Title
Computational approaches for metagenomic analysis of the microbiome

Author
Briscoe, Leah Pritnapah

Publication Date
2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Computational approaches for metagenomic analysis of the microbiome

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Bioinformatics

by

Leah Pritnapah Briscoe

2023

ABSTRACT OF THE DISSERTATION

Computational approaches for metagenomic

analysis of the microbiome

by

Leah Pritnapah Briscoe

Doctor of Philosophy in Bioinformatics

University of California Los Angeles, 2023

Professor Nandita Garud, Co-Chair

Professor Eran Halperin, Co-Chair

The microbiome is a community of microorganisms living in our bodies and throughout the environment. The genomic data researchers can extract from microbiomes, known as metagenomic data, can be used to predict traits about a host or environment. By identifying microbiome biomarkers associated with disease or health, researchers can develop better therapeutics for microbiome-associated diseases. However, metagenomic data is commonly affected by technical variables unrelated to the phenotype of interest, such as sequencing protocol, which can make it difficult to predict phenotype and find biomarkers of disease. Here, we evaluate methods to remove background noise due to technical variables unrelated to the phenotype of interest, such as sequencing protocol, and thereby improving our ability to find accurate biomarkers of human disease. Also crucial in understanding host health is elucidating the sources of their microbiomes, as it allows researchers to understand the dynamics behind how microbial communities form and how they respond to changing environments. In this work,

we introduce a method to use metagenomic variants obtained from hundreds of species in microbiome data to perform source tracking, which is a method of estimating colonization sources for a sample of interest. These analyses shed light on phenomena like the colonization of the early infant gut microbiome, or spatial patterns in the ocean microbiomes around the world. Lastly, we analyze metagenomic data to understand how genetic diversity changes along the human gut on the species, strain and gene level. In sum, this work leverages the genomic information contained in our microbiomes to find universal patterns in microbiomes, allowing us to better understand the relationship between microbiome and phenotypes, the colonization sources of microbiomes, and also the colonization dynamics on the species and strain level.

The dissertation of Leah Pritnapah Briscoe is approved.

Jingyi Li

Sriram Sankararaman

Nandita Garud, Committee Co-Chair

Eran Halperin, Committee Co-Chair

University of California Los Angeles

2023

To my family, for their relentless love and support

Table of Contents

List of Figures

List of Tables

Acknowledgements

Every scientist stands on the shoulders of giants. In my case, these giants were the brilliant scholars who have helped shaped me into who I am today. Firstly, I thank Dr. Nandita R. Garud for her incredible wisdom, patience and foresight as my faculty advisor and career mentor. I am endlessly amazed at Dr. Garud's talent and energy to ask fascinating scientific questions and teaching me to be the best scientist I can be. I never felt like I was standing alone with Dr. Garud there to cheer me on and advocate for me. Since joining as faculty at UCLA, she created a lab community that I have always dreamed being a part of. I thank my second faculty advisor Dr. Eran Halperin for jump starting my PhD career and allowing me to explore so many exciting areas of bioinformatics and precision medicine. I thank him for advocating for me early on in my program. I thank Dr. Sriram Sankararaman for his generous time in helping me work through many tough scientific problems and teaching me machine learning. I thank Dr. Jingyi Jessica Li for her enthusiasm as an educator and helping me to understand the complexities of modeling. I thank Dr. Brunilda Balliu for the many hours she spent at the chalkboard with me, scaffolding the ideas for my first paper as a doctoral student. When the science got tough, her optimism helped me through.

I would not be where I am today without the mentors who guided me early on in my science career. I thank Dr. Ann Hirsch for welcoming me into her lab – it was a dream come true running experiments with plants and learning how to be a microbiologist. Dr. Hirsch gave me the spark to want to pursue Bioinformatics and be able to use science for the good of humanity. I thank Dr. Eleazar Eskin who founded and developed the undergraduate minor program in Bioinformatics. Thanks to his mentorship I was able to apply to Bioinformatics PhD programs and meet so many

luminaries in genomics at Computational Genomics Summer Institute. I think Amanda Freise for her guidance in navigating graduate school.

From the first days of graduate school, the members of my cohort were a source of solace and laughter: Ha T. Vu, Ruthie Johnson, Sarah Spendlove, Crista Caggiano, Jesse Garcia, Juan De La Hoz, Alec Chiu, Kofi Amoah, Brandon Jew, Mike Thompson, and Tommer Schwarz. I could not have asked for better friends. I thank the more senior graduate students, Lisa Gai, Soo Bin Kwon, Yu-Ching Hsiao, and Chris Robles who encouraged me and inspired me.

I thank my incredibly intelligent lab mates: Ricky Wolff, Mariana Harris, Michael Wasney, Aina Martinez I Zurita, Peter Laurin and lab alum William Shoemaker. I've truly loved working in lab and the happy hours we all had together. I am continuously impressed by your scientific curiosity and questions. Additionally, Michael Wasney has been an amazing scholar to work alongside with.

I thank Sim-Lin Lau of the CS Department and Stephanie Caranica of the EEB Department for helping me with the financial aspect of being a graduate student. I thank Gene Gray for his support as my Student Affairs Officer, who was an amazing resource in navigating the PhD program and frequently planning events for the students.

I thank my life mentor Daisaku Ikeda for showing me how to enjoy life and all of its obstacles. I thank him for his encouragement and inspiring me to become a humanistic scientist.

I would like to thank my family for nurturing me from Day 1. I am so fortunate to have you as my parents, and for all that you have sacrificed to allow me to fight for my dreams. I thank my mother for intervening when my math scores in the 2nd grade were at the bottom of the class and doing everything she could to help me learn concepts that she herself didn't have the opportunity to learn in school. I thank my father for his creativity and warm words of encouragement for me when I did not always believe in myself.

I would like to thank my dog Yasmin for her unconditional love and support during my many late nights doing science and writing papers in the last year and a half of my PhD program.

Lastly, I would like to thank my partner, Alec, for his unwavering support for past last seven years. As a fellow PhD student, I felt I had a partner who could truly understand the challenges of being a graduate student and together we overcame the ups and downs of academic life. Each day, he inspires me more to fully realize my potential.

Chapter Two is a version of Briscoe, L., Balliu, B., Sankararaman, S., Halperin, E., & Garud, N. R. (2022). Evaluating supervised and unsupervised background noise correction in human gut microbiome data. *PLOS Computational Biology*, *18*(2), 1–25. DOI:10.1371/journal.pcbi.1009838. L.B., B.B., S.S., E.H., N.R.G. conceived and designed the experiments. L.B. and B.B. performed the experiments. L.B., B.B. : Analyzed the data. LB Coded the pipeline: E.H., N.R.G. ccontributed analysis tools: L.B., B.B., S.S., E.H., N.R.G wrote the manuscript.

Chapter Three is a version of Briscoe, L., Halperin, E., & Garud, N. R. (2023). SNV-FEAST: microbial source tracking with single nucleotide variants. *Genome Biology 2023 24:1*, *24*(1), 1–23. DOI:10.1186/S13059-023-02927-8. LB, EH, and NRG conceived of the study. LB implemented the method and experiments. LB and NRG designed and directed experiments and wrote the manuscript.

Chapter Four is in preparation for publication. LB and Michael Wasney (WS) implemented the experiments and analyses. WS, LB and NRG conceived of the study and wrote the manuscript.

This work was supported in part by the NSF Graduate Research Fellowship Program grant numbers DGE-1650604 and DGE-2034835 as well as NSF grant 1705197 and NIH grant NHGRI 5R01HG010505. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Vita

Education

2012 - 2016 BS, Molecular, Cell and Developmental Biology
University of California Los Angeles, CA, USA

Selected Publications

Briscoe, L., Halperin, E., & Garud, N. R. (2023). SNV-FEAST: microbial source tracking with single nucleotide variants. *Genome Biology 2023 24:1*, *24*(1), 1–23. https://doi.org/10.1186/S13059-023-02927-8

Briscoe, L., Balliu, B., Sankararaman, S., Halperin, E., & Garud, N. R. (2022). Evaluating supervised and unsupervised background noise correction in human gut microbiome data. *PLOS Computational Biology*, *18*(2), 1–25. https://doi.org/10.1371/journal.pcbi.1009838

Singh, G., Haileselassie, Y., **Briscoe, L.**, Bai, L., Patel, A., Sanjines, E., Hendler, S., Singh, P. K., Garud, N. R., Limketkai, B. N., & Habtezion, A. (2022). The effect of gastric acid suppression on probiotic colonization in a double blinded randomized clinical trial. *Clinical Nutrition ESPEN*, *47*, 70–77. https://doi.org/https://doi.org/10.1016/j.clnesp.2021.11.005

Mandric, I., Schwarz, T., Majumdar, A., Hou, K., **Briscoe, L.,** Perez, R., Subramaniam, M., Hafemeister, C., Satija, R., Ye, C. J., Pasaniuc, B., & Halperin, E. (2020). Optimized design of single-cell RNA sequencing experiments for cell-type-specific eQTL analysis. *Nature Communications*, *11*(1), 14. https://doi.org/10.1038/s41467-020-19365-w

Shenhav, L., Furman, O., **Briscoe, L.**, Thompson, M., Silverman, J. D., Mizrahi, I., & Halperin, E. (2019). Modeling the temporal dynamics of the gut microbial community in adults and infants. *PLoS Computational Biology*, *15*(6). https://doi.org/10.1371/journal.pcbi.1006960

Shenhav, L., Thompson, M., Joseph, T. A., **Briscoe, L.**, Furman, O., Bogumil, D., Mizrahi, I., Pe'er, I., & Halperin, E. (2019). FEAST: fast expectation-maximization for microbial source tracking. *Nature Methods*, *16*(7). https://doi.org/10.1038/s41592-019-0431-x

Estrada-de los Santos, P., Palmer, M., Chávez-Ramírez, B., Beukes, C., Steenkamp, E. T., **Briscoe, L.**, Khan, N., Maluk, M., Lafos, M., Humm, E., Arrabit, M., Crook, M., Gross, E., Simon, M. F., dos Reis Junior, F. B., Whitman, W. B., Shapiro, N., Poole, P. S., Hirsch, A. M., … James, E. K. (2018). Whole Genome Analyses Suggests that Burkholderia sensu lato Contains Two Additional Novel Genera (Mycetohabitans gen. nov., and Trinickia gen. nov.): Implications for the Evolution of Diazotrophy and Nodulation in the Burkholderiaceae. *Genes 2018, Vol. 9, Page 389*, *9*(8), 389. https://doi.org/10.3390/GENES9080389

Lopez, D., Montoya, D., Ambrose, M., Lam, L., **Briscoe, L.**, Adams, C., Modlin, R. L., & Pellegrini, M. (2017). SaVanT: A web-based tool for the sample-level visualization of molecular signatures in gene expression profiles. *BMC Genomics*, *18*(1). https://doi.org/10.1186/s12864-017-4167-7

De Meyer, S. E., **Briscoe, L.**, Martínez-Hidalgo, P., Agapakis, C. M., De-Los Santos, P. E., Seshadri, R., Reeve, W., Weinstock, G., O'Hara, G., Howieson, J. G., & Hirsch, A. M. (2016). Symbiotic burkholderia species show diverse arrangements of nif/fix and nod genes and lack typical High-Affinity cytochrome cbb3 Oxidase genes. *Molecular Plant-Microbe Interactions*, *29*(8). https://doi.org/10.1094/MPMI-05-16-0091-R

CHAPTER 1: Introduction

Scope of Research

The human gut microbiome is associated with a number of host phenotypes including colorectal cancer[1], obesity[2,3], and antibiotic consumption[4,5], among other traits[6]. Despite the promise of leveraging the microbiome as a diagnostic of disease, significant challenges still remain in accurately predicting human phenotypes and consequently identifying underlying causal mechanisms of disease. Among these challenges are that biological and technical covariates can confound the ability to detect associations between the microbiome and human phenotypes[7].

It is well known that major components of microbiome variability can often be attributed to technical or biological factors. Some of these factors introduce unwanted, systematic variability in the data that is unrelated to the biological variable of interest, e.g. body mass index and colorectal cancer status. Technical factors include differences in preservation [8], storage[9,10], the kit[11], lysis[12,13], extraction of DNA[13], primer[14], and several others [15–18]. Biological, or host-related factors can include diet, sex, age and medication use[7]. When these factors are correlated to the phenotype of interest, they can act as confounders of the phenotypic effects and correcting for such factors is crucial to improve the prediction accuracy of phenotypes.

There have been several efforts to address confounders in other domains including gene expression[19,20] and methylation[21,22]. Existing approaches to covariate correction are often inappropriate for microbiome data because features are often sparse[23,24], non-independent[25], and non-Gaussian[26]. Additionally, applying statistical methods that are not appropriate for microbiome data can yield spurious results that are not reproducible in follow-up studies. This calls for the development of microbiome-specific methods to correct for confounders.

Moreover, discerning the major contributors of a person's microbiome may reveal a large environmental influence that cannot be explained by health alone. Elucidating the sources of a microbiome can provide insight into the ecological dynamics responsible for the formation of these communities, and further understanding host health.

The gut microbiome is a dynamic ecosystem that changes with time and also along the tract of the gut. It is important that scientists understand how to adequately sample the gut to understand how the gut microbiome is actively changing in response to our health and our environment.

Contributions and Overview

In this dissertation, we propose quantitative and computational approaches to analyze metagenomic data from the human gut microbiome. Key advantage in our work is leveraging the high-specificity of metagenomic data that is missed in most studies that focus on amplicon sequencing data such as 16S.

In chapter 2, we comprehensively evaluate supervised and unsupervised approaches to remove background noise from microbiome data that is due to technical variables unrelated to the phenotype of interest, such as sequencing protocol, and thereby improving our ability to find accurate biomarkers of human disease. We perform our evaluation on four broad categories of datatypes: 16S taxa abundance, k-mers from 16S reads, metagenomic taxa abundance, k-mers from metagenomic data. Using a series of benchmarks, we show that the combination of certain data transformations and correction procedures can maintain and in some cases improve

phenotype prediction accuracy and reduce false positive associations. We demonstrate this with prediction of colorectal cancer using one 16S [cite] and one metagenomic dataset [cite], prediction of BMI using the Hispanic Community Health Cohort [cite], and lastly prediction of antibiotic history using the American Gut Project [cite]. Using another benchmark, we demonstrate that correction suppresses false positive associations when performing biomarking discovery for colorectal cancer.

In chapter 3, we propose a novel method to find single nucleotide variants for source tracking using metagenomic data. Previously, source tracking has been primarily done using species abundance and not single nucleotide variants, which may be more informative because of their high specificity to certain sources. Utilizing all SNVs for all sinks and sources of interest would exact a heavy computational burden, for that reason we design a signature SNV scoring method to produce features for input into a previously designed source tracking algorithm, FEAST. We apply our signature SNV method combined with FEAST to simulated infant microbiomes produced from mixtures of maternal microbiome data and find that our method accurately estimates mixture proportions compared to FEAST applied to species abundance. We then apply the approach to three case studies, infants over the first year of life, infants in the NICU, and ocean microbiome samples.

In Chapter 4, we study a novel dataset of humanized mice to understand how species, strain, and gene diversity change along the gut. We assess these levels of diversity using metagenomic data sampled along the tract of the gut from 6 genetically identical mice that are gavaged with a human decal sample and then raised on either a standard rodent diet or fiber-rich diet. Previously, most studies of diversity have used stool data. We show that diversity as represented in the stool may not represent the diversity in the gut. We also find that species

composition differs substantially between the upper and lower gut as well as between diet regimes. By contrast, strain composition is more uniform between segments of the gut. In comparing different hosts, we frequently find different sets of strains despite hosts being provided with the same inoculum of strains, suggesting that colonization is a stochastic process. Further, hosts that are cohoused are more likely to have the same strain, illustrating how shared environments can constrain strain diversity across individuals. Strains tend to be at more similar (although still variable) abundances within the same host, and less similar frequencies between hosts. Even when a mouse harbors only a single strain, gene content can differ predictably along the tract of the gut. In sum, we show that diversity in the gut microbiome is shaped by tissue, diet, and shared environments.

CHAPTER 2: Evaluating supervised and unsupervised background noise correction in human gut microbiome data

Evaluating supervised and unsupervised background noise correction in human gut microbiome data

Leah Briscoe[1]*, Brunilda Balliu[2], Sriram Sankararaman[3,4,5], Eran Halperin[3,4,5,6,7]*, and Nandita R. Garud[4,8]*

1. Bioinformatics Interdepartmental Program, University of California Los Angeles, Los Angeles, Los Angeles, CA, United States of America
2. Department of Computational Medicine, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, United States of America
3. Department of Computer Science, University of California Los Angeles, Los Angeles, CA, United States of America
4. Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, United States of America
5. Department of Computational Medicine, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, United States of America
6. Department of Anesthesiology and Perioperative Medicine, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, United States of America
7. Institute of Precision Health, University of California Los Angeles, CA, United States of America
8. Department of Ecology and Evolutionary Biology, University of California Los Angeles, CA, United States of America

*Co-corresponding authors. Correspondence should be addressed to:
leah.briscoe@ucla.edu; ehalperin@cs.ucla.edu; ngarud@ucla.edu

Abstract

The ability to predict human phenotypes and identify biomarkers of disease from metagenomic data is crucial for the development of therapeutics for microbiome-associated diseases. However, metagenomic data is commonly affected by technical variables unrelated to the phenotype of

interest, such as sequencing protocol, which can make it difficult to predict phenotype and find biomarkers of disease. Supervised methods to correct for background noise, originally designed for gene expression and RNA-seq data, are commonly applied to microbiome data but may be limited because they cannot account for unmeasured sources of variation. Unsupervised approaches address this issue, but current methods are limited because they are ill-equipped to deal with the unique aspects of microbiome data, which is compositional, highly skewed, and sparse. We perform a comparative analysis of the ability of different denoising transformations in combination with supervised correction methods as well as an unsupervised principal component correction approach that is presently used in other domains but has not been applied to microbiome data to date. We find that the unsupervised principal component correction approach has comparable ability in reducing false discovery of biomarkers as the supervised approaches, with the added benefit of not needing to know the sources of variation apriori. However, in prediction tasks, it appears to only improve prediction when technical variables contribute to the majority of variance in the data. As new and larger metagenomic datasets become increasingly available, background noise correction will become essential for generating reproducible microbiome analyses.

1. Keywords: **Batch correction; microbiome; metagenomics**

Author Summary

The human gut microbiome is known to play a major role in health and is associated with many diseases including colorectal cancer, obesity, and diabetes. The prediction of host phenotypes and identification of biomarkers of disease is essential for harnessing the therapeutic potential of the microbiome. However, many metagenomic datasets are affected by technical

variables that introduce unwanted variation that can confound the ability to predict phenotypes and identify biomarkers. Currently, supervised methods originally designed for gene expression and RNA-seq data are commonly applied to microbiome data for correction of background noise, but they are limited in that they cannot correct for unmeasured sources of variation. Unsupervised approaches address this issue, but current methods are limited because they are ill-equipped to deal with the unique aspects of microbiome data, which is compositional, highly skewed, and sparse. We perform a comparative analysis of the ability of different denoising transformations in combination with supervised correction methods as well as an unsupervised principal component correction approach and find that all correction approaches reduce false positives for biomarker discovery. In the task of predicting phenotypes, different approaches have varying success where the unsupervised correction can improve prediction when technical variables contribute to the majority of variance in the data. As new and larger metagenomic datasets become increasingly available, background noise correction will become essential for generating reproducible microbiome analyses.

Introduction

The human gut microbiome is associated with a number of host phenotypes including colorectal cancer [1], obesity [2,3], and antibiotic consumption [4,5,27,28], among other traits [6,29]. Despite the promise of the microbiome as a diagnostic tool, significant challenges remain for predicting phenotypes and finding reproducible biomarkers of human phenotypes from microbiome data. One major challenge is that technical covariates, including sample storage [9], cell lysis protocol [14,15], extraction method [13,30], DNA preservation and storage protocol [8], preparation kit [11,31], and

primer choice [14], are known to introduce unwanted variation and systematically bias the relative abundances of taxonomic features in microbiome samples [15–18,32–35].

