity in HEK293 cells expressing a luciferase reporter gene functionally linked to an FXR-responsive element (FXRE-Luc). (B) Dose-dependent activation of exogenous human TGR5 by amino acid conjugated cholic acid species. Luciferase activity in HEK293 cells expressing a luciferase reporter gene functionally linked to a cAMP-responsive element which is downstream of TGR5. (C-D) Dose-dependent proliferation of intestinal organoids from WT mice treated with indicated conjugated cholic acid, measured by luminescent cell viability assay. (E-F) Dose-dependent proliferation of intestinal organoids from $APC^{min/+}$ mice treated with indicated conjugated cholic acid, measured by luminescent cell viability assay. (G) Relative expression of FXR and TGR5 target genes, and intestinal stem cell marker genes in intestinal organoids from $APC^{min/+}$ mice treated with 10μM conjugated cholic acids. (H-I) Dose-dependent increases in TCF/LEF assay in HT29 cells treated with indicated conjugated cholic acid species, measured in Relative Luminescence Units (RLU). The TGR5 ligand INT-777 serves as a positive control. (J) Permeability coefficients (Papps) of indicated conjugated cholic acid species measured in Caco2 cells. Atenolol and propranolol serve as a negative and positive controls, respectively. Digoxin serves as a positive control for P-glycoprotein-mediated efflux. Data represent the mean ± SEM. Student's unpaired t-test or one-way Anova test followed by multiple comparation. *$p<0.05$; **$p<0.01$; ***$p<0.005$.