These covariates, when differentially distributed across phenotypes, can act as confounders. There are two potential outcomes of confounding in prediction accuracy: increased accuracy when confounders are consistently correlated with the phenotype, or decreased prediction accuracy when the confounder is oppositely correlated with phenotype from one subset of the data to another. In either scenario, confounding is problematic for detecting true associations between the microbiome and phenotype. The pooling of datasets is a major contributor of confounding yet combining datasets is an increasingly common [1,36–39] and powerful means to validate associations [6,40] in a discovery dataset with held out datasets [1,41,42]. Recent studies have shown that confounding covariates are widespread in genomic datasets. Gibbons *et al.* [43] found that combining datasets to detect members of the microbiome that are associated with colorectal cancer resulted in false positive detection of differentially abundant taxa. Confounding covariates were also pervasive [7] in one of the largest metagenomic datasets available, the American Gut Project (AGP) [44].

Despite the widespread effects of background noise in microbiome data, there is currently a dearth of methods specially equipped for removing unwanted variation in microbiome data. Initial steps in processing microbiome data often involve addressing differences in library sizes across samples by applying the variance-stabilizing transformation (VST) from DESeq2 [45] or the $\log_2$-counts per million (logCPM) from EdgeR [46] on taxonomic counts data [47–52]. However these transformations do not sufficiently address other contributors of unwanted variance such as study-specific covariates, which neccessatates explicit methods for correction. Existing methods repurposed from other domains for this purpose, including gene expression[39,40] and methylation

[53–55], generally fall into two categories: supervised methods, where the sources of variation must be explicitly specified, and unsupervised methods, where the sources of variation are first inferred and then removed before association or prediction analyses. The most popular supervised methods are batch mean centering (BMC)[43], which centers data batch by batch, and ComBat[44] and limma[45], which both use empirical Bayes. Many studies will apply a supervised method after applying one of the above transformations in microbiome data . However, since many sources of variation may be unknown, and moreover, the extent of variation they introduce may vary from dataset to dataset [17,43,56–58], unsupervised approaches [59–61] for covariate correction may be more effective in removing background noise. Among the unsupervised approaches are ReFactor [61], Surrogate Variable Analysis (SVA) [59], and Remove Unwanted Variation (RUV) [60] which were designed for methylation or gene expression data. These methods quantify "surrogate variables" that represent study-specific effects and regress them out of the data.

Despite their promise, the repurposed supervised and unsupervised approaches [59–61] are not suitable for microbiome data because most of them rely on assumptions that the data is normally distributed. However, taxonomic features are often sparse [23,24] due to taxa having abundances below the detection limit of sequencing [23], or taxa being absent in certain samples, resulting in skewed non-normal distributions. Additionally, because the microbiome data is usually transformed into measures of relative abundances, the data is compositional, or in other words, represented as relative frequencies of taxonomic features within a sample that sum to one. This representation also causes non-normal distributions.

Supervised methods proposed explicitly for microbiome data to reduce background noise include percentile normalization [38], Partial Least Squares Discriminant Analysis [62], and multiplicative bias correction [33]. Both percentile normalization [38] and Partial Least Squares

Discriminant Analysis [62] aim to find predictive features in fully labeled data with known batches and known phenotypes, and are not designed for prediction of phenotypes in unlabeled data, while multiplicative bias correction [33] requires either a reference sample in which the species abundance distribution is known or a term specifying the experiment label, and thus cannot account for multiple sources of background noise simultaneously. Given that these methods are supervised and thus cannot be applied to unlabeled data, there still remains a need in the microbiome field for unsupervised approaches that can adjust for both measured and unmeasured variables. Additionally, there is little published research comparing adapted approaches head-to-head in microbiome data.

To address the need for unsupervised approaches applicable to microbiome data, we examined a popular approach used in the field of population genetics known as Principal Components Analysis (PCA) correction. Population structure is often strongly reflected in the first principal components (PCs) calculated from genotype data [63–65]. By removing the effect of the first few PCs in a regression approach, association testing can be done to find potential genetic biomarkers of phenotype rather than biomarkers of population structure [63–65]. PCA correction has been effective in correcting for confounding covariates in human genetic data [63,65] and morphological data[66], but to date has not been applied to microbiome data. Yet, we and others find that top principal components in multiple datasets are correlated with numerous confounding variables like host genetics [67], ethnicity of the host [68], and also abiotic factors like temperature [69], suggesting that PCA correction may be an effective unsupervised correction approach.

In this paper, we evaluated the ability of PCA correction to remove background noise in microbiome data and compared its performance to supervised background noise correction

approaches [70–72] that are commonly used for microbiome data. Specifically, we tested the impact of regressing out principal components (PCs) from microbiome data featurized as abundances of taxonomic features or *k*-mers. Abundance of taxonomic units are determined by aligning or binning reads based on reference genomes, whereas *k*-mer abundances are calculated by counting appearances of short substrings of length *k* in raw sequences. While taxonomic features have immediate biological interpretability, the use of *k*-mers is beneficial because they do not rely on a reference genome. Additionally, we assess the impact of applying a variance stabilizing transformation (VST) or logCPM (log counts per million), and compare this to application of the centered log ratio (CLR). CLR is more widely used for compositional data, particularly in microbiome contexts [25,40,73–77], and is a suggested transformation prior to factor analysis such as PCA because it breaks the dependence between features [25] and makes data more normally distributed [62]. This transformation can make the PCs more interpretable because the transformed value is the abundance relative to the mean value for a sample.

By performing a comparative analysis of PCA correction and existing supervised correction approaches, we evaluate the merits of repurposing the PCA correction approach from the field of population genetics to the microbiome, as well as assess the strengths and limitations of various methods. Throughout this study, we highlight important considerations for phenotype association studies from large cohort and cross-study metagenomic analyses, which we hope paves the way for higher reproducibility across microbiome studies.

Results

We analyzed four metagenomic datasets for evidence of technical covariates that could introduce noise or confounding that, as a result, interfere with biomarker discovery and prediction accuracy. We evaluated the ability of three popular supervised approaches for microbiome data (ComBat [72], limma [71], and batch mean centering (BMC) [70]), three transformations (CLR, VST from DESeq2 [45] and logCPM from EdgeR [46]), and an unsupervised approach, PCA correction, to correct for noise and confounding. We focused on three phenotypes of interest: body mass index (BMI), colorectal cancer (CRC), and antibiotic consumption (**Table 1**). The datasets we analyzed included: (i) the American Gut Project [44] (AGP), which has known confounding variables [7], (ii) a pooled dataset composed of three 16S datasets with healthy and CRC individuals (hereafter referred to as 'CRC-16S') [38], (iii) a pooled dataset composed of seven whole metagenome sequenced datasets (WGS) with healthy and CRC individuals (hereafter referred to as 'CRC-WGS') [1,78], and (iv) the Hispanic Community Health Study (HCHS) [79] consisting of 16S samples from over one thousand individuals from several Hispanic countries. These datasets allowed us to assess noise and confounding both within datasets (AGP and HCHS) and across pooled datasets (CRC-16S and CRC-WGS).

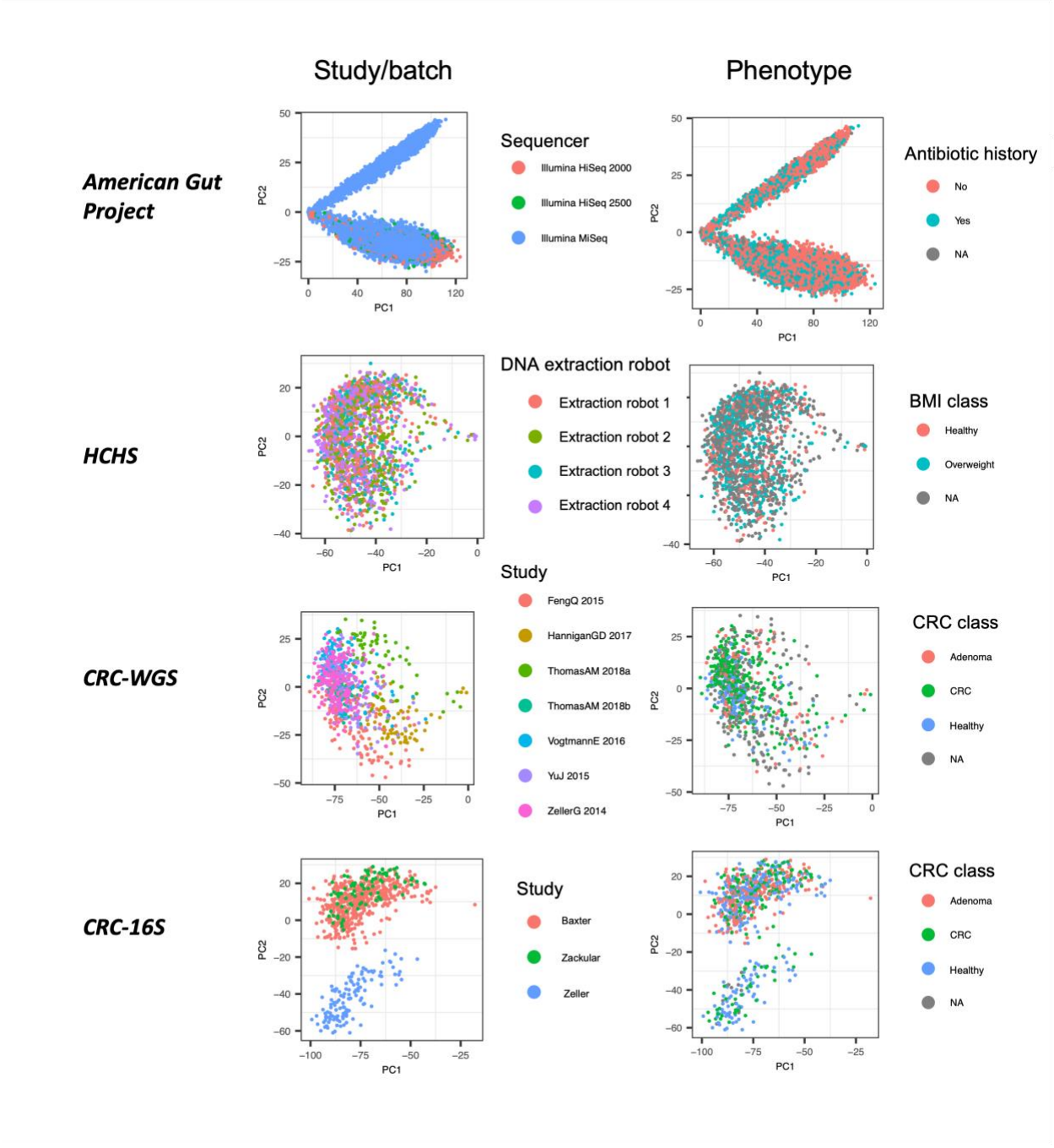
Phenotype	Joined dataset	Number of samples	Number of studies	Sequencing method	Published Sources
Body mass index	American Gut Project (AGP)	6,722	1 (multiple sequencing batches)	16S	McDonald et al.
Antibiotic history	American Gut Project (AGP)	12,619	1 (multiple sequencing batches)	16S	McDonald et al.
Body mass index	Hispanic Community Health Study (HCHS)	1,769	1 (multiple sequencing batches)	16S	Kaplan et al.
Colorectal Cancer	CRC-16S	574	3	16S	Baxter et al. Zeller et al. Zackular et al.

Colorectal Cancer	CRC-WGS	813	7	WGS	Feng et al. Yu et al. Vogtman et al. Hannigan et al. Thomas et al. Zeller et al.

T 1 **Table 1. Datasets used in this study**. Two pooled datasets composed of multiple studies are abbreviated as CRC-16S [80–82] and CRC-WGS [1,81,83–86], whereas the American Gut Project (AGP) [44] and the Hispanic Community Health Study (HCHS) [79] are each from a single source study and have several potential confounders [7].

Background noise detected by principal component analysis

To assess the extent of microbiome variation attributable to technical covariates, we performed PCA on CLR-transformed (see Methods) taxonomic abundance profiles and short *k*-mers (between sizes 5 and 8) derived from the raw metagenomic reads (see Methods). In most cases, for the first two PCs, samples cluster by dataset and not the primary phenotype of interest (**Fig 1** and **Fig S1**), consistent with previous findings [13] that technical factors have a strong effect on the microbiome.



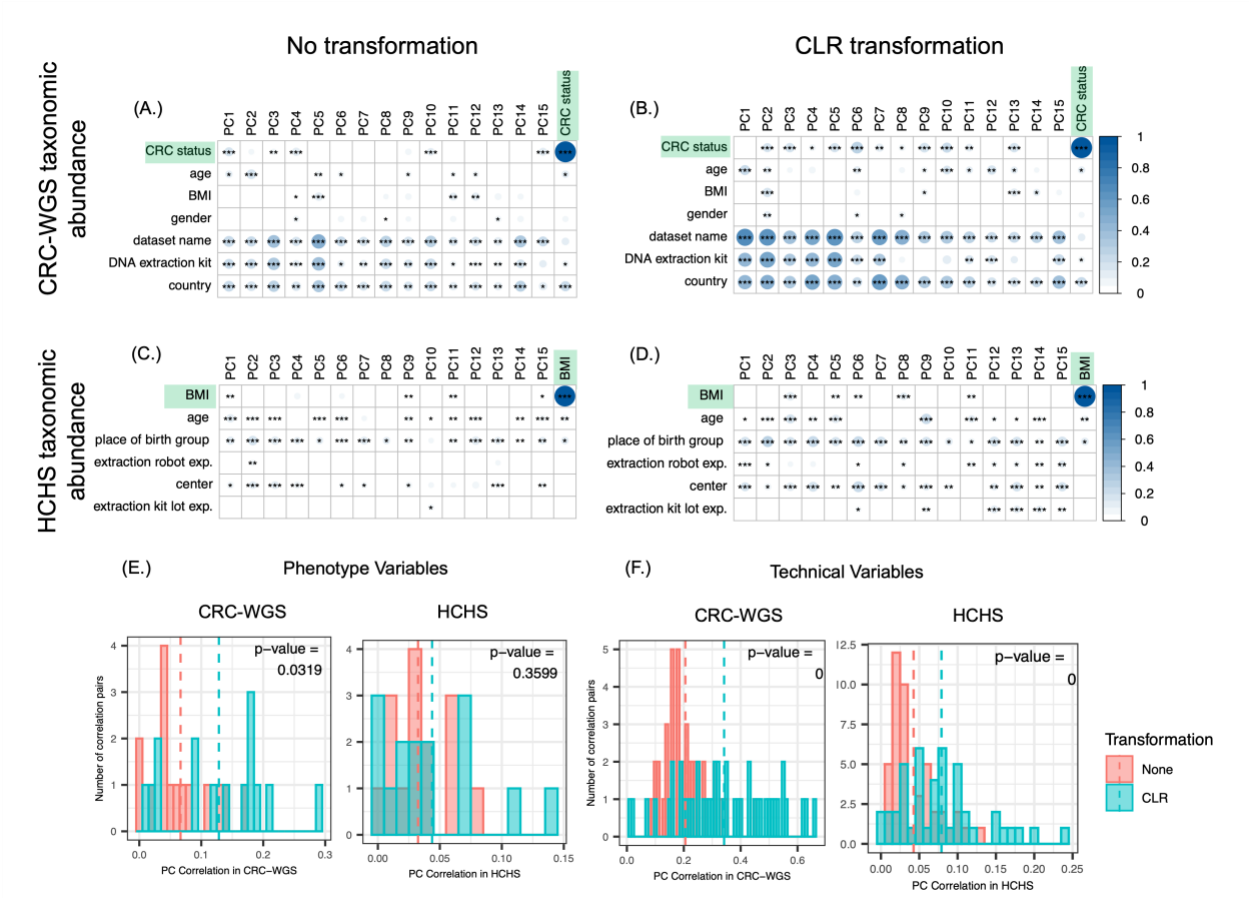
1 **Figure 1. First two principal components of across datasets.** PCA applied to CLR-transformed taxonomic abundance data from the four datasets of the study. Each point represents a single microbiome sample colored by either study or batch and by phenotype group.

More generally, the top 15 PCs in each dataset are more correlated with technical variables than the phenotypes of interest (**Fig 2A and Fig S2**). For example, in the CRC-WGS

dataset, PCs one through five on average have a 0.28 mean correlation with dataset label but only 0.072 mean correlation with CRC status (**Fig 2A**). It is worth noting that these first five PCs collectively explain 84% of the variance in the CRC-WGS data and that the strongest correlations with CRC status are in the first five PCs. In the HCHS dataset, the top 5 PCs have significant correlation with demographic information such as place of birth (0.13 mean correlation of top 5 PCs) and sequencing center (0.09 mean correlation of top 5 PCs), but only a mean 0.04 correlation with BMI. In this dataset, the first five PCs collectively explain 59% of the variance in the HCHS dataset but only the first PC is significantly correlated with BMI, where PC1 explains 24 % of variance (**Fig 2D**).

We next assessed the impact of CLR-transformation on the correlation of top PCs with technical and biological covariates, and compare the correlations using a two-sample Wilcoxon signed-rank test. Firstly, across all datasets, CLR-transformation of taxonomic abundance and *k*-mer data results in more normally distributed data (**Fig S3**), making the data more suitable for PCA. However, the change in correlation of the top PCs with technical and biological covariates after application of the CLR transformation varies from dataset to dataset. In the case of both AGP and CRC-WGS datasets, the CLR transformation results in siginificantly increased correlation of the top PCs with both biological and technical covariates (**Fig 2A and 2B, Fig S4**) (in the CRC-WGS dataset, median correlation of PCs with CRC increased from 0.05 to 0.14 with Wilcoxon signed-rank p-value = 0.03 and median correlation with technical covariates increased from 0.19 to 0.32 with Wilcoxon signed-rank p-value $< 2.22 \times 10^{-3}$; in the AGP dataset, median correlation of PCs with BMI and antibiotic history increased from 0.16 to 0.31 with Wilcoxon signed-rank p-value = 1×10^{-4} and median correlation with technical covariates increased from 0.05 to 0.07 with Wilcoxon signed-rank p-value = 8.7×10^{-3}) (**Fig 2B and 2C**). In the CRC-16S

dataset, neither biological or technical variates showed significantly increased correlation after CLR transformation variables (CRC median correlation increased from 0.05 to 0.10 with Wilcoxon signed-rank p-value 0.084; technical covariate median correlation changed from 0.09 to 0.08 with Wilcoxon signed-rank p-value 0.12). Unlike all the other datasets, application of the CLR transformation to the taxonomic abundances of the HCHS dataset results in a significantly increased correlation with technical variables, but not biological variables (BMI median correlation increased from 0.029 to 0.033 with Wilcoxon signed-rank p-value = 0.36; technical covariate mean correlation increased from 0.03 to 0.07 with Wilcoxon signed-rank p-value < 2.22×10^{-3}) (**Fig 2E and 2F)**. These correlations are all the more striking given the high percentage of variance explained by the first five PCs alone: 80% of variance in the CRC-WGS dataset, 64% of variance in the AGP dataset, and 65% of variance in the HCHS dataset. We similarly assessed the impact of logCPM and VST transformations on the correaltions of the top 15 PCs with technical and biological variables in **Fig S6** and found that correlations with study covariates also increase.



2 **Figure 2. Microbiome data is affected by technical and biological variables.** (A-D) Heatmaps of canonical correlations between the first 15 PCs and study covariates in CRC-WGS with (A) no transformation and (B) after CLR transformation; and in HCHS with (C) no transformation and (D) after CLR transformation. **(E,F)** Histograms of the correlations in (A-D) where the distributions were compared using a paired Wilcoxon signed-rank test to test whether the distribution of correlations from PCs of CLR-transformed are greater than the untransformed. The size and color of the circles in each cell in A-D indicate the magnitude of correlation and black asterisks indicate the significance of the Pearson correlation of the PCs with each of the variables. The color bar at right represents the range of correlations observed across all datasets. [*,**,*** indicate Wilcoxon signed-rank *p*-values as follows: $10^{-2} < p < 0.05$, $10^{-3} < p < 10^{-2}$, $p < 10^{-3}$]. See **Figs S2** and **S5** for similar analyses for the other datasets, and **Fig S6** for other transformations.

We also assessed the impact of *k*-merization on the correlation of variables with top PCs. Unlike for taxonomic abundances, CRC-WGS does not show significant change as a result of CLR transformation on *k*-mers (**Fig S5**). In the AGP dataset, median correlations with BMI and antibiotic history increase from 0.55 to 0.57 (Wilcoxon signed-rank p-value = 8×10^{-4}), and in the HCHS and CRC-16S datasets, correlations with technical variables increase from a median

of 0.04 to 0.07 (Wilcoxon signed-rank p-value = 0.001) and a median of 0.1117 to 0.1125 (Wilcoxon signed-rank p-value = 0.0498) after the CLR transformation. Through these analyses on taxonomic abundance and *k*-mers, we show that technical variables introduce considerable variation in microbiome data sets, that this variation is often larger than variation explained by phenotypes of interest. Transformations like CLR can additionally make this variation explained by technical variables more apparent.

Reduction of false positive biomarker discovery as a metric of background noise correction

Pooling of datasets is frequently done to augment power to detect associations with or make predictions about host phenotype [1,36,37,39,78]. However, this practice can also result in false positive associations due to confounding between study-specific variables and phenotype [38]. Thus, we tested the ability of different background noise correction methods to reduce false positive biomarker discoveries. To do so, we performed a titration experiment similar to that described in Gibbons et al. [43] in which control groups from two different studies in the CRC-WGS dataset were mixed at different proportions to create a new control group of equal size that was then compared with cases to identify taxa significantly associated with disease using a Wilcoxon rank sum test with false discovery rate correction (q-value < 5%). Without correction, spurious associations are expected to increase with increasing proporion of control samples coming from a different study (**Fig 3**). We compare correction approaches by ascertaining the number of likely false positive associations at different titration levels (proportions of control samples from another study) ranging from 0% to 100%. In the scenario where 100% of controls are from a second study, the study variable is a complete confounder for case-control status.

To assess the efficacy of transformations to reduce false positive associations, we first compared the untransformed and uncorrected relative abundance data to each of three data

denoising transformations: logCPM, VST and CLR applied to feature counts. As expected, when the data is untransformed, the number of new taxa identified that are likely false positives steadily increases as the number of control samples added from a second study increases, reaching 42 when 100% of controls are from the second study. When the data is transformed with logCPM, VST, or CLR, the number of likely false positives reaches 20, 52, and 44, respectively (**Fig 3A** and **3B, Fig S7** and **Table S1**), indication that transformation alone does not always reduce false positives.

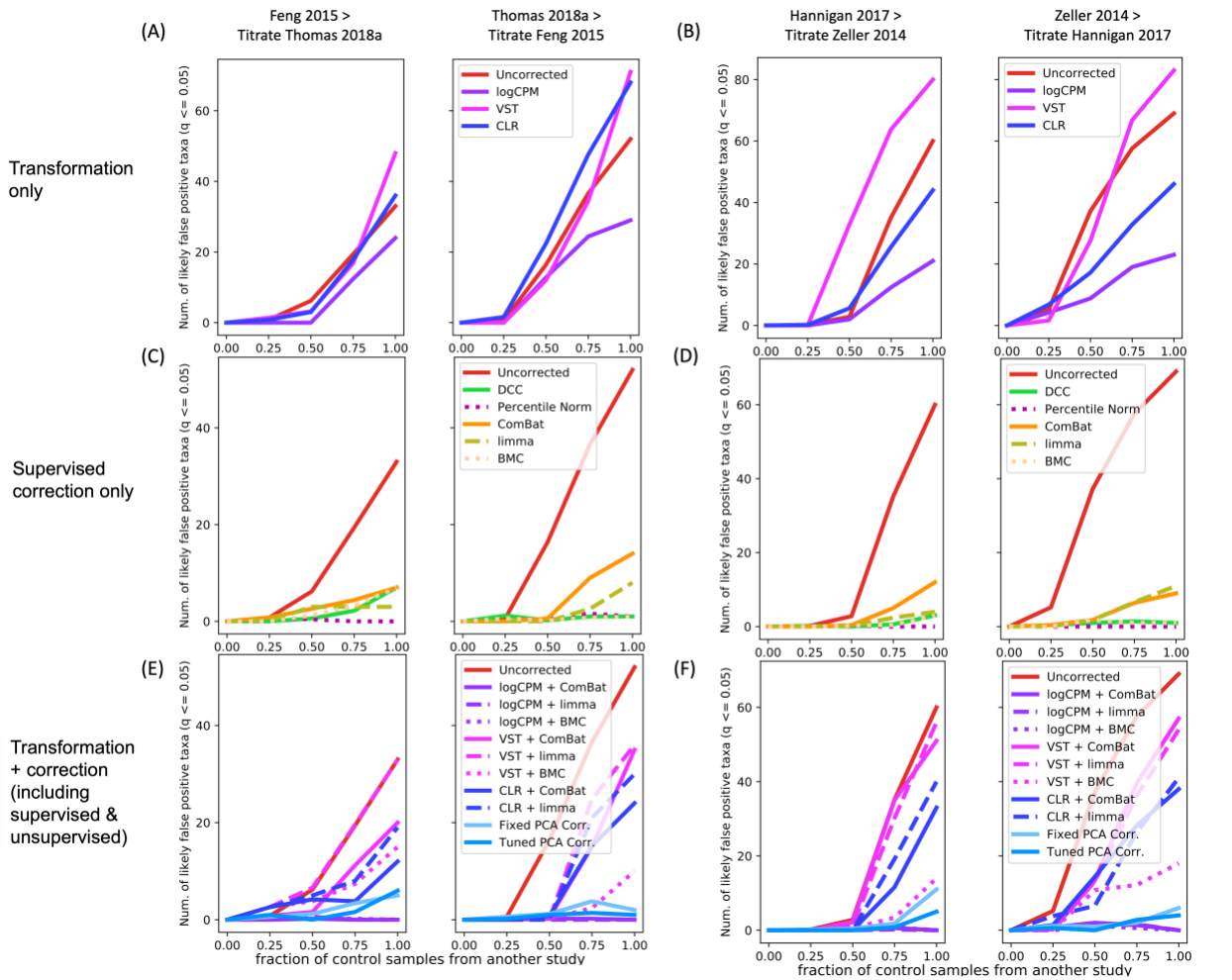
Next, we assessed the ability of supervised background noise correction methods to suppress false positives. These methods included percentile normalization [38], BMC [70], ComBat [72], and limma [71] which require a batch variable to be specified. Thus, in these cases we corrected for the variables that are the most correlated with the top PCs in each dataset: sequencing instrument in the AGP dataset, processing robot in the HCHS dataset, and source study in the CRC dataset. We additionally included a supervised correction approach in which these same primary contributors of heterogeneity were directly regressed out, an approach we term in this paper as Direct Covariate Correction (DCC) (see Methods). When 100% of controls are from the second study, the number of likely false positives drops to 5, 0, 5, and 6 respectively for the DCC, percentile normalization, ComBat, and BMC methods (**Fig 3C** and **3D, Fig S7** and **Table S1**).

Next, we evaluated the effectiveness of applying the logCPM, VST, and CLR tranformations in combination with the supervised approaches ComBat, limma, and BMC (**Fig 3E and F**), a practice which is currently done in the literature for microbiome studies [47–52]. We also compared these approaches to two variants of unsupervised correction in which PCA correction is applied after CLR: one in which the optimal number of top PCs are identified via

cross-validation and regressed out from the data and another in which data is corrected for a fixed and arbitrary number of PCs. We refer to these two variants as tuned PCA and fixed PCA, respectively (see Methods). Tuned PCA uses a validation set to determine the optimal number of PCs that maximize prediction accuracy while fixed PCA correction corrects for the first three PCs (Methods). The choice of three PCs for this analysis was arbitrarily selected to avoid completely throwing away the signal associated with the phenotype of interest.

When 100% of controls are from the second study, logCPM applied prior to ComBat, limma, or BMC results in 1, 2, and 2 likely false positive associations, respectively (**Fig 3D** and **3E, Fig S7** and **Table S1**). When the VST transformation is applied prior to ComBat, limma, or BMC, we find 45, 55, and 25 likely false positive associations (**Fig 3D** and **3E, Fig S7** and **Table S1**). When the CLR transformation is applied prior to ComBat, limma, or BMC, we find 26, 35, and 173 likely false positive associations (**Fig 3D** and **3E, Fig S7** and **Table S1**). Lastly, when Fixed PCA and Tuned PCA is applied along with CLR, we find 14 and 11 likely false positive associations, respectively.

Overall, these results suggest that data transformations should not be applied alone and that a transformation like logCPM can be applied before applying a supervised correction in order to reduce the appearance of false positive associations. Alternatively, unsupervised approaches where CLR is applied prior to PCA correction can also reduce false positive associations.



3 **Figure 3. Spurious association of taxa with case-control status without appropriate correction.** (A) We tested the number of associations identified after replacing the controls from the CRC-WGS study sequenced by [1] referred to as Thomas et al. 2018a with controls from Feng et al. at increasing proportions and vice versa. (B) Similarly, controls in the CRC-WGS study Hannigan et al. [86] were replaced with controls from Zeller et al.[81] and vice versa (**S7 Fig**). BMC + CLR was an outlier and excluded for clarity of visualization, but the summary of mean associations of BMC + CLR is in **Table S1**.

Cross-study prediction after background noise correction

A successful predictive model is transferable across datasets. To assess the impact of background noise correction on phenotype prediction, we performed a leave-one-dataset-out (LODO) analysis. For this analysis, we utilized a nested cross-validation scheme where one dataset was set aside for testing of a prediction model that was trained and validated on the

remaining datasets using either a Random Forest classifier or linear regression model (see Methods). We evaluated the impact of supervised and unsupervised background noise correction approaches, with and without data transformations, on prediction of host phenotype using taxonomic abundance profiles and *k*-mers (see Methods), where binary phenotype prediction accuracy is assessed by Area Under the Curve (AUC) and continuous phenotype prediction accuracy is assessed by Pearson correlation.

We first compared the effect of the different transformation and corrections on prediction of BMI, a continuous phenotype. When applying a transformation only to taxonomic abundances, logCPM and CLR resulted in significantly better Pearson correlations between the true and predicted BMI (0.04 under uncorrected increased to 0.14 and 0.13 median Pearson across batches with one-sided Wilcoxon rank-sum p-value = 0.014 for both), but VST did not show any significant improvement (one-sided Wilcoxon rank-sum p-value = 0.443) (**Fig 4A, Fig S8**). When applying supervised correction approaches without transformations to taxonomic abundance data, we found that ComBat and limma significantly improved prediction to 0.13 median Pearson (one-sided Wilcoxon rank-sum p-value = 0.014 for both) while DCC and BMC did not (one-sided Wilcoxon rank-sum p-value = 0.557). Finally, applying a transformation followed by supervised correction, logCPM or CLR followed by ComBat, limma, or BMC resulted in significantly improved prediction (one-sided Wilcoxon rank-sum p-value = 0.014 for all). Applying Fixed or Tuned PCA correction, which includes a CLR transformation prior to regressing on PCs, also significantly improves prediction (one-sided Wilcoxon rank-sum p-value = 0.014 for both). Because DCC is the only method that explicitly adjusts for primary confounders, we also compared Fixed PCA correction directly to DCC and found that Fixed PCA is significantly better than DCC with median Pearson increasing from 0.045 to 0.089 (one-sided

Wilcoxon rank-sum p-value = 0.014) suggesting that unsupervised correction may more broadly correct for noise that interferes with BMI prediction.

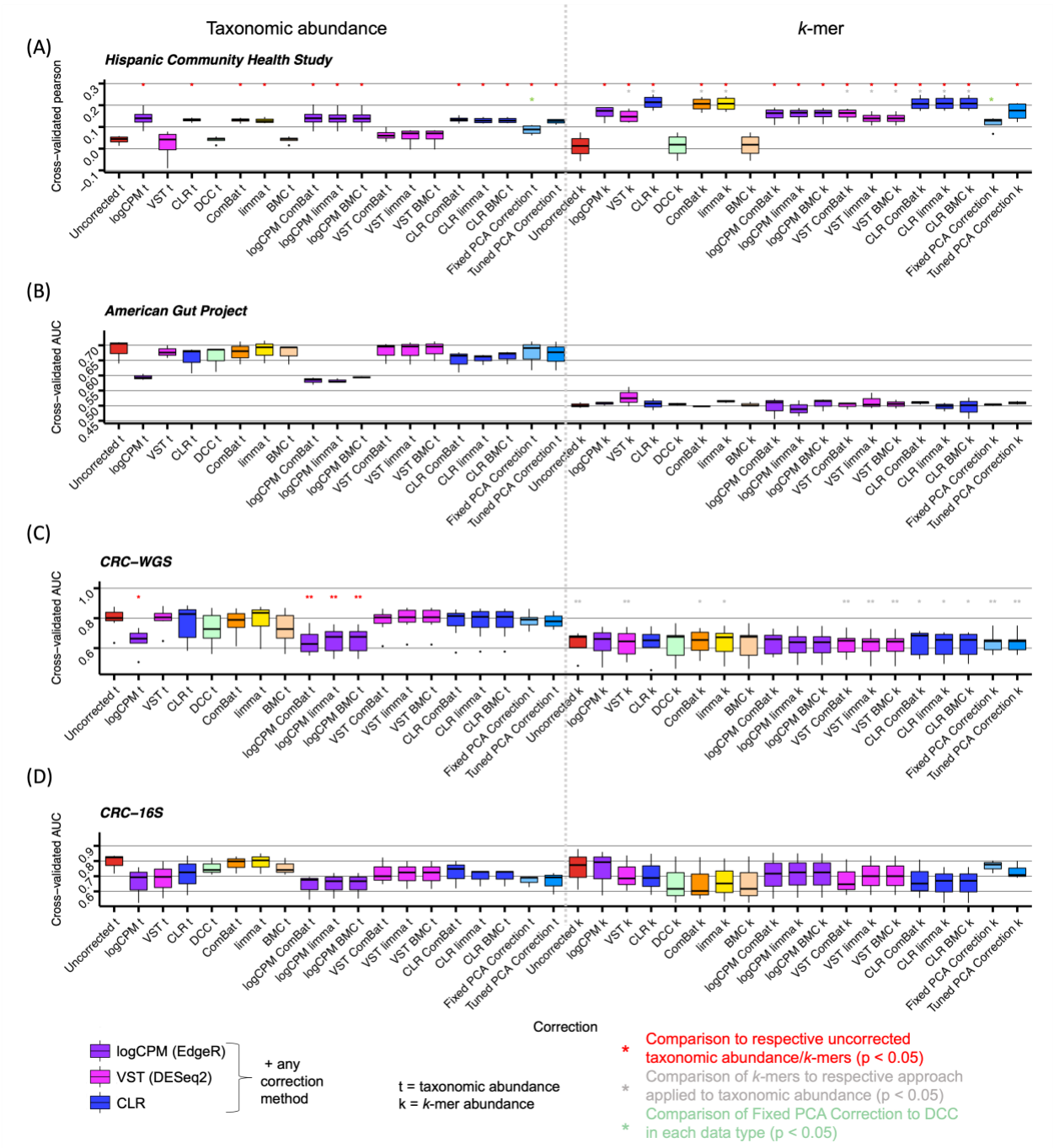
We next assessed the prediction performance using *k*-mers instead of taxonomic abundances. Uncorrected *k*-mer abundances have worse prediction accuracy than taxonomic abundances. However, when *k*-mer abundances are transformed with logCPM, CLR, ComBat or limma alone, or a combination of VST or CLR with a supervised correction, the prediction improves significantly compared to using taxonomic abundance with the highest median Pearson of 0.21 resulting from applying CLR alone (one-sided Wilcoxon rank-sum p-value = 0.014) (**Fig 4A, Fig S8 and S9**). In particular, the use of *k*-mers with a CLR transformation and any correction method, supervised or unsupervised, surpasses prediction accuracy using taxonomic abundance. CLR combined with supervised correction results in a median Pearson correlation of 0.21 and CLR combined with Tuned PCA correction results in a median correlation of 0.17 (one-sided Wilcoxon rank-sum p-value = 0.014 for comparison with uncorrected *k*-mers) (**Fig 4A, Fig S8 and S9**). As with taxonomic abundance, Fixed PCA is significantly better than DCC applied to *k*-mers with median Pearson increasing from 0.018 to 0.13 (one-sided Wilcoxon rank-sum p-value = 0.029).

Next, we evaluated prediction ability with two binary phenotypes: whether an individual had consumed an antibiotic in the previous year and whether an individual has been diagnosed with colorectal cancer (CRC). For the taxonomic abundance profiles of the AGP, CRC-WGS and CRC-16S datasets, applying a data transformation alone did not significantly change the AUC results with the exception of logCPM in the CRC-WGS dataset where accuracy decreased significantly (median AUC went from 0.80 to 0.66, one-sided Wilcoxon rank-sum p-value = 0.0055) (**Fig 4B-D, Fig S8**). Applying any supervised correction method by itself or after a data

transformation did not result in any change in prediction ability, except when logCPM was applied with any supervised correction method to the CRC-WGS dataset, resulting in decreased accuracy (median AUC went from 0.80 in uncorrected to 0.76 for all supervised methods, one-sided Wilcoxon rank-sum p-value = 2×10^{-3}, 3.5×10^{-3}, 2×10^{-3} for ComBat, limma, BMC) (**Fig 4c, Fig S8**).

Unlike for the BMI phenotype, *k*-mers showed significantly lower prediction accuracy than taxonomic abundances irrespective of correction method in the CRC-WGS and AGP datasets (**Fig 4C, S8**). Fixed and Tuned PCA correction on *k*-mers were able to maintain prediction accuracy of uncorrected *k*-mers for all three binary phenotype datasets (**Fig 4B-D, Fig S8**). For the CRC-16S dataset, application of both data transformation and correction methods to *k*-mer abundances resulted in increased accuracy of CRC prediction, but there is insufficient data to find significant increases, with both PCA corrections resulting in the highest accuracy (**Fig 4D)**.

The benefit of utilizing *k*-mers is most apparent in predicting BMI in HCHS, whereas in other datasets, taxonomic abundance data is better. These results indicate that for some phenotypes, correction can improve prediction accuracy, and in most cases accuracy is at least maintained.



4 **Figure 4. Phenotype prediction models generalize across studies after application of noise correction methods.** Cross-study prediction of **(A)** body mass index (BMI) in the HCHS dataset across different extraction robots **(B)** antibiotic consumption in the past year in the AGP dataset across different Illumina sequencing models, **(C)** CRC status in the CRC-WGS dataset across different studies and **(D)** CRC status in the CRC-16S dataset across different studies. The boxplots in (A) indicate leave-one-dataset-out Pearson correlation between true and predicted BMI, for each batch. (B-D) indicate leave-one-dataset-out AUC for each held-out study or batch. p-values comparing each boxplot were computed using a one-sided Wilcoxon signed-rank test. A red * indicates a significant difference in prediction ability compared to uncorrected data in the respective taxonomic or *k*-mer group. A grey * indicates a significant

difference in prediction between the *k*-mer (k) and taxonomic abundance (t) groups for a given approach. A green * indicates a significant difference in prediction between the Fixed PCA correction and DCC for a given data type. Due to the low number of folds in LODO prediction (3 to 7 values per box plot), many tests did not yield a p-value.

Discussion

The ability to predict human phenotypes from metagenomic data is important for the discovery of biomarkers of disease and the subsequent development of therapeutics. However, a major issue that impacts prediction and biomarker discovery is the presence of confounders and systemic background noise both within [7] and across studies [33,38]. In this paper, we investigated the ability of different denoising transformations in combination with supervised correction methods to correct for sources of background noise in microbiome data and evaluated the utility of an unsupervised approach – PCA correction on CLR-transformed data. We recognize that fully correcting for background noise and population-specific factors, particularly in an unsupervised manner, is extremely difficult if not impossible. Further, biological variables associated with population-specific factors can be helpful for prediction of phenotype and applying correction approaches can potentially remove the effect of these variables. For that reason, we do not advocate for one approach over the other, but instead we highlight the issues that can arise when study-specific effects are not appropriately accounted for and demonstrate several approaches to combat these effects.