Sup Figure 7

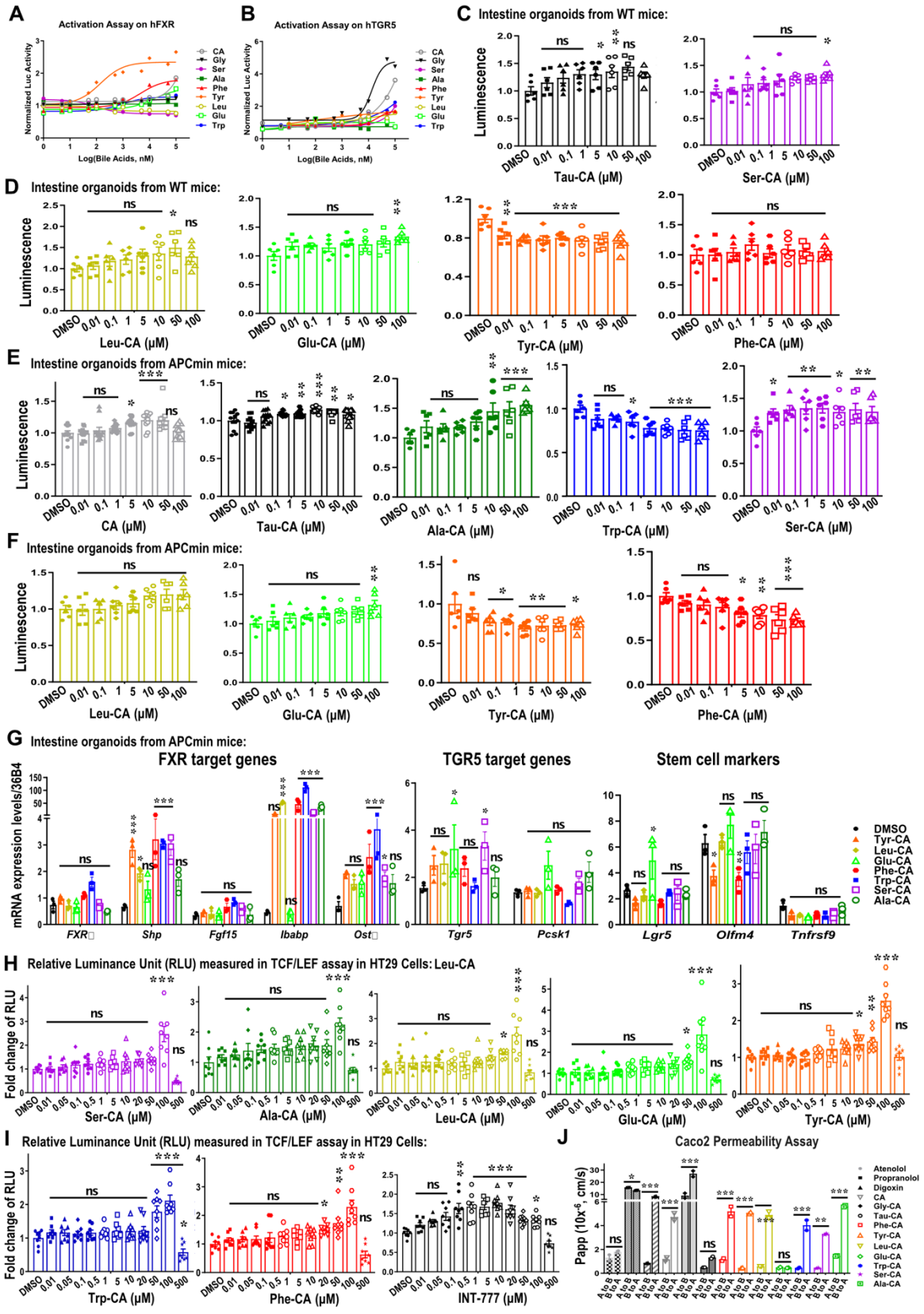
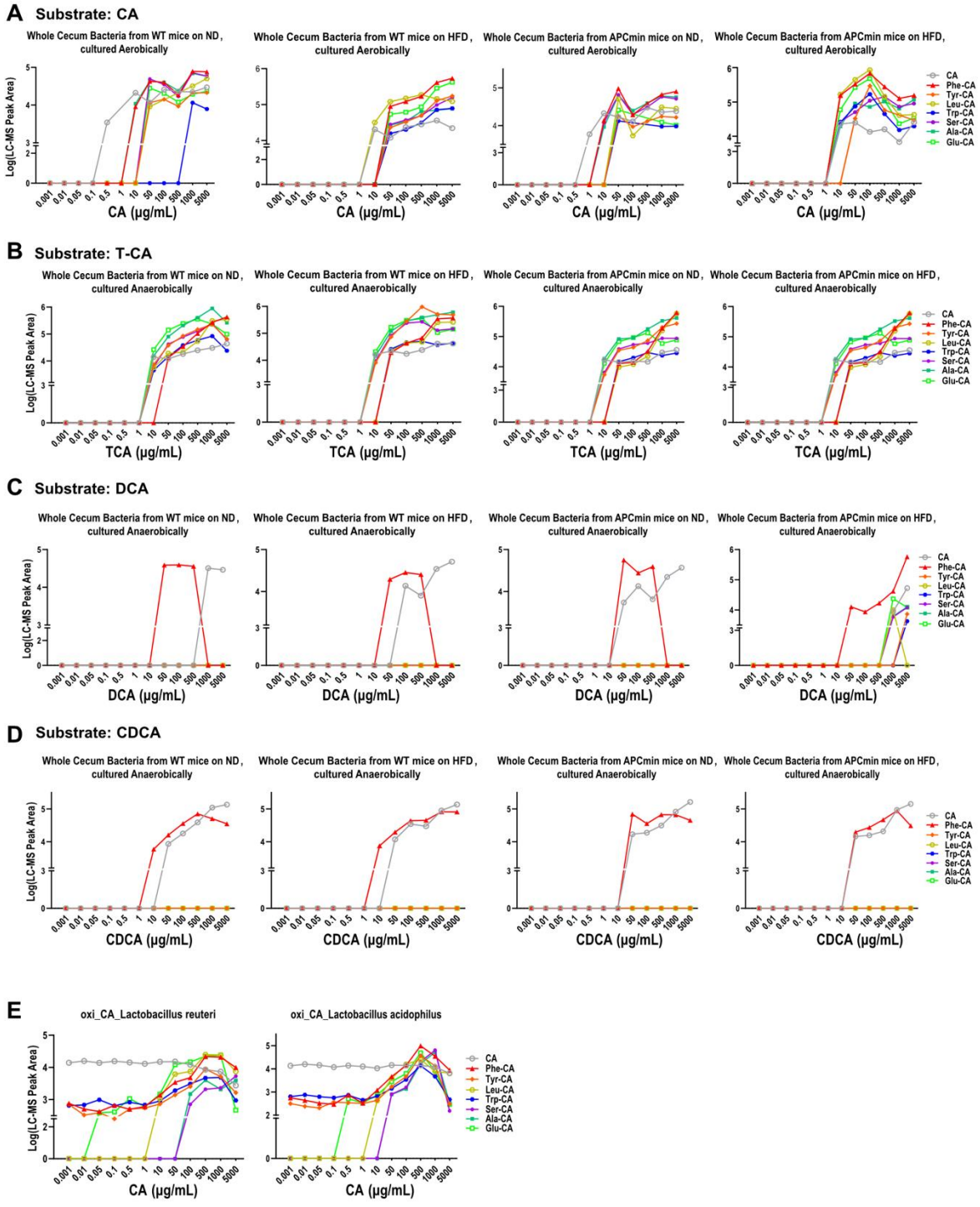