In this study, we analyze four datasets: AGP, HCHS, CRC-WGS, and CRC-16S. The AGP and HCHS datasets provided the opportunity to evaluate intra-study heterogeneity, whereas the CRC-WGS and CRC-16S datasets provided the opportunity to evaluate inter-study

heterogeneity. These are particularly unique datasets because they are either very large (AGP and HCHS), or they are comprised of several datasets measuring the same phenotype (CRC-WGS and CRC-16S), which is uncommon. For example, our decision to focus on CRC-WGS was motivated by important findings in Wirbel et al. [78] and Thomas et al. [1], two studies which compiled a collection of metagenomic samples from healthy and CRC individuals across a total of seven cohorts. Both these studies were able to find a core set of CRC-associated microbes despite differences in ethnicity, diets, and other host factors across studies. Both Wirbel et al. [78] and Thomas et al. [1] found that CRC classification models generalized effectively across studies and reported similar mean LODO AUCs of 0.81. We were able to also predict CRC with a similar accuracy of AUC 0.79 both before and after correction. In addition to CRC, we found prediction of BMI to be a useful analysis because it is notoriously difficult to predict accurately [87–89].

Given the diverse range of datasets available, there is not one data denoising transformation or correction method that outperforms the others universally, and multiple methods should be tested for phenotype analysis. This motivated a broad comparison of popular transformations and correction approaches. PCA correction has been effective in correcting for unwanted variation in human genetic data and morphological data [63–66], but to date has not been evaluated for correction of such noise in microbiome data. Yet, we and others have shown that top principal components in multiple datasets are correlated with numerous potential sources of unwanted noise such as host genetics [67], ethnicity of the host [68], and also abiotic factors like temperature [69], suggesting that PCA correction may be an effective unsupervised correction approach. We found that regressing out the top PCs after applying a CLR transformation may address multiple issues simultaneously: first, this approach can prevent inflation of false

positives associations (**Fig 3**), second, can maintain and, in the case of BMI, increase prediction accuracy of host-associated phenotypes in a LODO analysis (**Fig 4**).

Our comparison of correlations between PCs and study covariates sheds light on which datasets are good candidates for PCA correction. In the HCHS dataset, where PCA correction was most successful, correlation of technical covariates and not biological covariates with the top PCs increased significantly after CLR transformation (**Fig. 2**, **Fig S2**, and **Fig S4**). This potentially allowed for removal of technical noise without sacrificing phenotype signal, perhaps even enhancing the phenotype signal. The result was that application of CLR along with any correction method to both taxonomic abundances and *k*-mers was successful in increasing prediction accuracy (**Fig 4A**). On the other hand, the CRC-WGS and AGP datasets had an increased correlation of both biological and technical covariates with the top PCs after CLR transformation (**Fig S4**), making the removal of technical noise without removing phenotypic signal difficult. In these cases, applying any transformation or correction approach did not improve accuracy and instead in most cases resulted in similar performance to uncorrected data. Thus, the extent of background noise differs from one dataset to another, and the success of an unsupervised versus supervised method varies for each dataset (**Table 2**).

Despite correction approaches having limited effect on prediction ability for most datasets, these same correction approaches had a large impact on reducing false positive biomarker associations in our titration analysis. Specifically, we found that when performing association analyses, a supervised correction applied after a denoising transformation may be best and that transformations alone are insufficient to reduce false positive discoveries (**Fig 3**).

In this work, we show that CLR has comparable ability to other denoising transformations both when used alone and in combination with other correction approaches. The

application of CLR transformation can address many attributes of microbiome data that make it difficult to model including sparsity and non-normality, which existing unsupervised approaches designed for non-microbiome data [59–61] are ill-equipped to deal with (**Table 2**). As PCA assumes features are normally distributed, we produced Q-Q plots (**Fig S3**) showing that the quantiles of CLR-transformed data are close to the quantiles of a theoretical normal distribution. The application of CLR to microbiome data has been broadly recommended [25,90] and is part of a suite of methods known as Compositional Data Analysis (CoDA) [91,92] to address the dependency between features inherent to compositional data. However, the adoption of CLR in the microbiome field has not been uniform. Recently, McLaren at al. [33] discussed that CoDA methods' ability to make microbiome data invariant to multiplicative bias has been underappreciated within the field. Specifically, McLaren et al. [33] found that that ratio-based analyses could remove intra-study bias, though did not address its effect on multiple datasets that are pooled together or large datasets with heterogeneous sampling procedures such as the AGP. Here, we provide the first systematic investigation into the effect of how CLR in combination with PCA can remove inter-study and intra-study bias. We hypothesized that applying CLR transformation will more readily reveal the covariates that introduce technical background noise across and within heterogeneous datasets because these contributors of bias (e.g. DNA extraction method, sequencing instrument, etc.) have a multiplicative effect on relative abundances [33]. We found that indeed relationships between the microbiome and such variables is more apparent after CLR transformation, our observation of this in taxa abundance profiles makes sense in the context of multiplicative bias expounded by McLaren et al. [33] because the multiplicative bias becomes additive in log space, such that PCA is able to capture the bias in the top PCs as a shift in the centroid of samples plotted for a given dataset (**Fig 2**). Just as we found CLR

transformation can significantly effect PC correlations with covariates, the application of data transformations like variance-stabilizing from DESeq2 [45] and the log counts-per-million (logCPM) transformation from EdgeR [46] applied to taxonomic abundance also affect the correlation of variables with top PCs (**Fig S6**). Simiarly, these transformations can be helpful for phenotype prediction (**Figs S8** and **S9**).

We also compared the impacts of correction on *k*-mers and taxonomic features (**Table 2**). *K*-mers are a useful way to featurize data because they are not dependent on reference genomes. Moreover, short *k*-mers of size 5-8 have the added benefit of a Gaussian-like distribution (**Fig S5**) and low sparsity, unlike taxonomic features. However, *k*-mers have inherent limitations because they are usually not directly interpretable biological features. This limitation may be a reason why taxonomic feature abundance outperforms *k*-mers in phenotype prediction accuracy (**Fig S9**). It is crucial to note however, that *k*-mers may provide a better signature of technical artifacts like PCR bias [93,94] and are also known to be protocol specific [95]. Thus, this may explain why for both 16S and WGS data, *k*-mers had higher correlations with technical variables compared to taxonomic features (**Fig 2, Figs S2 and S5**). This aspect of *k*-mers offers a potential explanation for why PCA correction was particularly effective with *k*-mers for the HCHS dataset. Of note, these correlation analyses may reveal associations between linear effects of PCs and covariates, but not for non-linear effects. Other have also found that *k*-mers performed poorly compared to counts of reads aligned to reference genomes [96]. In predicting CRC and antibiotic consumption status, species profiles were more predictive whereas in predicting BMI, *k*-mers were more predictive under the majority of correction approaches when compared to application of the same approach to taxonomic abundance.

The supervised approaches [70–72] are beneficial in that they directly remove known confounding, potentially at the cost of phenotype prediction, while unsupervised approaches are can correct for both measured and unmeasured factors of microbiome (**Table 2**). Correcting for confounders and PCs both can result in the removal of phenotype signal, as is the case in ComBat [72] and fixed PCA (**Fig 4 and Fig S9**). Tuned PCA may reduce the removal of phenotypic variance by removing up to, but not including, the first PC that would significantly impact phenotype signal. However, caution must be taken when using tuned PCA in the presence of strong confounding as it may not remove all confounding to protect the phenotype effect. In these scenarios, one should consider either a liberal correction of confounding by correcting for more PCs or subsampling the data such that cases and controls are matched for known confounders as is done in Vujkovic-Cvijin et al. [7].

Background noise correction is becoming increasingly important as the microbiome field matures and new datasets become available. One exciting future application of correction that we foresee is in microbiome wide association studies in which microbiome genomic polymorphisms are associated with human phenotypes [97,98]. Such a scenario may benefit from background noise correction since population structure may play a considerable confounding role [99]. As researchers consider the best approach for background noise correction for their specific research questions, they must weigh the tradeoffs between addressing confounding while also maintaining as much of the phenotype signal as possible. There is no single solution that will address all problems, but at minimum researchers should perform careful forensics to investigate the nature and pervasiveness of confounders in their data. In this manner, consistent and robust inferences can be made across multiple studies, moving us towards the goal of accurate phenotype prediction from microbiome data.

Taxonomic features	*K*-mer features
• Pro: Find directly interpretable biomarkers of phenotype • Pro: May be better for prediction of binary phenotypes like colorectal cancer • Con: features are often rare, resulting in a sparse feature matrix unless features we are grouped to genus or family level	• Pro: Not reliant on reference genomes • Con: Features not immediately interpretable • Pro: May be better for prediction of certain continuous phenotypes like BMI • Pro: Short *k*-mer sizes are more Gaussian distributed and non-sparse
No transformation of features	**CLR transformation of features**
• Pro: Useful for compositional analysis. Sufficient when feature distribution meets assumptions regarding normality • Con: Compositional data does not meet assumptions of many types of differential abundance analyses.	• Pro: Useful to apply to compositional data before PCA for interpretability[100] • Pro: Produces a Gaussian-like distribution (log transformation may also accomplish this) • Con: May be problematic for correlation-based analyses[101] • Note: Other transformations (edgeR and DESeq2) may be useful
Supervised Correction	**Unsupervised Correction**
• Pro: Correction is targeted and most influential batch effects are explicitly accounted for • Con: Need metadata on experimental setup (batches or study-effect groups)	• Pro: Do not need information on batches or study-effect groups, but helpful for assessing signal of study effects • Pro: Multiple sources of noise can be corrected for simultaneously • Con: Correction is less targeted and biological signal may be sacrificed.

T 2 **Table 2.** Key considerations when performing background noise correction in metagenomic data.

Methods

Datasets

Raw 16S fastq files were downloaded from the NCBI Sequence Read Archive (SRA) with study accessions PRJEB11419 for the American Gut Project, and PRJNA290926 (Baxter et al.[80]) and PRJEB6070 (Zeller et al.[81]) for CRC-16S. Fastq files for Zackular et al.[82] from CRC-16S were obtained from http://mothur.org/MicrobiomeBiomarkerCRC/. The raw WGS fastq files

for CRC-WGS were downloaded from SRA with study accessions PRJEB12449 (Vogtmann et al.[84]), PRJEB10878 (Yu et al.[85]), PRJEB7774 (Feng et al.[83]), PRJNA447983 (Thomas et al. Italian validation cohorts[1]), PRJEB6070 (Zeller et al.[81]), and PRJNA389927 (Hannigan et al.[86]). Processed OTU data for the AGP was obtained from Qiita study id 10317 (EBI submission ERP012803). OTU profiles from CRC-16S were obtained from the MicrobiomeHD database (Duvallet et al.[6]). Taxonomic profiles for CRC-WGS were obtained through the R package curatedMetagenomicData[41] which used MetaPhlAn2[102]. In both 'MicrobiomeHD' and 'curatedMetagenomicsData', taxonomic abundances were computed in the same pipeline for each set of studies.

k-mer Processing

Features in metagenomic data can be defined in two broad ways, both high-dimensional: reference-based approaches and reference-free approaches. Reference-based approaches cluster sequenced reads based on a defined threshold and assign taxonomy by aligning reads to reference genomes. Reference-free approaches, sort reads into bins that are defined independently of known genomes, i.e. *k*-mers, short strings of length *k* that can be obtained directly from read sequences, which are increasingly popular in microbiome data analyses and have been used by several studies to do prediction . *K*-mers offer a powerful alternative approach to more commonly used taxonomic features, because they do not rely on a reference database of genomes and do not require identifying a set of parameters to determine taxonomic features .

To compute *k*-mer abundances, raw sequences from either 16S or whole metagenome sequencing were input into the *k*-mer counting algorithm Jellyfish 2.3.0[103] with default

parameters except for a hash of 10 million elements and canonical *k*-mers with size of 5, 6, 7 or 8. Prior work has shown that *k*-mer sizes of 6 and 7 are predictive of phenotype[76]. The resulting *k*-mer abundance table is then converted to a composition such that each sample sums to 1 to account for different reads depths across samples. Taxonomic profiles were similarly converted to compositions.

Centered log ratio transformation

The centered log ratio (CLR) transformation is a compositional data transformation that takes the log ratio of between observed frequencies and their geometric means. This is done within each sample where relative frequencies of different taxa are measured and sum to 1. This can be written in mathematical form as:

$$\mathrm{clr}(x) = \left[\log\frac{\mathrm{x_1}}{\mathrm{G(x)}}, \log\frac{\mathrm{x_2}}{\mathrm{G(x)}}, \ldots, \log\frac{\mathrm{x_n}}{\mathrm{G(x)}}\right]$$

$$= [logx_1 - log\ G(x), logx_2 - log\ G(x), \ldots, logx_n - log\ G(x)]$$

$$\mathrm{G}(x) = \left(\prod_{i=1}^{N} x_i\right)^{1/N}$$

Here, $\boldsymbol{x}$ is a vector representing the abundance of microbiome features in a single sample, and $G(\boldsymbol{x})$ represents the geometric mean. The Gaussian-like distribution of CLR-transformed microbiome compositional data is shown in **Figure S3.** We added a pseudocount equal to 0.65 times the minimum non-zero relative abundance, following zero-replacement strategies as suggested by , prior to applying the CLR transformation.

Background noise correction methods

The existing supervised approaches for background noise correction compared in this study include percentile normalization[43], batch mean centering (BMC)[70], ComBat[72], and limma[71] applied to relative abundance data. ComBat[72] assumes data is cleaned and normalized prior to batch effect removal. We added a pseudocount equal to 0.65 times the minimum non-zero relative abundance, following zero-replacement strategies as suggested by . It's common to add a pseudocount to 0 relative abundance observations so that one can apply a log transform in the normalization prior to ComBat[72] (as described in Gibbons et al.[38]). We followed this same procedure with both OTU and *k*-mer, and applied ComBat[72] and limma[71] to the log of relative abundance data. For percentile normalization, batch mean centering (BMC), and Direct Covariate Correction (DCC) we used the relative abundance.

For phenotype prediction and titration analysis, a relative-abundance feature is needed. ComBat, limma, and PCA corrected data will often produce non-positive data that does not resemble counts. To create count-like data we took the exponent of the resulting ComBat and limma corrected data produces count like features.

The CLR transformation and PCA-Correction used the relative abundance of *k*-mers and taxonomic features. The equation used to regress out confounding covariates in DCC is as follows:

$$X^{m\times n} \sim \beta^{\mathrm{m\times b}} C^{b\times n} + \epsilon^{\mathrm{m\times n}}$$

Where the original feature matrix X with m features and n samples is the outcome of a linear model with covariate associated coefficient matrix β, dummy matrix C with each row representing one of the b possible values of the confounding covariate, and ϵ, the residual matrix. The residual matrix ϵ is the covariate-corrected feature matrix. To perform titration and

downstream prediction analysis on PCA-corrected data, we performed an inverse-clr as implemented in the compositions R package to convert data to relative abundance.

In PCA correction, top PCs computed from the CLR transformed *k*-mer or OTU relative abundance tables are regressed out. The CLR transformation cancels out the multiplicative bias within each study by taking a ratio of features to the geometric mean of features that are all impacted by the same study-specific multiplicative bias. The transformation accentuates the difference in bias across studies by smoothing out the intra-study bias, thereby allowing PC regression to account for the confounding across studies. In the fixed PCA correction, a set number of PCs are regressed out from the microbiome data. In the main figures we show results after regressing up to three PCs. Alternatively, the tuned PCA correction uses a train-validation-test approach to tune two hyperparameters: the optimal number of PCs to regress out *p*, and, when using *k*-mers, the optimal *k.* The same portion of data used for validation in the Random Forest tuning is used for tuning the PCA correction hyperparameters, where the tuned Random Forest hyperparameters are fixed before tuning *p* and *k*. To determine the number of PCs that optimize phenotype prediction, PCs 1 through *p* were regressed out of the input data with *p* ranging from 1 to 20. The *p* that produces the highest AUC or Pearson correlation in phenotype prediction (method of prediction model described below) in validation was selected. The same procedure is done with *k* where values between 5 and 8 are tested (only *k*-mer sizes 6 and 7 were tested for CRC-WGS) The reported performance is based on the remaining 20% set aside for testing.

Correlation analyses

To compute the correlation of PCs with covariates before and after CLR correction, we used canonical correlation analysis using the 'canCorPairs' function in the R package variancePartition[104]. We used canonical correlation because several covariates were categorical, with the result that only positive correlation values can be calculated. The distribution of correlations before and after CLR transformation were statistically compared using the two-sample Wilcoxon signed-rank test.

Phenotype prediction

In CRC-16S and CRC-WGS, we predicted whether a sample comes from a host with colorectal cancer or a healthy host. For the American Gut Project, we predicted whether a sample comes from a host who took antibiotics in the previous year or a host who has not taken antibiotics in the previous year. We also use the American Gut Project to predict body mass index (BMI).

We performed prediction of binary traits using Random Forest implemented in Scikit-learn[105], which has been previously employed successfully for predicting binary outcomes from microbiome data[1,41,106,107]. We tuned four hyper-parameters of the Random Forest model in a grid search using a train-validation-test strategy. In the LODO framework, one study was reserved for testing while the remaining studies were split such that 70% of samples were used for training and 30% for validation of model hyper-parameters. In the non-LODO framework, 56% of samples in the meta-cohort were used for training, 24% for validation of model hyper-parameters, and 20% reserved for testing, where the distribution of studies or sub-cohorts were similar in the test, train, and validation sets. Six hyperparameters where four were tuned in a grid search: estimator trees (100, 1000, or 1500), criterion (entropy only), minimum samples per split

(2, 5, or 10), minimum samples per leaf (1, 5, or 10). Two hyperparameters were trained using the following settings: max depth of trees was set at 'None' (nodes are expanded until all leaves contain only one class or until all leaves contain less than min_samples_split samples[105]) and maximum features was set to "auto" (set to square root of number estimator trees[105]), and default parameters otherwise. This was performed in five-fold cross validation repeated ten-times to obtain confidence intervals on the area under the ROC curve (AUC), our metric of prediction accuracy. A similar train-validation-test strategy was used for the linear regression model to select coefficients of the model where accuracy was measured using Pearson correlation of the true BMI to the predicted BMI. The difference in the distribution of prediction accuracy for both prediction tasks was quantified statistically using a Wilcoxon rank-sum test.

Titration

Following the procedure from Gibbons et al.[43], samples from different studies were pooled together to assess the inflation of false positive associations. The minimum class membership across two studies was used as the set sample size drawn from the case and controls for each study for a given titration experiment. A fraction of 0, 25, 50, and 100% controls in the first study were replaced with controls from a second study. The filtering of features as implemented in Gibbons et al. required features resembling relative abundance, and we therefore, applied the appropriate transformations to convert ComBat, limma, and PCA-corrected data to relative abundance. For ComBat and limma, we applied the natural exponent of the matrix. For CLR-transformed data (including PCA-corrected data), we applied the 'inverse clr transform' as implemented in the 'compositions' package in R[77].

Acknowledgements

We thank members of the Halperin Lab and Garud Lab, as well as Michael R. McLaren for helpful discussions.

CHAPTER 3: SNV-FEAST: microbial source tracking with single nucleotide variants

Leah Briscoe[1,8]*, Eran Halperin[2,3,4,5,6,9], Nandita R. Garud[3,7,10]*

9. Bioinformatics Interdepartmental Program, University of California Los Angeles, Los Angeles, CA, United States of America
10. Department of Computer Science, University of California Los Angeles, Los Angeles, CA, United States of America
11. Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, United States of America
12. Department of Computational Medicine, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, United States of America
13. Department of Anesthesiology and Perioperative Medicine, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, United States of America
14. Institute of Precision Health, University of California Los Angeles, Los Angeles, CA, United States of America
15. Department of Ecology and Evolutionary Biology, University of California Los Angeles, Los Angeles, CA, United States of America
16. Twitter handle @leahbriscoe
17. Twitter handle @halperineran
18. Twitter handle @nanditagarud

*Correspondence to leahpbriscoe@gmail.com and ngarud@ucla.edu

Abstract

Elucidating the sources of a microbiome can provide insight into the ecological dynamics responsible for the formation of these communities. Source tracking approaches to date leverage species abundance information, however, single nucleotide variants (SNVs) may be more informative because of their high specificity to certain sources. To overcome the computational burden of utilizing all SNVs for a given sample, we introduce a novel method to identify signature SNVs for source tracking. Signature SNVs used as input into a previously designed source tracking algorithm, FEAST, can more accurately estimate contributions than species and provide novel insights, demonstrated in three case studies.

2. Keywords: **Source tracking, microbiome, single nucleotide variants, transmission, strains**

Background

Understanding the sources that could contribute to the formation of a given microbiome is of great interest in elucidating the ecological processes that give rise to these complex communities and the impact of these communities on human and environmental health. For example, a hospital environment may introduce antibiotic resistance genes to an infant gut microbiome, and local selective pressures may result in vastly different microbial compositions in different parts of the ocean. Approaches for determining the proportion of a microbiome of interest (the "sink") that is attributed to different microbiomes (the "sources") is known as "source tracking" [108,109]. Source tracking is useful for forensics, categorization of samples, detecting contamination, and tracing transmissions between different hosts or environments. While source tracking was developed as a way to quantitatively characterize a sample based on a set of samples with known origin, in most studies, the true source of samples may never be collected. In these cases, source tracking approaches are useful in identifying similarities between microbiome samples even if they cannot be used to definitively identify the true source of origin.

Current approaches for source tracking include the Bayesian approach, SourceTracker [108] and more recently the expectation-maximization approach, FEAST [109]. These source tracking methods use species abundance profiles of the sample of interest (the sink) and of potential sources and compute percentages of sinks that are attributable to each potential source. However, species abundance profiles miss important sub-species single nucleotide variants (SNVs), which may provide higher resolution information than species about transmission patterns. For example, Nayfach et al. 2016 [110] found that the sharing of microbiome SNVs private to mothers and their infants decreases over the first year of the infant's life while species sharing increases.

This suggests that while the infant microbiome increasingly resembles the adult microbiome ecologically, sources other than the mother also colonize the infant. Thus, species-level resolution may obscure true sources of microbes while SNVs can reveal actual transmissions to the infant.

While tracking strain transmissions with SNVs has been highly successful in a number of studies [110–116] current approaches to strain tracking are limited. These methods provide binary information by inferring whether or not a strain transmission has occurred per species but they do not shed light on the relative proportions of microbiomes that are similar. A specific example of this is inStrain [113] which computes a pairwise population-level average nucleotide identity (popANI) between two samples. If an infant harbors several strains derived from the mother at low frequency, these shared strains will have high popANI values, but they will represent a relatively small proportion of the infant's microbiome. By contrast, source tracking allows us to simultaneously infer the putative proportions for multiple sources contributing to a given sink, integrated over all community members in the sink. As shown in **Fig. 1**, one may be able to estimate that an infant microbiome is explained 25% by their mother, 10% by their dog, and 30% by unknown sources [108,109]. In other words, source tracking with SNVs leverages not only the genetic variants within species, but also the relative abundances of the species that carry the SNVs.

Here, we evaluate whether source contributions estimated with SNVs are more accurate than with only species when they are provided as input to FEAST [109] (hereafter referred to as SNV-FEAST and species-FEAST, respectively). FEAST [109] is faster and more accurate than previous source tracking tools [108] and therefore, is ideal for adaptation to SNV source tracking since it can accept larger numbers of features and input sources. Despite this improved

computational efficiency, the potentially millions of single nucleotide variants across all microbiome species in a given host still can computationally overwhelm FEAST. To address this, we introduce a novel approach to determine signature SNVs that can be used as input to FEAST. This both reduces memory requirements and computation time in the FEAST estimation, allowing us to optimally estimate the source contribution of a sink. We find that SNV-FEAST and species-FEAST yield different outcomes when applied to simulated data, with SNV-FEAST frequently out-performing species-FEAST. We apply SNV-FEAST to three real-world case studies, including source tracking between infants and their mothers in the first year of life, between infants and the neonatal intensive care unit (NICU), and between oceans around the world. We confirm the ability of SNV-FEAST by recapitulating several previously published findings in our case studies, as well as discover new source tracking patterns across oceans. In sum, we show that SNVs can be used to estimate potential transmissions across hosts and across environments.

Results

SNV-FEAST algorithm

Here we adapt FEAST to accept SNV abundance instead of species abundance as input. A computational challenge in using SNVs instead of species as input to FEAST is that SNVs contribute a significantly larger feature space. The number of different species comprising a microbiome can range from a few hundred to a few thousand, while the number of possible SNVs for a given species alone can be in the thousands [117]. This difference in number of input

features can result in FEAST runtimes that last several hours instead of a few minutes and memory intensive storage of read counts at all sites of variation.

We devised a likelihood-based approach for selecting a set of informative or "signature" SNVs for a given source tracking analysis, allowing us to overcome the time and memory intensive challenges of utilizing SNV-level data. We identify these informative SNVs by computing a signature score (**Fig. 1A**) (see **Methods)** that quantifies the extent to which SNVs in the sink are most likely derived from one of the potential sources. This is analogous to identifying SNVs private to sources and their sinks, but more generalized to include SNVs that may be found in multiple sources, albeit at higher frequency in one of the potential sources (see **Methods**).

To compute a signature score for a given SNV, two hypotheses are compared for each potential source: (1) that one source solely explains the observed allele counts in the sink and (2) all sources except that one source collectively explain the observed allele counts in the sink. For each hypothesis, we calculate the binomial log-likelihood for the estimate of the allele frequency in the sink, θ.

Hypothesis 1: Source *i* with allele frequency p_i explains the allele counts in the sink.

$$\hat{\theta} = p_i$$

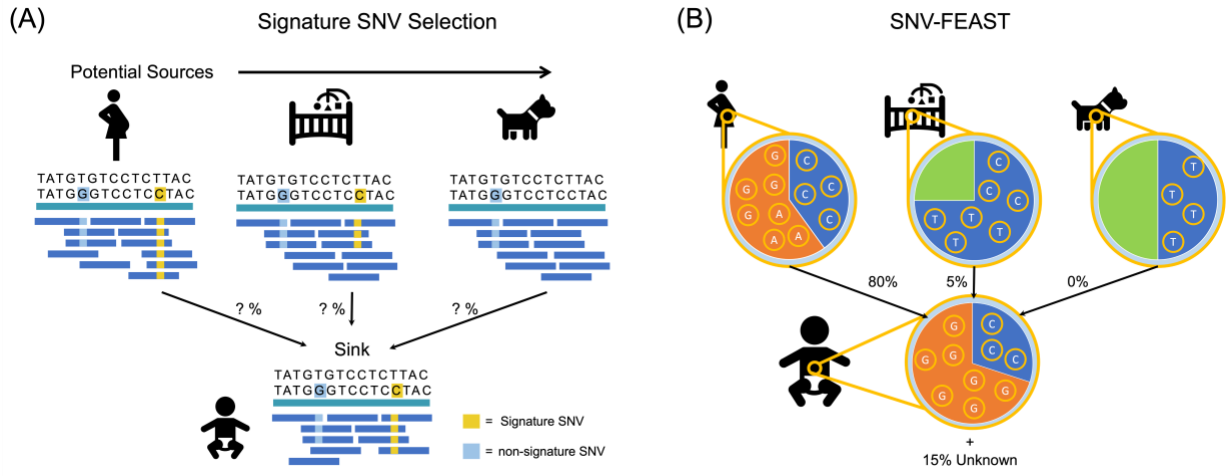
Hypothesis 2: A combination of all other sources except *i* (sources $j \neq i$) explain the observed allele count distribution in the sink. The estimate of the sink allele frequency is computed using a mixture of the allele frequencies p_j from those sources. The mixing parameter α_j is learned using Sequential Least Squares Programming with the constraint that $\sum_{j \neq i} \alpha_j = 1$.

$$\hat{\theta} = \sum_{j \neq i} \alpha_j p_i$$

The binomial log-likelihood is calculated as follows, where there are *n* reads with the reference allele and *m* reads with the alternative allele in the sink.

$$LL(\hat{\theta}) = n\, log\, \hat{\theta} + m\, log(1 - \hat{\theta})$$

A log likelihood ratio representing the support for hypothesis 1 relative to hypothesis 2 is calculated per site per potential source. The maximum log likelihood ratio per site is the signature score for that SNV, representing how favorably one of the sources explains the sink over all other sources. Signature SNVs are those with scores greater than two standard deviations over the mean signature score computed for all SNVs (**Methods**).



5 **Figure 1.** Signature SNV selection and SNV-FEAST. (A) A signature SNV is present in one or few but not all sources. By contrast, a non-signature SNV is generically present in multiple sources and thus provides little discriminating information. (B) SNV-FEAST estimates the proportion a given sink derived from various sources using the read counts for each allele in sinks and sources.

Evaluation of SNV-FEAST in simulations

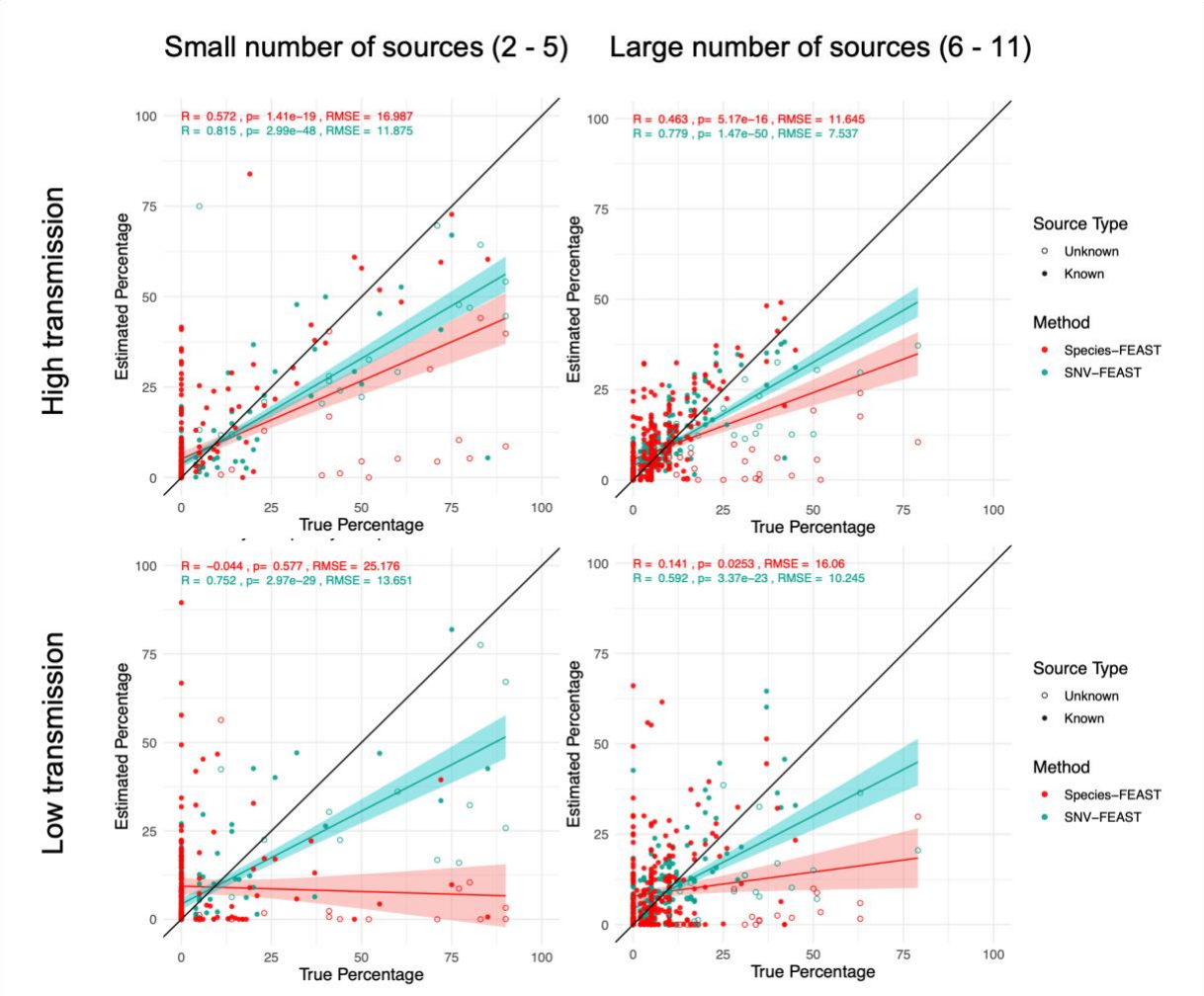
To compare the accuracy of species-FEAST and SNV-FEAST, we performed simulations mimicking mother-infant transmissions with the goal of estimating contributions of different sources to an infant sink. Our simulations tested the ability of SNVs and species to recapitulate the true source composition in synthetic samples comprised of a mixture of reads drawn from multiple real fecal adult samples. To construct these synthetic infant microbiomes, we mixed metagenomic data from mothers sampled in a mother-infant dataset [118] at various proportions as described below (**Methods**).

The difficulty of source tracking increases with the number of contributing sources [109]. Thus, we simulate infants that have a small (<=5) versus large (6 – 10) number of contributing sources (Additional file 1: **Table S1**), including an unknown source (e.g. a randomly selected unrelated mother). Known source contributions to the simulated gut microbiome sample of the infant were varied between 1 and 90% while the unknown contribution varied between 10 and 90%. The unknown source was not presented to FEAST as a potential known source.

Additionally, not all species in a mother are transmitted to the infant [111,112,114,119,120]. Thus, in our simulations, species transmission rates were determined using a beta distribution, which is a natural model for values between (0,1) and often proposed for microbial abundance data [121–124] (see **Methods**). We therefore consider four simulated scenarios: a combination of low versus high number of sources and low versus high transmission rates (see **Methods**).

Fig. 2 compares the performance of SNV-FEAST and species-FEAST in estimating the true contribution of sources. FEAST using SNVs has equal if not better performance than species in most scenarios and performs especially well when transmission rates are low and unknown source proportions are high. SNVs have a lower root mean squared error (RMSE) compared to

species in three of the four scenarios and higher Pearson correlation between true and estimated contributions in all four scenarios. The difference in these correlations for SNVs versus species is significant in all four cases when using a paired Wilcoxon signed rank test (high transmission: p-value = 0.00560, 0.00251 for small and large number of sources, low transmission: p-value = 0.00024, 0.002340 for small and large number of sources). These results suggest that SNVs may offer useful signatures of transmission.



6 **Figure 2. Ability of SNV and species-FEAST to recapitulate true contributions in simulations.** Estimated known and unknown source proportions for infant microbiomes simulated with in silico mixtures of real maternal fecal microbiomes under different scenarios: either low number of contributing sources (<=5) or high number of sources (6-11), and high transmission rate of species or low transmission

rate. The transmission rate is the probability of an infant being colonized by a given species, simulated using a beta distribution centered on the relative abundance of species in sources (**Methods**). 23 infants were simulated with five or fewer sources and 19 infants were simulated with a large number of sources (**Table S1**). The black line indicates the ground truth for proportions. For each simulated infant, there are 11 points plotted, whereby 10 correspond to known sources (some of which have zero contribution), and one corresponds to an unknown source which are indicated by hollow circles in the plot.

To assess whether all species and all signatures SNVs in the sink are needed for accurate source tracking, we varied the proportion of species (10%, 50% or 100%) and SNVs (10%, 50% or 100%) included as inputs to the algorithm (Additional file 1: **Fig. S1**). We used Pearson correlation between the true and estimated proportions to represent accuracy of SNV-FEAST. When decreasing the percentage of SNVs used, there is no statistically significant change in the performance. However, when decreasing the percentage of species used, there are statistically significant decreases in the performance (Additional file 1: **Fig. S1**).

To illustrate the advantage of SNV-FEAST over traditional strain tracking approaches such as inStrain [113], we used the same synthetic communities produced in the above simulation for inStrain profiling between each infant and each of their potential contributing sources (Additional file 1: **Fig. S2**). InStrain computes a popANI score, which represents the average nucleotide identity between two different metagenomic samples for a given species. As per the inStrain paper, popANI values > 99.999% represent the same strain being shared between samples for a given species (**Methods**). However, this approach provides a binarization as to whether or not a strain was transmitted and does not account for the relative abundance of the strain in the sink. Thus, we computed the fraction of each infant's species that have popANI ≥99.999%, with each potential source.

As expected, both SNV-FEAST and inStrain produce estimates of sharing that correlate positively with the ground truth mixture proportions of the contributing source samples in each

infant (Additional file 1: **Fig. S2**). We found inStrain results yielded a 0.742 Pearson correlation (p-value < $1x10^{-12}$) with the true mixture proportions, whereas SNV-FEAST has a 0.866 Pearson correlation (p-value < $1x\ 10^{-12}$) with the true proportions. The higher correlation values for SNV-FEAST likely reflect that relative abundances of strains and their genomic identities are simultaneously taken into account for source tracking, whereas inStrain only accounts for genomic identities. Finally, several of the shared species in the simulations had popANI values < 99.999%, reflecting the complex mixtures from multiple sources.

We next compared SNV-FEAST with the strain tracking procedure in Nayfach et al. 2016 [110]. Again, we used the same synthetic communities produced in the simulation to determine marker alleles as defined in Nayfach et al. 2016 (**Methods**). Here a marker allele is determined to be a SNV that is private to mother, infant, or the mother-infant dyad, and absent from the background population, which consisted of other samples in the dataset as well as samples from United States adults in the Human Microbiome Project [125,126] (**Methods**). Species with ≥ 5% marker allele sharing between mother and infant were deemed to share a strain (**Methods**). We found a high correlation between the true mixture proportions (on x-axis) and the percentage of species with transmission events (y-axis) (Pearson correlation 0.915, p-value < 1×10^{-16}) (Additional file 1: **Fig. S3A**). The higher correlation for the Nayfach et al. 2016 approach compared to the inStrain approach possibly reflects horizontal gene transfers between lineages residing in infants and mothers. By contrast, there was a lower correlation between the true mixture proportions (x-axis) and the sharing for all marker alleles across species present in the infant (y-axis) and (0.575 Pearson correlation, p-value < 1×10^{-16}) (Additional file 1: **Fig. S3B**).

Source tracking in infants over the first year of life

Having assessed the abilities of SNV-FEAST in synthetic data, we next estimated the contribution from the true mother over time to the true infant with SNV and species-FEAST in the Backhed et al. 2015 dataset. This dataset is composed of metagenomic samples from infants collected at four days, four months, and 12 months after birth, as well as their mothers at the time of delivery. Previous analyses on this data have shown that even while species similarity increases, infants and their mothers share fewer proportions of strains over time as revealed by sharing of SNVs private to mother-infant dyads [110]. Thus, SNVs belonging to strains shared only by the infant and their mother may be more informative of the true source compared to species. Here we sought to test whether SNV and species-FEAST recapitulate these results (**Methods**).

In applying FEAST to the Backhed et al. 2015 dataset, we estimated the proportion of the infant sample at birth attributable to their own mother. For 4 month-old infants, we estimated the proportion attributable to the mother and itself at birth. For 12 month-old infants, we estimated the proportion attributable to the mother and itself at birth and four months [109]. This allowed "unknown" to be more strictly defined as the component of the infant microbiome that could not be explained by the mother. It also allowed us to better discern if completely new strains were acquired at the 4^{th} and 12^{th} months of life (that were not already acquired during previous life stages).