A
Activation Assay on hFXR
Normalized Luc Activity
Log(Bile Acids, nM)
CA
Gly
Ser
Ala
Phe
Tyr
Leu
Glu
Trp
B
Activation Assay on hTGR5
C Intestine organoids from WT mice:
Luminescence
Tau-CA (μM)
Ser-CA (μM)
D Intestine organoids from WT mice:
Leu-CA (μM)
Glu-CA (μM)
Tyr-CA (μM)
Phe-CA (μM)
E Intestine organoids from APCmin mice:
CA (μM)
Tau-CA (μM)
Ala-CA (μM)
Trp-CA (μM)
Ser-CA (μM)
F Intestine organoids from APCmin mice:
Leu-CA (μM)
Glu-CA (μM)
Tyr-CA (μM)
Phe-CA (μM)
G Intestine organoids from APCmin mice:
FXR target genes
TGR5 target genes
Stem cell markers
mRNA expression levels/36B4
FXR□
Shp
Fgf15
Ibabp
Ost□
Tgr5
Pcsk1
Lgr5
Olfm4
Tnfrsf9
DMSO
Tyr-CA
Leu-CA
Glu-CA
Phe-CA
Trp-CA
Ser-CA
Ala-CA
H Relative Luminance Unit (RLU) measured in TCF/LEF assay in HT29 Cells: Leu-CA
Fold change of RLU
Ser-CA (μM)
Ala-CA (μM)
Leu-CA (μM)
Glu-CA (μM)
Tyr-CA (μM)
I Relative Luminance Unit (RLU) measured in TCF/LEF assay in HT29 Cells:
Trp-CA (μM)
Phe-CA (μM)
INT-777 (μM)
J Caco2 Permeability Assay
Papp (10⁻⁶, cm/s)
Atenolol
Propranolol
Digoxin
CA
Gly-CA
Tau-CA
Phe-CA
Tyr-CA
Leu-CA
Glu-CA
Trp-CA
Ser-CA
Ala-CA

Figure AB.1.S8: Putative bacterial origins of non-classic amino acid-conjugated cholic acid. (A) Dose-dependent generation of conjugated cholic acid species in aerobic cultures of cecal bacteria from WT and APCmin/+ mice on ND and HFD. Cultures were supplemented with increasing concentrations of cholic acid (CA) for 48h prior to mass spectral analysis. (B) Dose-dependent generation of conjugated cholic acid species in anaerobic cultures of cecal bacteria from WT and APCmin/+ mice on ND and HFD. Cultures were supplemented with increasing concentrations of taurocholic acid (T-CA) for 48h prior to mass spectral analysis. (C) Dose-dependent generation of conjugated cholic acid species in anaerobic cultures of cecal bacteria from WT and APCmin/+ mice on ND and HFD. Cultures were supplemented with increasing concentrations of deoxycholic acid (DCA) for 48h prior to mass spectral analysis. (D) Dose-dependent generation of conjugated cholic acid species in anaerobic cultures of cecal bacteria from WT and APCmin/+ mice on ND and HFD. Cultures were supplemented with increasing concentrations of chenodeoxycholic acid (CDCA) for 48h prior to mass spectral analysis. (E) Dose-dependent generation of conjugated cholic acid species in anaerobic cultures of Lactobacillus reuteri and Lactobacillus acidophilus. Cultures were supplemented with increasing concentrations of cholic acid (CA) for 48h prior to mass spectral analysis.