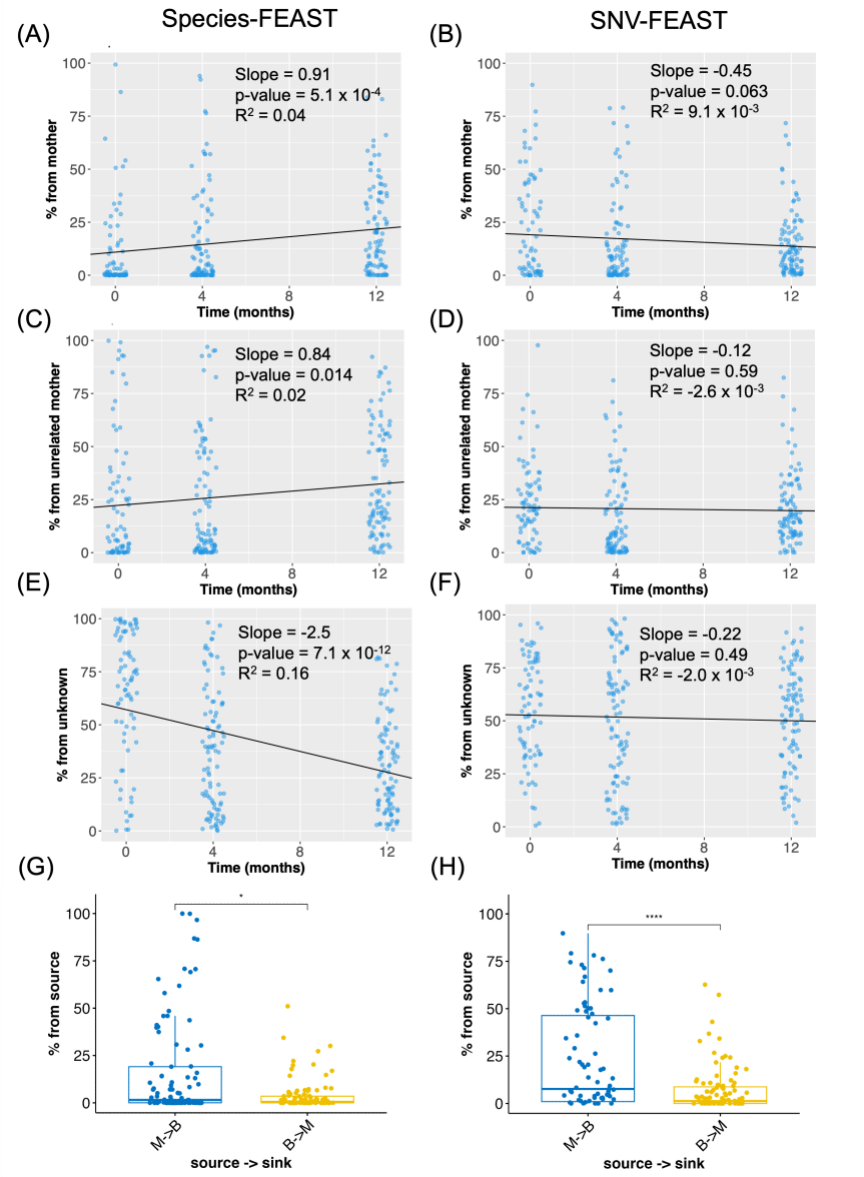
First, consistent with previous findings made with species and SNVs [110], species-FEAST estimates an increasing contribution of the mother over time (t-test p-value = 5.1×10^{-4}), but SNV-FEAST estimates a decrease over time (p-value = 0.063) (**Fig. 3**).

Second, we assessed the ability of species and SNV-FEAST to distinguish the true mother from three randomly selected unrelated mothers. Species-FEAST estimates an increasing

contribution of unrelated mothers over time (t-test p-value = 0.014) while SNV-FEAST estimates no significant change over time (t-test p-value = 0.59) (**Fig. 3**). The increase in contribution from unrelated mothers with species-FEAST does not suggest that these particular unrelated mothers are seeding the infant. Rather, the opposing trend observed with SNVs suggests that similarity at the species level is consistent with the maturation of the infant microbiome over time.

Finally, we estimated contributions from unknown sources, i.e. the proportion of the infant microbiome not explainable by the true mother, the three randomly selected unrelated mothers, or any previous time point. Species-FEAST estimates a sharp decline in contribution of unknown sources over the first year of life (t-test p-value =7.1×10^{-12}) (**Fig. 3**). This significant decrease in unknown at the species level reflects the infant microbiome maturation over the first year of life. By contrast, SNV-FEAST estimates little change in the contribution of unknown sources (t-test p-value = 0.49) (**Fig. 3**). Note that this unknown component reflects what was gained since a previous time point. In other words, at 12 months, the infant on average acquired the same fraction of unknown as it did at 4 months and birth. When source tracking is run without including previous time points as sources, the unknown component increases over the first year of life for SNVs only (Additional file 1: **Fig. S5)**.

Next, we sought to understand the effect of swapping sink and source in the re-analysis of Backhed et al. 2015 data. In **Fig. 3G and H**, the infant at birth is the potential source and mother is the sink. The estimated contribution from baby to mother is significantly smaller (species-FEAST: 11.9 difference, Wilcoxon rank sum test p-value = 0.013; SNV-FEAST: 16.0 difference, p-value = 2.2×10^{-5}) compared to that of mother to baby. This trend may be suggestive, but is not conclusive, of directionality, whereby a less diverse source is seeded by a more diverse source.



7 **Figure 3. Source tracking in the infant gut microbiome over the first year of life.** Species- and SNV-FEAST were applied to Backhed et al. 2019 data to estimate the contribution of (A, B) mother, (C, D) unrelated mothers and (E, F) unknown sources to infants sampled at birth, four months, and twelve months. The black line and inset statistics pertain to the linear regression fit for the source estimates as a function of age of the infant. (G, H) are swapped source tracking analyses with mother and infant swapped when using species-FEAST and SNV-FEAST, respectively. Additional file 1: **Fig. S4** shows the species that were included in species-FEAST and species that had SNVs included in SNV-FEAST. Additional file 1: **Fig. S5** shows the estimate of the unknown component when previous time points of the infant are excluded from the sources.

Contribution of the NICU built environment to infant microbiomes

Next, we re-analyzed a metagenomic dataset studying the contribution of the hospital

environment to the infant gut microbiome in the neonatal intensive care unit (NICU) (Brooks et

al. 2017). This dataset is composed of microbiomes of infant stool, as well as the NICU rooms of the same infants at frequently touched surfaces, sink basins, the floor, and isolette-top sampled over an 11-month period [127]. We applied SNV and species-FEAST to assess the contribution of the infant's own NICU room as well as a different NICU room in the vicinity to the infant's gut microbiome (**see Methods**).

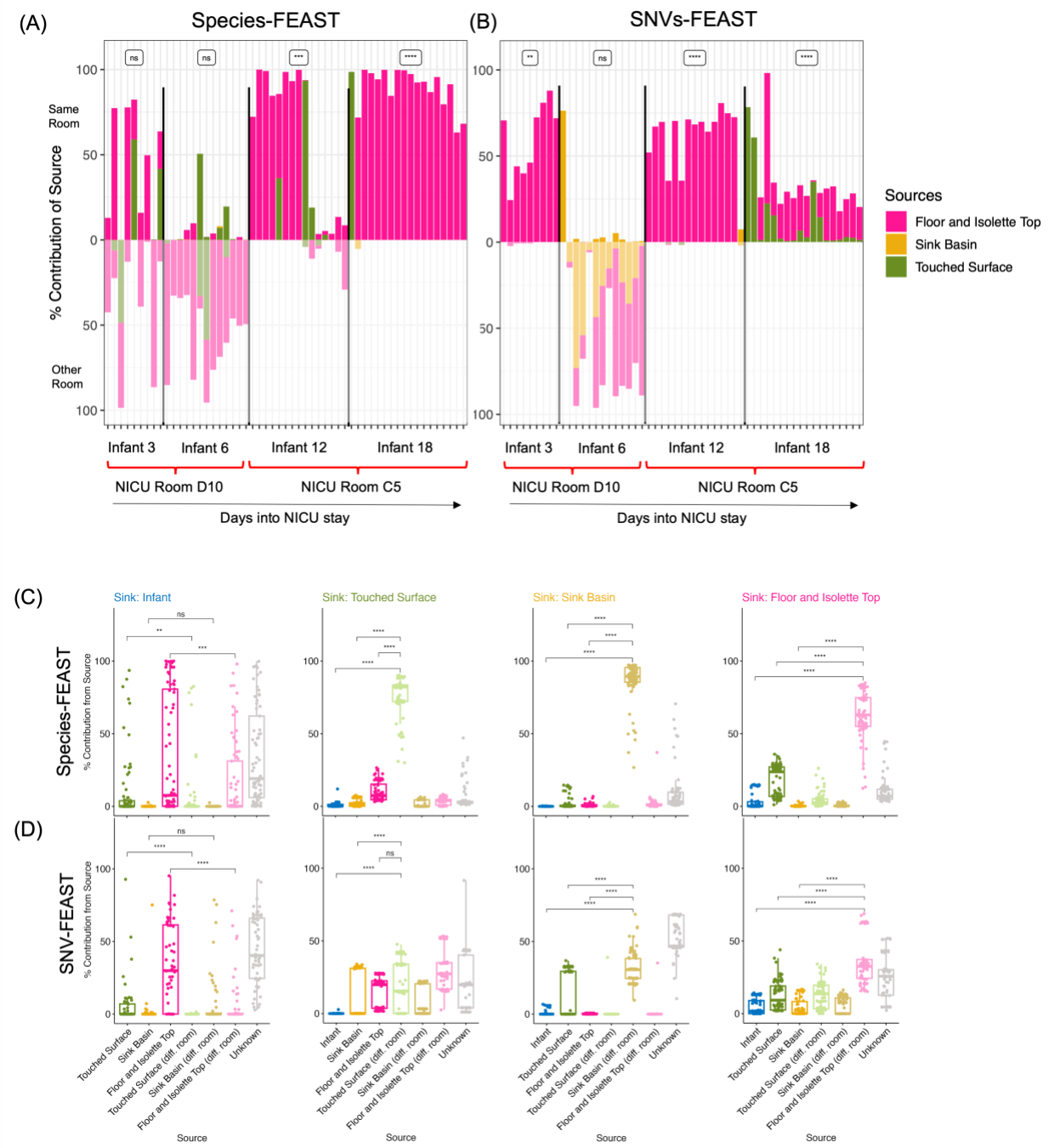
Concordant with the findings of Brooks et al., both SNV and species-FEAST detected that the most common source contributing to the infant microbiome was the floor and isolette-top from the infant's own room (**Figures 4A and B**). SNV-FEAST found Infant 18 also had large contributions from their own room's touched surfaces at multiple time points (**Fig. 4B**), which is consistent with a finding by Brooks et al. that three strains found in Infant 18 perfectly matched (> 99.999% average nucleotide identity) strains found in the touched surfaces samples of Infant 18's own room. Lastly, both species-FEAST and SNV-FEAST found Infant 6's microbiome was explained almost entirely by samples from a different room with SNV-FEAST finding a sizeable contribution from both the floor and isolette top and the sink basin in this different room. This is concordant with Brooks et al.'s finding of multiple cases of strain sharing across rooms of Infant 6 and 12 for the different surfaces. FEAST with both data types can quantify the extent to which Infant 6's microbiome was influenced by strains present in the built environment.

Through application of SNV and species-FEAST, we can quantify any time trends for the influence of the built environment on the infant microbiome (**Figures 4A and B**). SNV-FEAST more consistently finds that contribution from the infant's own room exceeds contributions from a different room over time (paired Wilcoxon signed rank test for same room > different room: Infant 3: p-value = 1.95×10^{-9}, Infant 6: 1.0, Infant 12: 3.05×10^{-5}, Infant 18: 3.81×10^{-6}) as compared to species-FEAST (Infant 3: p-value = 0.41, Infant 6: 1.0, Infant 12: 5.8×10^{-4}, Infant

18: 3.81 x 10^{-6}). Interestingly, species-FEAST assigns one dominant source primarily, whereas SNV-FEAST more often finds a combination of sources for a given sample.

Additionally, both SNV and species-FEAST estimated a large unknown component for all four infants, with Infant 18 showing the largest mean unknown component across the NICU stay based on SNVs (Additional file 1: **Fig. S6**). This unknown component is important because it signifies the extent to which other sources such as the mother and diet impact infant gut colonization.

We then asked the question is the infant more explained by the built environment rather than vice-versa, the built environment is more explained by the infant. We tested this by swapping the infant and each of the three built environment sources (**Fig. 4C and D**). The estimated contribution of room to infant is significantly higher than the estimated contribution of infant to room, but this asymmetry is more pronounced with SNV-FEAST. SNV-FEAST showed significantly higher contribution of room to infant for two of the three surface types (floor and isolette top: Wilcoxon rank sum test p-value = 7.00x 10^{-9}, touched surface: p-value = 0.0058, sink basin: p-value = 0.274) while species-FEAST found this to be true for one of the three surface types (floor and isolette top: Wilcoxon rank sum test p-value = 7.1x 10^{-5}, touched surface: p-value = 0.968, sink basin: p-value = 0.998). Interestingly, the built environments of different rooms highly resemble each other. This is especially apparent with species-FEAST, suggestive of similar ecological forces operating in similar built environments. By contrast, SNV-FEAST reveals a higher diversity of contributing sources of the built environment samples to other NICU built environments, once again highlighting the utility of performing source tracking with SNVs.



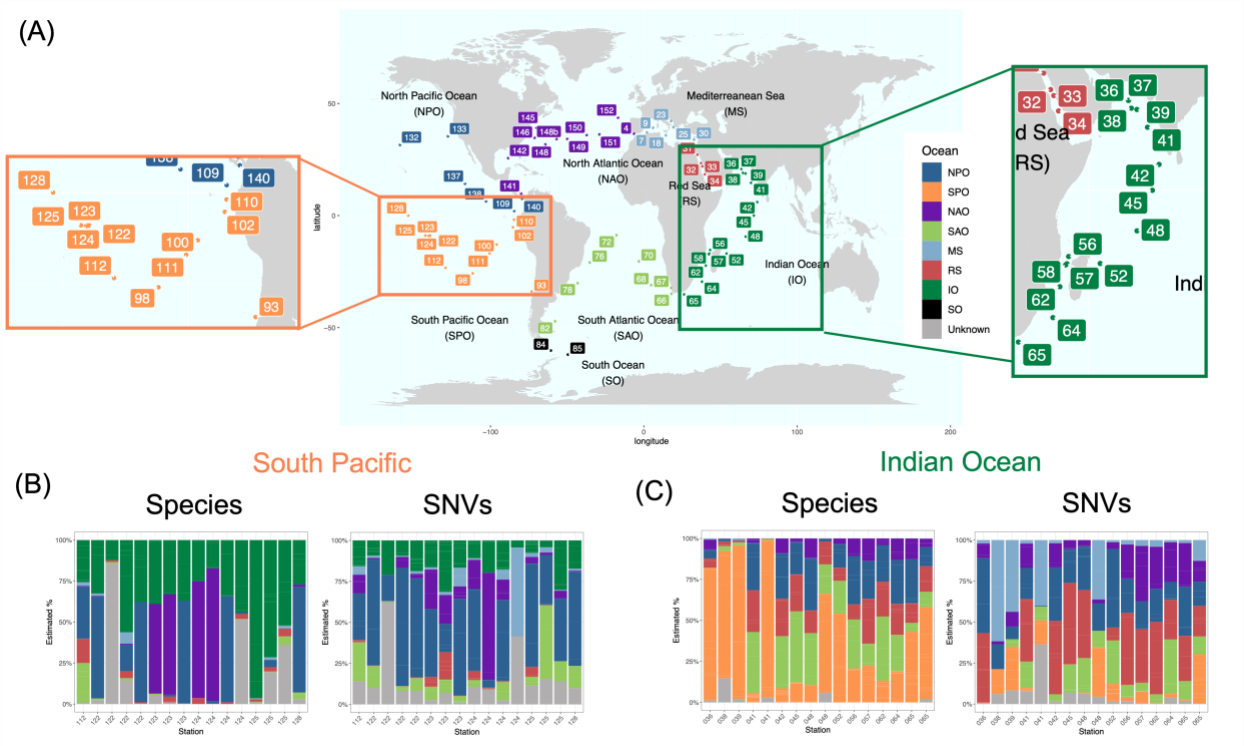
8 **Figure 4. Source tracking of infant gut microbiome in the NICU.** (A) species-FEAST and (B) SNV-FEAST applied to infants in the NICU. Each bar represents one sampling day in the NICU stay of an infant. Infants 3 and 6 stayed in the same room, but at different times. The same applies to Infants 12 and 18. The contribution of a different room was determined by using samples from Infant 12's room for Infants 3 and 6, and samples from Infants 6's room for Infants 12 and 18 for each of the categories of surfaces per infant: touched surface, sink basin, or floor and isolette top surface. The asterisks represent the result of a paired Wilcoxon signed rank test indicating whether the total contribution of surfaces from the infant's own room were higher than contributions from the other room. Iterative swapping of the infant sink and each potential source for source tracking with (C) species-FEAST and (D) SNV-FEAST. The first column shows source tracking results in which the infant was treated as the sink. In each column

after the first column, a different environmental source was swapped with the infant and considered as a sink. The brackets indicate the pairs of results that are compared using a paired Wilcoxon signed rank test. For all results, the following symbols represent the results of the statistical test: **** for p-value < 0.0001, *** for p-value < 0.001, ** for p-value < 0.001, * for p-value < 0.05, and n.s. for p-value > 0.05.

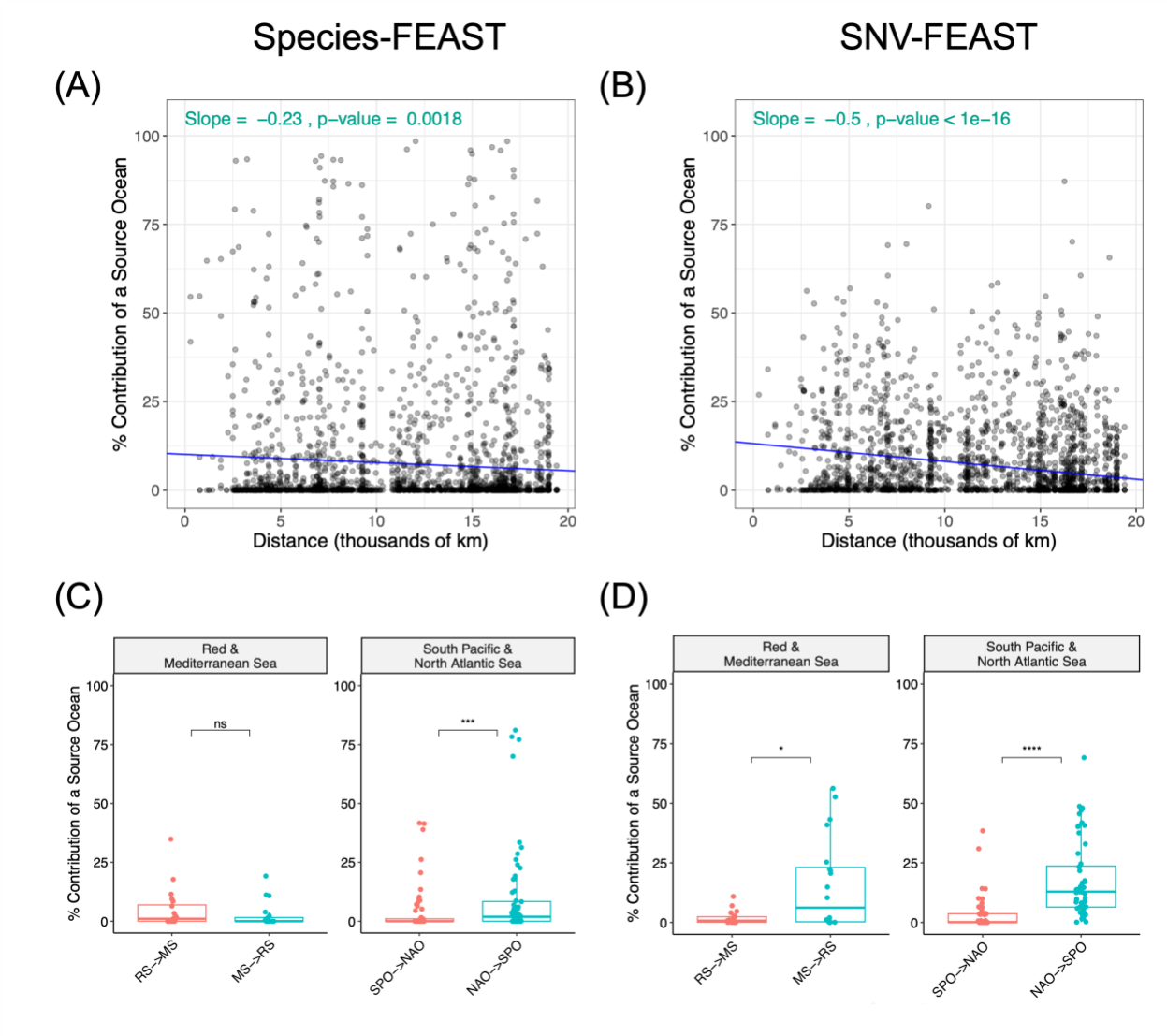
Global source tracking of ocean microbiomes

The ocean microbiome is a complex community that displays biogeography at the species and functional levels [110,128]. To further understand global patterns of ocean microbiomes, we applied SNV and species-FEAST to the Tara Oceans microbiome dataset [128]. In the source tracking context, rather than defining sharing as evidence of a transmission event (which is more likely in mother-infant data), estimated source contributions at best explain the extent to which a given ocean sample resembles other ocean samples. On one extreme, an ocean sample might be entirely explainable by a single ocean's samples, and at the other extreme, an ocean sample might be explainable by multiple oceans at the same time. Another alternative is for an ocean sample to not be explainable by any of the provided sources, resulting in a high unknown component and potentially suggesting high endemism. These source tracking estimates could be indicative of the extent to which oceans mix or may be reflective of similar niches.