Sup Figure 8


A Substrate: CA
Whole Cecum Bacteria from WT mice on ND, cultured Aerobically
Whole Cecum Bacteria from WT mice on HFD, cultured Aerobically
Whole Cecum Bacteria from APCmin mice on ND, cultured Aerobically
Whole Cecum Bacteria from APCmin mice on HFD, cultured Aerobically
B Substrate: T-CA
Whole Cecum Bacteria from WT mice on ND, cultured Anaerobically
Whole Cecum Bacteria from WT mice on HFD, cultured Anaerobically
Whole Cecum Bacteria from APCmin mice on ND, cultured Anaerobically
Whole Cecum Bacteria from APCmin mice on HFD, cultured Anaerobically
C Substrate: DCA
Whole Cecum Bacteria from WT mice on ND, cultured Anaerobically
Whole Cecum Bacteria from WT mice on HFD, cultured Anaerobically
Whole Cecum Bacteria from APCmin mice on ND, cultured Anaerobically
Whole Cecum Bacteria from APCmin mice on HFD, cultured Anaerobically
D Substrate: CDCA
Whole Cecum Bacteria from WT mice on ND, cultured Anaerobically
Whole Cecum Bacteria from WT mice on HFD, cultured Anaerobically
Whole Cecum Bacteria from APCmin mice on ND, cultured Anaerobically
Whole Cecum Bacteria from APCmin mice on HFD, cultured Anaerobically
E
oxi_CA_Lactobacillus reuteri
oxi_CA_Lactobacillus acidophilus
Log(LC-MS Peak Area)
CA (μg/mL)
TCA (μg/mL)
DCA (μg/mL)
CDCA (μg/mL)
0.001 0.01 0.05 0.1 0.5 1 10 50 100 500 1000 5000
CA
Phe-CA
Tyr-CA
Leu-CA
Trp-CA
Ser-CA
Ala-CA
Glu-CA

AB.2. Supplemental Tables

Table AC.2.S1: Primer lists.

16s-515F	GTGCCAGCMGCCGCGGTAA
16s-805R	GACTACCAGGGTATCTAAT
16s-515F	GTGCCAGCMGCCGCGGTAA
16s-805R	GACTACCAGGGTATCTAAT
Ileibacterium valens -F1	GCCAGAAAGTCACGGCTAAC
Ileibacterium valens -R1	CGCCCTCTCCTGTACTCAAG
Ileibacterium valens -F2	CCCTAGTAGTCCACGCCGTA
Ileibacterium valens -R2	TAAGGTTCTTCGCGTTGCTT
Ruminococcus gnavus-F1	CACATTGGGACTGAGACACG
Ruminococcus gnavus-R1	TAAATCCGGATAACGCTTGC
Ruminococcus gnavus-F2	CTTGCTGGACGATGACTGAC
Ruminococcus gnavus-R2	CTCCGATTAAAGAGCGGTCAGA
Clostridium scindens-F1	TAGTCCACCTGGGGAGTACG
Clostridium scindens-R1	CGATGTTCCGAAGAAAGAGC

Appendix C. Supplemental Material for Chapter 4 BIRDMAn: A Bayesian differential abundance framework that enables robust inference of host-microbe associations

AC.1. Supplemental Figures

Figure AC.1.S1: Extended antibiotics analysis.(a) Phylogenetic tree of all OTUs with a heatmap of posterior means for each time-interval contrast. OTUs assigned to one of the top 8 most abundant genera are annotated through the colored strip. (b) When BIRDMAn is used to account for per-subject variation, log-ratio comparisons of the top 40 OTUs vs. bottom OTUs are associated with the difference between each time point and the next one. For each of these contrasts, the log-ratios of the samples between the two time intervals were compared using a one-sided t-test. Plots are annotated with p-values. Different taxa contribute to the log ratios for each contrast. (c) Overall performance of BIRDMAn classifier on predicting the antibiotics time interval using the log-ratios. The classifier prediction accuracies shown are aggregated across folds and repeats from repeated k-fold cross-validation. (d) Accuracy of the multinomial classifier by number of OTUs used in log-ratio calculations. Points represent mean accuracy across cross-validation iterations and shaded areas represent ±1 standard deviation. (e) Distribution of Gram positive and Gram negative OTUs associated with FirstCp and SecondCp log-ratios

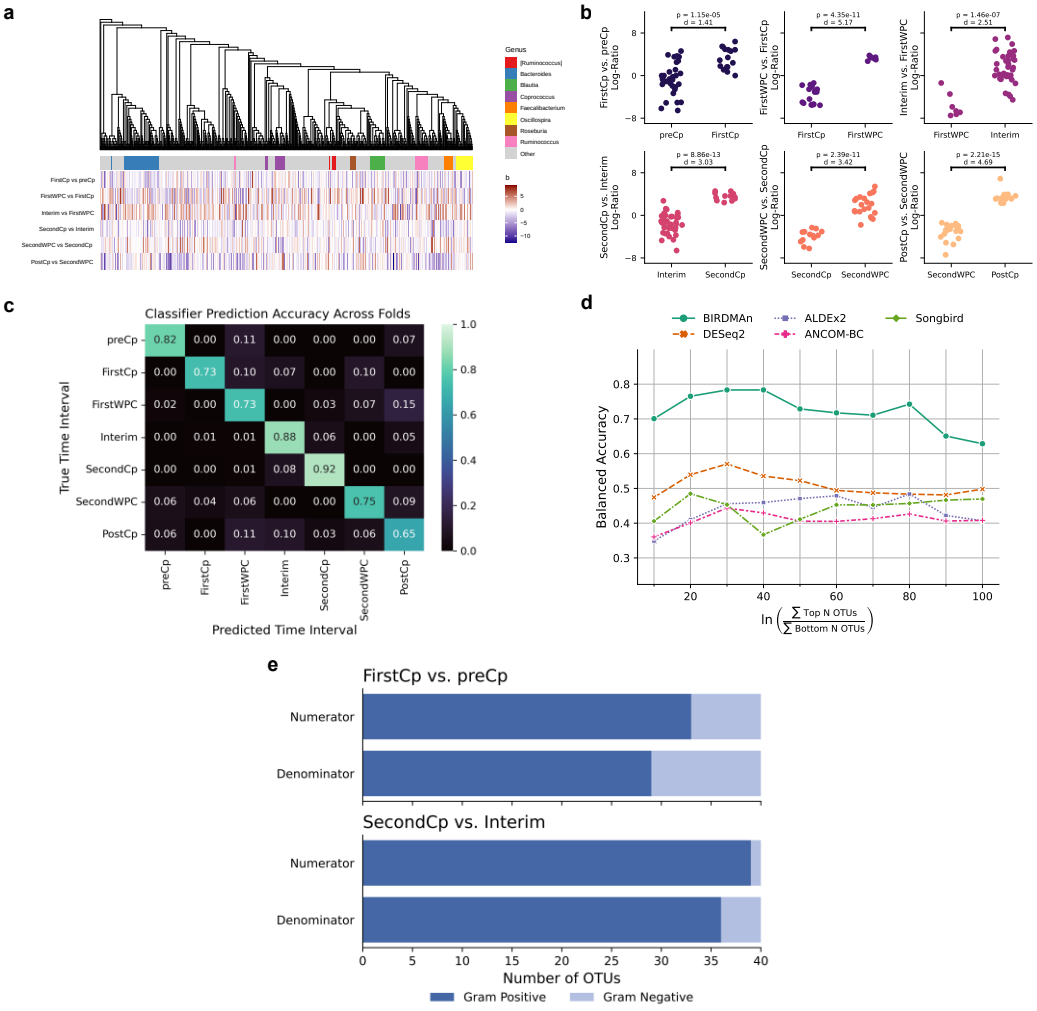
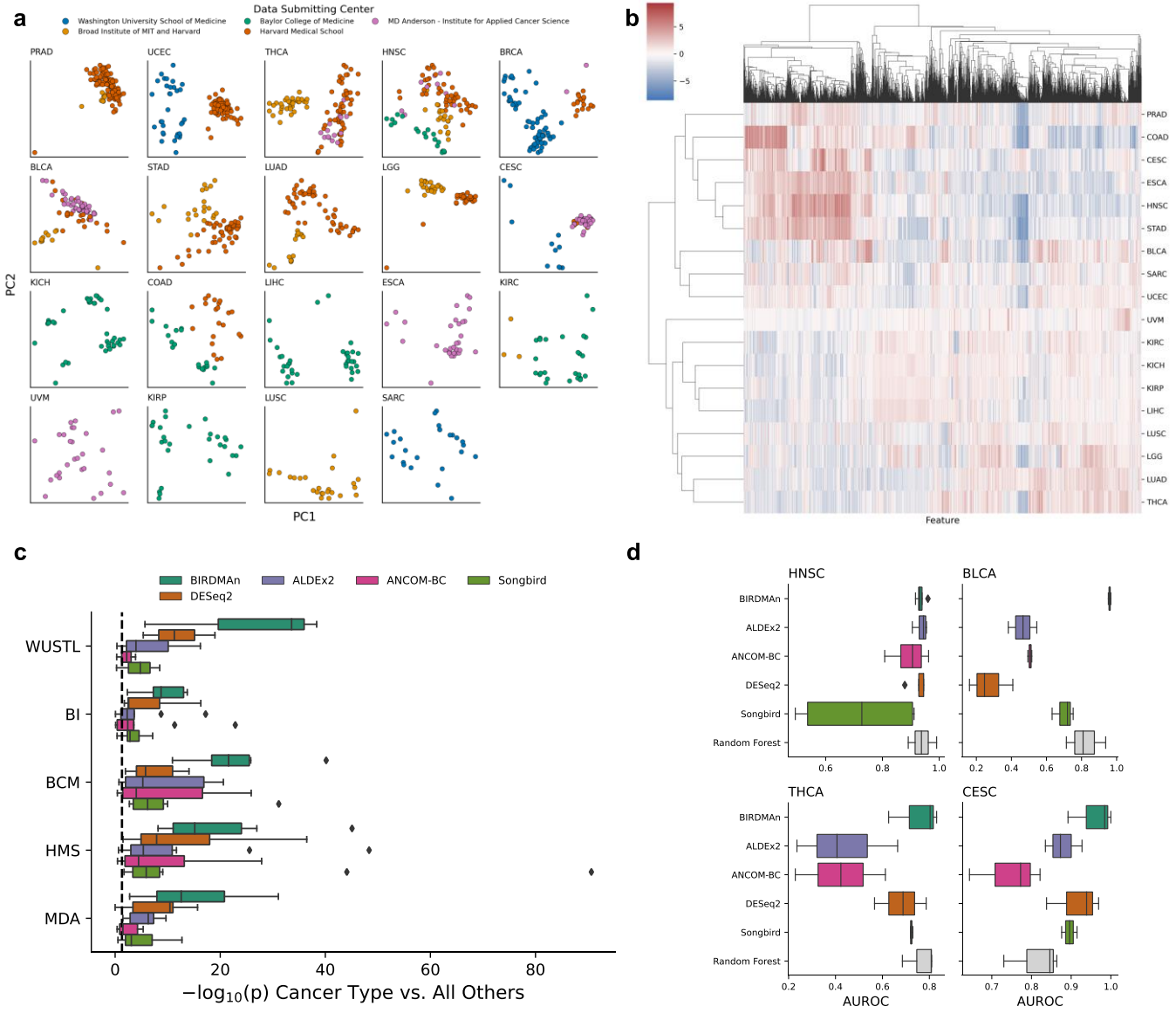