Tara Oceans is composed of 182 whole metagenomic sequencing samples derived from 64 stations at multiple depths. Previous research indicates that temperature is one of the highest drivers of variability in microbial composition in the ocean [128,129]. For this reason, we restricted the source tracking analysis to sinks and sources from the same temperature and depth range: above 20 degrees Celsius and within an average of 5 meters below the surface.



9 **Figure 5. Microbial source tracking in the Tara Oceans dataset with SNV and species-FEAST.** World map indicating the location of sampling sites (A). Source tracking estimates for the contribution of different oceans to the South Pacific (n=16) (B) and Indian Oceans (n=16) (C) are depicted with vertical bars. In each experiment, all stations around the world excluding those from the "sink" ocean are considered potential sources. Light blue, for example, represents the total contribution of the four stations from the Mediterranean Sea that had samples in the surface layer that were also greater than 20°C in temperature.



10 **Figure 6. Source tracking with ocean samples**. Distance decay in contribution of a "source" ocean to a "sink" ocean when using (A) species-FEAST and (B) SNV-FEAST. In each experiment, only stations from one ocean were considered as sources for a given sink station. For example, when performing source tracking between the Mediterranean and North Atlantic, for each Mediterranean station, the 10 available North Atlantic stations were considered as potential sources. Thus, plotted are 10 points for a given Mediterranean sink, where each point represents the contribution of a source station from the North Atlantic to the Mediterranean sink station in question. Shown in inset text are the slope and t-test p-value for the slope. (C) and (D) are flipped source tracking analysis with the Red Sea and Mediterranean, as well as the South Pacific Ocean and North Atlantic Ocean using species-FEAST and SNV-FEAST, respectively.

First, we performed source tracking between oceans using SNV and species-FEAST. We treated each station around the world as a sink and estimated the contribution of different oceans

around the world to that sink (**Methods**). Unknown represents any portion of the microbiome in these sink samples that cannot be explained by any of the provided source samples. We found that species and SNV-FEAST estimate different amounts of sharing between oceans, where SNVs estimate a higher unknown on average, potentially indicative of endemism. The finding that SNV-FEAST estimates a higher unknown contribution on average is most evident in the North Pacific, North Atlantic, South Atlantic, and Mediterranean oceans (Additional file 1: **Fig. S7**). Additionally, in some oceans, SNVs identify contributions from oceans that species-FEAST does not detect (**Fig. 5,** Additional file 1: **Fig. S7**). For example, in applying FEAST to Indian Ocean samples we find that there is measurable sharing of microbes with the Mediterranean Sea, but this is not detected with species (**Fig. 5C**). Such differences were found in samples from other oceans as well (Additional file 1: **Fig. S7**).

Next, we assessed whether source tracking estimates display a distance-decay relationship. Previous studies found that genetic distance, such as that represented by fixation index F_{ST}, increases with geographic distance between populations [130,131]. Based on these findings, our expectation was that samples that are further away from a given station will have reduced resemblance to that station. To assess this distance-decay relationship, we plotted pairwise source tracking results across all possible pairs of ocean samples (**Fig. 6A and B**). We found that indeed as the distance increases, the % explainability of a given source ocean decreases -0.23 % per thousand km according to species-FEAST (t-test p-value $< 1 \times 10^{-16}$), and -0.5% per thousand km according to SNV-FEAST (t-test p-value = 0.0018). The steeper slope for SNV-FEAST suggests that SNVs may be more sensitive to distance decay signals on a global level.

Finally, we investigated whether some oceans have higher estimated contributions to other oceans than vice versa, potentially indicative of the directionality of transmissions (though see Discussion). Specifically, we investigated the relationship between the Red Sea to the Mediterranean Sea (**Fig. 6C and D**). Migration from the Red Sea to the Mediterranean, known as Lessepsian migration, is well-documented for not only microorganisms but also macroorganisms like fish [132–134]. Additionally, recent studies may suggest that anti-Lessepsian migration of bacteria (Mediterranean to Red Sea) is more common than Lessepsian migration [135]. Studies find that Mediterranean has brine pools that produce a similar environment to the Red Sea's [136], allowing for bacteria from the MS to potentially thrive in the RS.

By swapping the Red Sea and Mediterranean as source and sink, we found that there was indeed a significant difference in the estimated contribution from one direction to another with SNVs but not species (**Fig. 6C and D**). SNV-FEAST found the Mediterranean explained an average of 15% of the Red Sea, while the Red Sea explained an average of 1.8% of the Mediterranean (Wilcoxon rank sum test, p-value =0.02), consistent with anti-Lessepsian migration. Meanwhile, a similar analysis with species-FEAST found the Mediterranean explained 2.5% of the Red Sea and the Red Sea explained 4.9% of the Mediterranean (Wilcoxon rank sum test, p-value = 0.25). In a similar analysis between North Atlantic and South Pacific we found that both species and SNVs supported significantly greater contributions from the North Atlantic to the South Pacific, with SNV-FEAST estimating a greater contribution (17%, Wilcoxon rank sum test p-value = 5.1×10^{-11}) than species-FEAST (10%, Wilcoxon rank sum test p-value =1.8×10^{-4}). The same analysis performed in the other oceans is presented in Additional file 1: Fig. S8.

Together, these results suggest that on average, SNV and species FEAST generate similar source tracking results in the Tara Oceans dataset, with SNVs displaying stronger signals of endemism, distance-decay relationships, and potential directionality of transmission.

Discussion

Source tracking provides insight into potential source contributions to a metagenomic sample as well as similarities between metagenomic samples. While species abundances have been informative in source tracking in several studies [108,109,137–139], they may be limited in their resolution. SNVs provide a potential alternative because of their ability to distinguish sources of strain transmissions. Here we compared the ability of a previously published source tracking algorithm FEAST using species versus SNVs as input data. In application of species and SNV-FEAST to simulations as well as three case studies, we demonstrate that the two input types can provide distinct insights into microbial sharing and similarities across different environments. As a hypothetical example, two unrelated samples may have very similar species composition due to similar colonization processes and similar environmental influences without any actual microbial sharing. It would be unlikely for these two unrelated samples to share rare SNVs, however. This distinction suggests that SNVs indeed can provide insight into the ecological processes shaping microbial communities that species information alone cannot, and our three case studies are able to demonstrate this.

In the first case study, we confirmed previous findings that SNV sharing between mothers and infants decreases over the first year of life while species sharing increases [110], suggesting that while the infant microbiome matures to resemble adults at the species level, sources other than the mother may seed the infant over time. In the second case study, we confirmed source contributions from the NICU built environment to the infant microbiome [127], and found that

SNVs detect a more consistent estimate in source contributions over time compared to species as well as detecting contribution from sources not detectable by species-FEAST.

In the third case study, we perform source tracking in the Tara oceans dataset [128] and found SNVs display a stronger distance decay relationship than species. These distance-decay results parallel recent findings made with gene content [140]. While previous studies have examined the biogeography of the ocean using species profiles, genes [110,140] or amino acid variants from a single species (SAR11) [141], for the first time, we leverage the use of SNVs across all detected prevalent species in the ocean microbiome to identify proportions of sharing across oceans. A benefit of using SNVs in the ocean microbiome is that SNVs can track fragments of DNA that have moved due to horizontal gene transfer in the distant past rather than relying on inference of whole genomes or presence of private SNVs that may been transmitted in the recent past. This global-level source tracking is analogous to admixture estimation with human genotypes [142,143].

We note that source tracking provides insights into similarities between microbiomes and potential transmissions, though the directionality is less conclusive. It is possible that increased contributions in one direction but not the other is suggestive of directionality of transmission. For example, in the case of the mother-infant data from Backhed et al. 2015, FEAST predicted higher contribution from mother to baby than vice versa. This is consistent with work done on crAss-like phage transmissions between mother and infant in the same dataset that showed evidence of directionality by tracking the accumulation of mutations over time that are private to the infant and absent from the mother [144]. But in the case of the ocean, it is possible that over longer time periods, differences in relative contributions from one part of the world to another (e.g. Mediterranean to Red Sea) are more reflective of local selection pressures that permit

certain species and genotypes [141]. Thus, source tracking in certain instances, such as the ocean microbiome, at best reflects the extent of similarity between samples and is less conclusive about directionality.

A popular approach used to track strain transmissions is by detecting high average nucleotide identity (ANI) for species shared between source and sink. For example, inStrain [113] identifies a match between samples for a given species when ANI exceeds 99.999%. However, it is to be noted that inStrain provides distinct and complementary information from FEAST given its binarization of whether or not a strain is shared. For illustration purposes, if an infant harbors 100 species, of which only 1 came from their mother, but that species' strain's relative abundance is 50% of the infant's microbiome, SNV-FEAST would infer that the mother's contribution is 50%, while inStrain would infer that only 1/100th of the infant's species are derived from the mother.

Other studies rely on tracking transmissions of strains with private SNVs shared only between the sink and putative source [110,114,116,118]. The private marker allele tracking approach in Nayfach et al. 2016 provides an improved estimate of true percentage of species that share some portion of their genome with putative sources compared to inStrain (Additional file 1: **Fig. S2, S3**). It is possible that requiring only 5% of marker alleles to be shared rather than a 99.999% ANI permits detection of horizontal gene transfers between lineages residing in mothers and infants [145,146]. However, in FEAST, by using any SNV with an informative distribution across sources as determined by our signature scoring method, we are able to quantify the relative contribution of all the sampled environments and assign a proportion to these putative sources. Another advantage to FEAST is that the contribution of unknown sources can be quantified. For

example, the significant fraction of marine biodiversity estimated to be 'unknown' may be endemic, as previously noted in the Mediterranean [147].

A drawback, however, with using SNVs over species is deeper, whole genome sequencing is required to accurately call SNVs. Moreover, even when there is sufficient coverage, there is still the challenge of a large number of SNVs that make FEAST computationally prohibitive. We demonstrate one way to subset SNVs that uses a scoring method for informativeness, but there may yet be other methods for filtering SNVs to the most informative set. Another potential caveat of SNV filtering is that not all species present will be represented in the final signature SNV set (Additional file 1: **Fig. S4**). Species with higher abundance are more likely to be represented in the signature SNV set. However, we show that not all species need to contribute signature SNVs in order to make accurate inferences, and likewise, not all SNVs are needed to make accurate inferences (Additional file 1: **Fig. S1)**.

Ascertainment of SNVs from metagenomic data in a high-throughput manner, especially common SNVs with microbiome genotyping technology [148], is becoming an increasing priority for the field as metagenomic datasets become more abundant. A genotyper for prokaryotes has already been developed and tested on a catalog of over 100 million SNVs in order to characterize population structure [148]. Such a catalog of informative SNVs could be invaluable for source tracking. With source tracking enabling us to characterize samples by their relationship to known samples, we have a powerful tool to explore samples in new contexts we have yet to discover.

Conclusions

SNV-FEAST is a novel approach to accurately perform source tracking using metagenomic data. By using our algorithm for determining signature SNVs, one can identify relevant SNVs that can be directly provided to FEAST, an existing source tracking approach that can successfully estimate sources using species abundance data. We demonstrate that SNV-FEAST not only accurately quantifies ground truth proportions in simulations but can also recapitulate previous findings in real-world infant datasets. In each test scenario, SNV-FEAST and species-FEAST yield different outcomes, with SNV-FEAST frequently out-performing species-FEAST. Finally, in applying SNV-FEAST to ocean metagenomic data, we uncovered distance-decay relationships between putative sources and sinks. With low computational cost, SNV-FEAST is able to leverage the increasing availability of shotgun metagenomic data to ask fascinating questions about microbiomes in the environment and hosts.

Methods

Data

For simulations and analyses of infant microbiomes in the first year of life , we downloaded the raw shotgun metagenomic sequencing reads from public read archives under accession number PRJEB6456 [118]. We downloaded the raw sequence reads for the NICU analysis from accession number PRJEB323631 [127], and the equivalent for the Tara Oceans analyses were downloaded from accession number PRJEB402 [128]. Data from the HMP Consortium [149] and Lloyd-Price et al [126] was downloaded from the following URL: https://aws.amazon.com/datasets/human-microbiome-project/ [125].

Estimation of species and SNV content of metagenomic samples

We used MIDAS (Metagenomic Intra-Species Diversity Analysis System, version 1.2, downloaded on November 21, 2016 [110] to estimate species abundance and SNV content per species in each metagenomic shotgun sequencing sample. The database we used to apply MIDAS consisted of 31,007 bacterial genomes that are clustered into 5,952 species. The parameters we used to estimate species abundances and SNVs were described in [150]. A species was considered present if there are at least 3 reads mapping to a set of single copy marker genes on average. To call SNVs, we used the default MIDAS settings in order to map reads to a single representative reference genome. The mapping was done with Bowtie 2 [151]: global alignment, MAPID≥94.0%, READQ≥20, ALN_COV≥0.75, and MAPQ≥20, where species with reads mapped to less than 40% of the genome were excluded from the SNV calls. We excluded samples with depth lower than 5 reads, and excluded genetic sites using the default site filters of MIDAS (e.g. ALLELE_FREQ≥0.01, with the exception of SITE_DEPTH which was set to 3.

Application of FEAST algorithm

FEAST, originally introduced by Shenhav et al. [109], is an R-based method that models the mixture proportions for various "source" microbial samples for a given "sink" [109]. This method utilizes expectation maximization to estimate the proportions when given any sort of count-based feature matrix representing the potential sources and sinks. The intuition behind the estimation process is that a source with a similar species distribution to the sink would have a higher contribution estimate to the sink. A species with non-zero counts only in source *j* and the sink would increase the estimated contribution of source *j*. However, in many cases, the same species are found in multiple sources simultaneous. The algorithm does not uniquely assign a species to a source but rather simultaneously utilizes all species information to infer the source

contributions. The method was originally tested and evaluated on species and not on more fine scale genetic data such as SNVs. The number of different species, on average, range in number from a few hundred to a few thousand, while the number of possible nucleotide sites that vary across different sources can number in millions. For this reason, a SNV-filtering process is necessary so that the algorithm can run within a reasonable time and with reasonable memory requirements.

Application of FEAST to the Backhed et al. 2015 dataset

For both species and SNV-FEAST, the same set of sources and sinks were fed into the FEAST algorithm. In the case study of infants in the first year of life [118], the sink consisted of the infant fecal sample at either four days, four months, or 12 months and the potential sources consisted of fecal samples from the true mother, three randomly selected mothers from the same dataset, and also any previous time points for the infant.

Species-FEAST utilized all species present in the infant whereas SNV-FEAST used signature SNVs from the subset of species that had signature SNVs. Shown in Additional file 1: **Fig. S4** are the distribution of species included in species and SNV-FEAST.

Application of FEAST to the Brooks et al. 2017 dataset

For the case study of infants in the NICU [127], the sink consisted of the fecal sample of the infant at a given time point and the potential sources consisted of pooled reads from the touched surfaces, the sink basin and the floor and isolette top from both the infant's own room as well as a different room. The different room was Infant 12's room for Infants 3 and 6, Infants 6's room for Infants 12 and 18.

Application of FEAST to the Sunagawa et al. 2015 dataset

For the Tara Ocean [128] samples, the sink consisted of the surface water sample from the ocean station of interest while the sources consisted of surface water samples from every other station from every other ocean in the world. To study the relationship between source tracking estimates and geographic distance, we analyzed all oceans as either a sink or source against all other possible oceans. To compute geographic distance between stations, we applied the Haversine distance to the longitude and latitude of the sampling sites provided by [128] using the package "geosphere" [152]. Source tracking estimates were computed as described above using either SNV-FEAST or Species FEAST. The regression line for the distance decay analysis was computed using a linear mixed model "contribution ~ distance + (1| sink_ocean)".

Determining the signature SNV set

Signature SNVs were identified as described in the main text. We provide specific steps for determining signature SNVs:

(1) Filter sites: only sites of the genome with at least the required number of reads mapping to the site are considered. In the case study of infants in the first year of life [118] and infants in the NICU [127], the minimum coverage requirement is 10 across the sink and J sources. For the Tara Ocean [128] samples, the minimum coverage is five reads [128]. Additionally, sites that are biallelic must have more than one read mapped to each allele to be considered.

(2) Perform per site per source parameter estimates: for each potential source compute the estimated allele frequency in the sink θ under two different hypotheses:

Hypothesis 1: Source *i* with allele frequency p_i explains the allele counts in the sink.

$$\hat{\theta} = p_i$$

Hypothesis 2: A combination of all other sources except *i* (sources $j \neq i$) explain the observed allele count distribution in the sink. The estimate of the sink allele frequency is computed using a mixture of the allele frequencies p_j from those sources. The mixing parameter α_j is learned using Sequential Least Squares Programming (scipy.minimize()) with the constraint of summing to 1 with bounds of 0 to 1 inclusive: $\sum_{j \neq i} \alpha_j = 1$.

$$\hat{\theta} = \sum_{j \neq i} \alpha_j p_i$$

(3) Compute per site per source log likelihoods: Compute the binomial log-likelihood under hypotheses 1 and 2, given *n* reads with the reference allele and *m* reads with the alternative allele in the sink:

$$l(\hat{\theta}) = n\, log\, \hat{\theta} + m\, log(1 - \hat{\theta})$$

(4) Compute per site per source log likelihood ratio:

$$l_1(\theta) - l_2(\theta)$$

(5) Compute per site summary signature score: The maximum log likelihood ratio per site is the signature score for that SNV, representing how favorably one of the sources explains the sink over all other sources

(6) Filtering of SNVs using signature score: One signature score for that SNV represents how favorably one source explains the sink better than all other sources. All the scores

are ranked across SNVs and SNVs with scores that are greater than two standard deviations over the mean signature score within each 200 kbp window of the genome are retained as signature SNVs. This window size was chosen for to optimize run time and memory requirements.

Note, if only one source passes minimum coverage filtering, $l_2(\theta) = 0$ resulting in a very high likelihood ratio as represented by $l_1(\theta)$ for the one source. These SNVs are more likely to pass the signature score filtering. One exception for SNVs that are included in the signature SNV set without passing signature score filtering are SNVs with an allele that is completely unique to the infant, as these represent SNVs that are potentially derived from an unknown source. Signature SNVs are obtained from the SNV profile of every species for which there is MIDAS output.

Simulating mother to infant transmission

The mixture proportions for 28 simulated infants is shown in **Table S1**. Four possible scenarios are simulated using a combination of either low or high number of sources and low or high transmission probabilities of species. High transmission of species was simulated by drawing separate transmission probabilities for each species in each contributing source based on a beta distribution with a mean equal to the species relative abundance and variance equal to 0.1, a value selected to emulate Backhed et al.'s mean relative abundance and variance. For the low transmission scenario, transmission probabilities were drawn from a beta distribution with mean 0.1 times the relative abundance of that species in the source sample and variance at 0.1. To determine if a species from each source was transmitted to a given infant, a binomial draw was

performed J times, where J = number of sources, and the probability of a mother transmitting the species is p_j based on the beta-drawn transmission probability. If any of the draws yields value 1, that species is transmitted to the infant from all sources. The same simulated data under these scenarios is used for both SNV and species source tracking.