a
Genus
[Ruminococcus]
Bacteroides
Blautia
Coprococcus
Faecalibacterium
Oscillospira
Roseburia
Ruminococcus
Other
FirstCp vs preCp
FirstWPC vs FirstCp
Interim vs FirstWPC
SecondCp vs Interim
SecondWPC vs SecondCp
PostCp vs SecondWPC
b
FirstCp vs. preCp Log-Ratio
FirstWPC vs. FirstCp Log-Ratio
Interim vs. FirstWPC Log-Ratio
SecondCp vs. Interim Log-Ratio
SecondWPC vs. SecondCp Log-Ratio
PostCp vs. SecondWPC Log-Ratio
p = 1.15e-05
d = 1.41
p = 4.35e-11
d = 5.17
p = 1.46e-07
d = 2.51
p = 8.86e-13
d = 3.03
p = 2.39e-11
d = 3.42
p = 2.21e-15
d = 4.69
preCp
FirstCp
FirstWPC
Interim
SecondCp
SecondWPC
PostCp
c
Classifier Prediction Accuracy Across Folds
True Time Interval
Predicted Time Interval
preCp 0.82 0.00 0.11 0.00 0.00 0.00 0.07
FirstCp 0.00 0.73 0.10 0.07 0.00 0.10 0.00
FirstWPC 0.02 0.00 0.73 0.00 0.03 0.07 0.15
Interim 0.00 0.01 0.01 0.88 0.06 0.00 0.05
SecondCp 0.00 0.00 0.01 0.08 0.92 0.00 0.00
SecondWPC 0.06 0.04 0.06 0.00 0.00 0.75 0.09
PostCp 0.06 0.00 0.11 0.10 0.03 0.06 0.65
d
BIRDMAn
ALDEx2
Songbird
DESeq2
ANCOM-BC
Balanced Accuracy
ln(∑ Top N OTUs / ∑ Bottom N OTUs)
e
FirstCp vs. preCp
Numerator
Denominator
SecondCp vs. Interim
Numerator
Denominator
Number of OTUs
Gram Positive
Gram Negative

Figure AC.1.S2: Extended cancer analysis.(a) RPCA projection of the original feature table subset to each individual cancer type. Points are colored by data submitting centers, showing that many cancer types exhibit strong separation by batch. (b) Posterior means (CLR) of feature differentials clustered by cancer type. (c) Log-ratios identified by BIRDMAn separate each tumor type from all others when stratified by center. Dashed line represents a t-test p-value at p = 0.05. (d) Performance of leave-one-center-out cross-validation logistic regression classifier AUROC of all methods.


a
Data Submitting Center
Washington University School of Medicine
Broad Institute of MIT and Harvard
Baylor College of Medicine
Harvard Medical School
MD Anderson - Institute for Applied Cancer Science
PRAD
UCEC
THCA
HNSC
BRCA
BLCA
STAD
LUAD
LGG
CESC
KICH
COAD
LIHC
ESCA
KIRC
UVM
KIRP
LUSC
SARC
PC2
PC1
b
PRAD
COAD
CESC
ESCA
HNSC
STAD
BLCA
SARC
UCEC
UVM
KIRC
KICH
KIRP
LIHC
LUSC
LGG
LUAD
THCA
Feature
c
BIRDMAn
ALDEx2
ANCOM-BC
Songbird
DESeq2
WUSTL
BI
BCM
HMS
MDA
0
20
40
60
80
−log10(p) Cancer Type vs. All Others
d
HNSC
BLCA
THCA
CESC
BIRDMAn
ALDEx2
ANCOM-BC
DESeq2
Songbird
Random Forest
AUROC
AUROC