The source tracking estimates are compared to the true mixing proportions using Spearman correlation. The significance of correlation is calculated using the stat_cor function in the 'ggpubr' package [153].

Comparison to inStrain

We ran inStrain [113] on the same synthetic samples as described above. InStrain "profile" [113] and inStrain "compare" [113] were run for every possible infant-source pair. For example, for simulated infant 1 there were 10 putative sources, therefore inStrain compare was run 10 times for each putative source. InStrain reports popANI calculated per scaffold for a given species. To compute a single statistic per species, we computed the average popANI across scaffolds for a given species. The percent infant microbiome species that had strains shared with mother was computed as the number of species in which popANI was >= 99.999% divided by the total number of species with coverage >= 5. PopANI was only calculated in scaffolds that had >=5 coverage in both samples of the pair.

Comparison with strain tracking approach in Nayfach et al. 2016

We applied the strain tracking approach in Nayfach et al. 2016 [110] on the same synthetic communities described above. In Nayfach et al. 2016, strain transmissions are tracked by identifying 'marker alleles' which are private to the infant, mother, or infant-mother dyad, and absent from the broader population. A strain is considered to be shared if at least 5% of all

marker alleles for a mother-infant dyad are shared. Note that the approach for strain tracking proposed in Nayfach et al. 2016 utilizes SNV information outputted by MIDAS, but is not a part of MIDAS.

Each simulated infant had up to 10 sources that were real maternal samples from Backhed et al. 2015 For each possible pair of infants and maternal sources (10 pairings per infant, with 48 infants), we found the set of infant-only marker alleles, mother-only marker alleles, and mother-infant dyad marker alleles. As described in Nayfach et al, 2016, only sites with minimum 30 reads and only alleles that were supported by at least 10% of the total reads aligned to that site were considered. The infant marker allele and mother marker allele were defined as alleles that were present only in the focal sample and absent from the background samples (or below 3 reads = 10% * 30 reads). For the infant, the background consisted of all mothers (including mothers that were used to simulate the infant), real infant samples (excluding infants of mothers used to simulate the infant), and 337 samples of adults from the United States in the HMP (which includes 180 unique adults) that were obtained from the metagenomics repository of HMP under project ID SRP002163 and SRP056641 [126,149]. For the mother, the background consisted of all mother and infant samples in addition to the HMP samples. For computing shared marker alleles, an allele must be present in both the mother and infant but absent from the background, which consisted of all mothers and the HMP samples.

To compute sharing, two quantities were considered: “total sharing”, defined as % shared marker alleles/ (infant marker alleles + mother marker alleles + shared marker alleles) and proportion of infant marker alleles that are shared: % shared marker alleles/ (infant marker alleles + shared marker alleles). The first quantity compared to FEAST estimates was the percentage of infant species in which the “total sharing” was at least 5%. The second quantity

compared to FEAST was the pooled proportion of infant marker alleles that are shared across all species.

Availability of Data and Materials

SNV-FEAST signature SNV selection is implemented in Python and available for pip installation via https://pypi.org/project/Signature-SNVs [154]. It's source code as well as analyses in this paper are available at https://github.com/garudlab/Signature-SNVs [155], licensed under GPL3. The version used in this manuscript is permanently available at https://doi.org/10.5281/zenodo.7515044 [156].

All metagenomic data was obtained from public repositories. The applicable accessions numbers are PRJEB6456 for Backhed et al. 2015 (mother-infant) [118], PRJEB323631 for Brooks et al. 2017 (NICU) [157], PRJEB402 for Sunagawa et al. 2015 (Tara Oceans) [81], and SRP002163 and SRP056641 for HMP [126,149].

Acknowledgements

We thank Nicole Zeltser for processing the Tara Oceans data through the MIDAS pipeline. We thank Richard Wolff for his advice on simulations. We thank Michael Wasney for testing the associated software. We thank members of the Garud lab for their feedback on the manuscript.

CHAPTER 4: Effects of diet, spatial location, and shared environment on microbiome diversity along the mammalian gut

Michael Wasney[1]*, Leah Briscoe[2]*, Ricky Wolff[3], Hans Ghezzi[4], Carolina Tropini[4,5,6], and Nandita R. Garud[1,3]

[1] Department of Human Genetics, University of California Los Angeles, CA
[2] Interdepartmental Program in Bioinformatics, University of California Los Angeles, CA
[3] Department of Ecology and Evolutionary Biology, University of California Los Angeles
[4] Department of Microbiology and Immunology, University of British Columbia, Vancouver, Canada
[5] School of Biomedical Engineering, University of British Columbia, Vancouver, Canada
[6] Humans and the Microbiome Program, Canadian Institute for Advanced Research, Toronto, Canada.
*co-first authors
Correspondence: ngarud@ucla.edu

Keywords: spatial; microbiome diversity

Abstract

The species and genetic diversity of the human gut microbiome has been extensively quantified from stool and associated with important host phenotypes. However, it remains unclear whether diversity measured from stool reflects the spatial heterogeneity of the microbiome along the tract of the gut. Here, we quantify species, strain, and gene diversity along the tract of the gut from 6 humanized, genetically identical mice gavaged with feces from the same healthy human donor, and then maintained on either a standard rodent diet or a fiber-rich diet. We found that species composition differs substantially between the upper and lower gut as well as between diet regimes. By contrast, strain composition is often, but not always uniform across locations along the gut. When two or more strains of the same species colonized the gut, the strains could be found at roughly similar frequencies in the majority of cases, but occasionally strain frequencies

varied dramatically along the gut. After controlling for strain structure, gene content still differed predictably based on spatial location within the gut. Finally, although the mice were provided the same inoculum strain composition varied considerably between mice, indicating stochasticity in colonization success. Cohoused mice were generally found to have much more similar strain compositions except in some cases, illustrating the effect of a shared environment. To understand if humans display similar patterns, we investigated strain occupancy and genetic diversity in a human cohort sampled along the gut and found that strain frequencies are also relatively constant along the gut. In sum, we show that diversity in the gut microbiome is shaped by diet, gut region, and co-housing.

Introduction

The human gut is a complex environment colonized by hundreds to thousands of microbial species. The composition of the gut microbiome has been associated with numerous phenotypes including intestinal diseases [158–161] and extraintestinal diseases including autoimmune disease [160,162], cardiometabolic disease [163,164], liver disease [165] and many others [166].

Typical studies on the gut microbiome focus on data collected from fecal matter and not along the gastrointestinal tract directly. However, the gut microbiome is not spatially homogeneous: many studies have found that microbiome species composition varies considerably along the tract of the gut [167–171]. This spatial heterogeneity has been shown to be widespread in other animals such as mice [172], wild rodents [173], hummingbirds [174], zebrafish [175], pigs [176] and rhesus macaques [176].

While stool has been used as a proxy for the species that exist in the gut at large, there are a growing number of studies that indicate that the spatial organization of species along the gut has important health implications [158,171,177–180]. This between-site variation, which is functionally and pathogenically relevant, is potentially lost by examining the stool microbiome alone [158,169,170,173,174,181–183].

While many studies have quantified community composition between segments of the gut at the species level, there is potentially important heterogeneity at the strain, gene, and single nucleotide level along the gut that remains to discovered [116,181,184]. For example, Yang et al. 2022 found genetic adaptations that arise in one region of the gut, enabling bacterial translocation to liver and inducing inflammation [184]. Shalon et al. 2023 found that carbohydrate active enzyme and antimicrobial resistance gene abundances can differ along the gut [185]. Montassier et. al. 2021 found that antibiotics increase the number of antimicrobial resistance genes in the lower gut tract whereas probiotics can reduce the number of such genes [170].

There are a number of questions about the sub-species spatial organization of gut microbiota that have not yet been thoroughly investigated. First, how do strain frequencies compare across locations along the gut and across hosts? Second, are these strain frequencies generic across species? Finally, controlling for strain structure, does gene content vary spatially and with diet?

To effectively evaluate sub-species variation along the gut, direct sampling of the microbiome along the gastrointestinal tract is important. Due to the difficulty of obtaining and sequencing samples collected along the human gut, humanized mice serve as a useful model of the human

gut due to the ability to control for extraneous variables while testing bacteria transplanted from human fecal samples [172,186]. Mice and their microbial communities respond to high-fat, high-sugar diets in much the same way humans do [172], and they additionally show different changes in species composition along the gut with respect to diet [172]. Validating findings in a murine system with what can be collected from humans allows us to understand the generalizability of our findings.

In our study, we evaluate how spatial location, co-housing, diet and drug treatments affect species, gene, and nucleotide differences along the mammalian gut. We first study whole metagenome shotgun data collected from five different gut regions of six humanized mice on a standard murine diet or fiber-rich diet consisting of guar gum. Guar gum is a galactomannan polysaccharide that is not digestible by the mammalian host [187,188](Fettig et al. 2022, Ohashi et al. 2015) and will pass into the large intestine where it can be readily fermented by bacteria, which then decreases pH. We employ an array of analyses to understand how location in the gut, co-housing, and diet drives differences. We then evaluate species and strain diversity using previously published data collected from endoscopy and stool collection in humans.

We find that species composition differs drastically between the upper and lower gut as well as between diet types. Within a host, the same strains are often colonizing different regions of the gut and usually at similar frequencies at the regions in which the species is detected. However, we detect similar spatial organization patterns of *B. wexlerae* strains in three of our mice, indicating that some species can display replicable strain-level spatial organization along the gut. Despite being provided with the same inoculum of strains and, in some cases, we found that

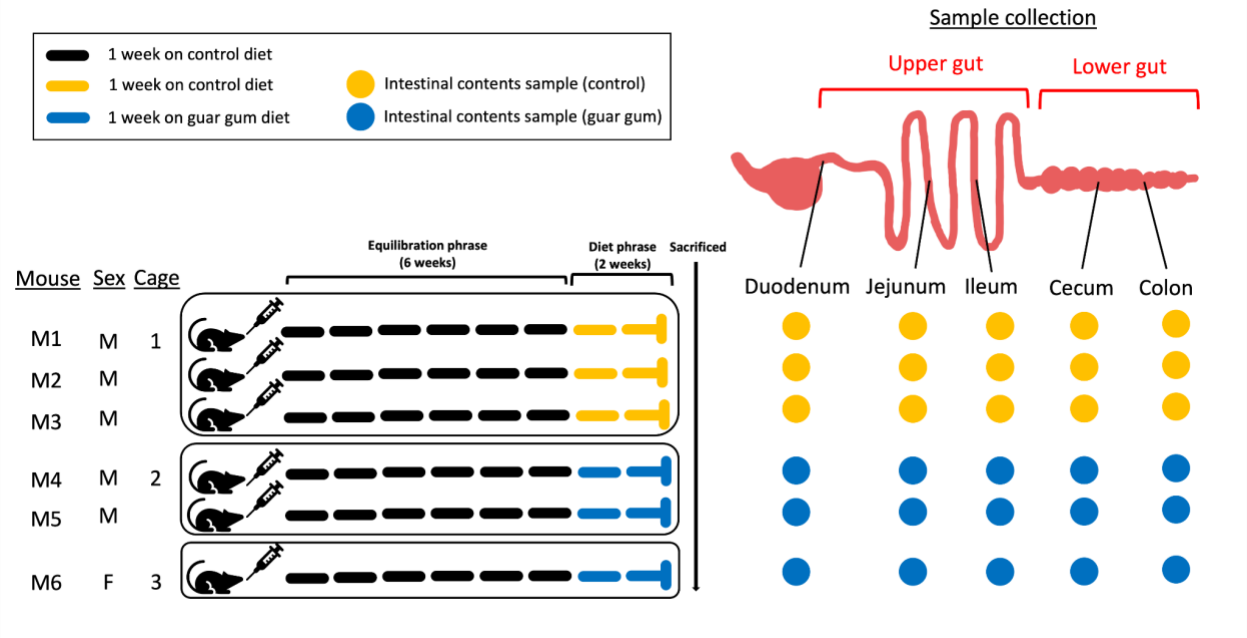
hosts that weren't cohoused were more likely to harbor different strains.. Even when strains are shared, they may be present at different frequencies. Despite the same strain backbone appearing throughout the gut of a host, gene content differed along the upper versus lower gut axis. We quantitatively show that the differing environmental conditions within the mammalian gut shapes variation of gut microbiomes.

Results

Data

To understand the effect of spatial location and diet on diversity along the gut, we analyzed six gnotobiotic mice that were orally gavaged with the same human stool sample (**Figure 1**). Mice were equilibrated for six weeks on a standard rodent diet and then divided into two treatment groups: one group (mice 1, 2, and 3) continued on the same diet, and a second group (mice 4, 5, and 6) was placed on a diet with 30% guar gum, both for two weeks (Ng et al. 2022). The mice on each diet were sacrificed and shotgun metagenomic sequencing was performed on five intestinal segments: duodenum, jejunum, ileum, cecum, and colon (**Methods, Table S1**). In our analyses, we classify duodenum, jejunum, and ileum as regions of the upper gut, and cecum, and

colon as regions of the lower gut.



11 **Figure 1. Schematic of humanized mouse experimental design.** Six germ-free Swiss Webster mice humanized over a period of 8 weeks. In the last 2 weeks, half the mice were switched to a guar gum diet.

Species diversity differs along the tract of the gut and in different diet regimes

The gut environment can vary considerably in terms of pH and osmolality (**Figure S1**), thereby having a significant impact on species relative abundances (Ng et al. 2022). As shown previously, guar gum diets significantly decrease pH in the cecum and colon (Wilcoxon rank sum test, p-value = 0.00033), and in the upper gut, pH is decreased to a lesser extent (Wilcoxon rank sum test, p-value = 0.042) (**Figure S1**) (Ng. et al. 2022). Microbial fermentation of complex carbs in the cecum and colon is known to produce short chain fatty acids which acidify the gut [189]. As a potential explanation for the lower pH in the lower regions of the gut (cecum and colon), Ng et al. 2022 found that mice on guar gum had a three fold increase in levels of

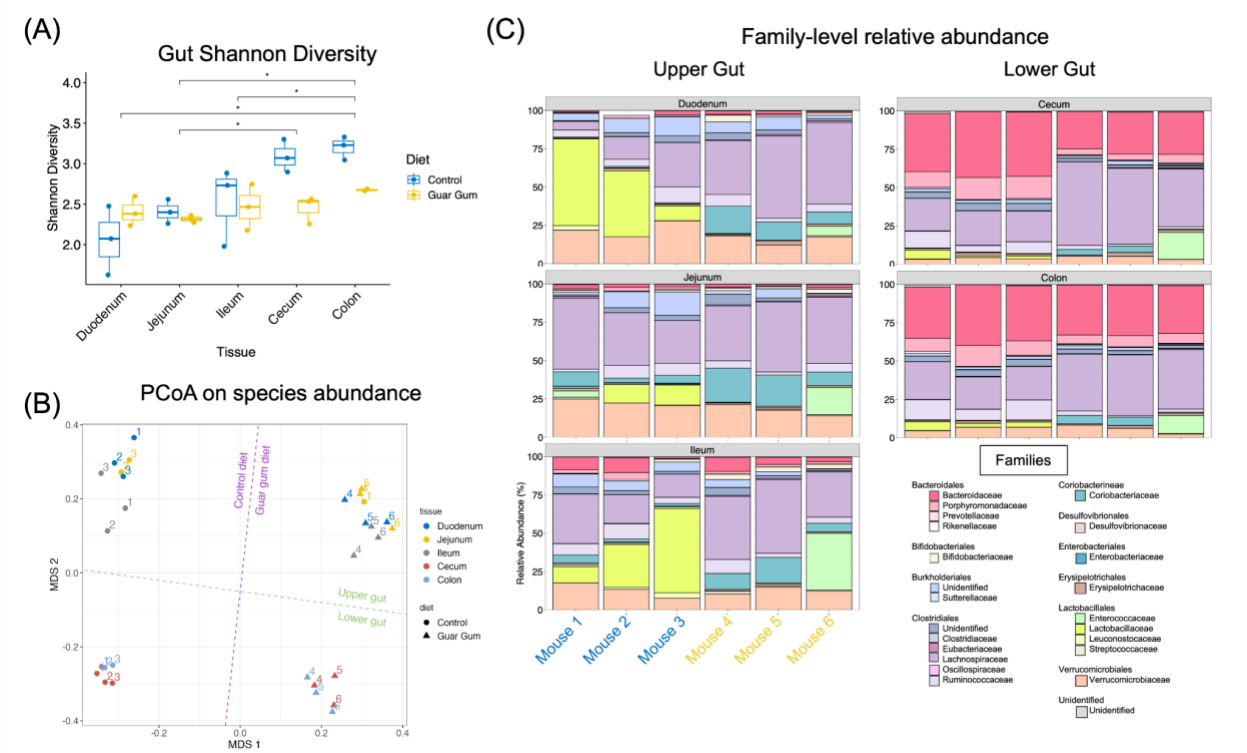
butyrate, a short chain fatty acid [189]. With 16S rRNA sequencing, Ng et al. 2022 detected family-level shifts in the composition of microbiota in the colon, suggesting that microbial communities can differ spatially within the gut and across different environmental regimes (e.g., diet).

Here we assessed the impact of diet and gut region on the species abundances ascertained from shotgun sequencing data instead of 16S data (**Figure 2A**). Diet is a major driver of species diversity, with guar gum mice displaying lower alpha diversity compared to control mice across all regions of the gut (paired Wilcoxon signed rank test, p-value = 0.036). Gut region is also a major driver of species of diversity, with the colon displaying significantly more diversity than all regions in the upper GI tract (paired Wilcoxon signed rank test, p-value = 0.016, 0.016, 0.031 for duodenum, jejunum and ileum respectively), irrespective of diet. We found similarity alpha diversity across the upper gut.

A principal coordinate analysis (PCoA) using beta diversity (Bray-Curtis dissimilarity index) computed on the species relative abundances in all samples further revealed the impact of diet and gut region on species diversity (**Figure 2B**). Control mice 1, 2, and 3, separate from guar gum-fed mice 4, 5, and 6 along axis 1 of the PCoA plot, indicating that diet is a key driver of species composition. This is consistent with findings from Ng et al. 2022 that diet drives changes in taxa composition primarily by altering pH and osmolality within the gut. While these results could in principle be due to a cage effect since mice 1, 2, and 3 were cohoused, this is unlikely because guar gum-fed mice 4 and 5 were in a different cage from mouse 6 and yet clustered on the PCoA plot (**Figure 2B,** Bray-Curtis values shown in **Figure S2A**). Additionally, samples separate along axis 2 of the PCoA plot by gut location (i.e., upper gut or lower gut) irrespective

of diet, indicating that gut region is also a major driver of species composition (**Figure 2B,** Bray-Curtis values shown in **Figure S2C-D**).

Diet and tissue community composition changes along the gut are also apparent at the family level (**Figure 2C**). For example, Verrucomicrobiaceae is enriched in the upper gut and Bacteroidaceae in the lower gut. This latter finding is consistent with Ng et al. 2022 findings with 16S data, where the expansion of the Bacteroidaceae family in the cecum and colon was thought to be the result of more acid-tolerant Bacteroides in the same niche. Additionally, like in Ng et al. 2022, guar gum mice experienced an increase in Lachnospiraceae in relative to control mice, which may reflect increase in butyrate producer Blautia as a result of the fiber rich diet. Unlike the Ng et al. 2022 16S findings, however, we did not observe a loss of the Erysipelotrichaeceae family on the guar gum diet. Additionally, the fiber-rich guar gum diet seems to suppress members of the Lactobacillaceae, resulting in Lactobacillaceae being at lower abundance in mice 4, 5, and 6. Together, our results show that diet and tissue drive broad-scale shifts in community composition that are evident at various taxonomic levels.



12 **Figure 2. Species diversity along the gut (A)** Shannon diversity estimates in different gut regions and diet regimes (two-sided Wilcoxon rank sum test between tissues using all mice across both diet groups, * for p-value < 0.05, ** for p-value < 0.01) **(B)** PCoA using species beta diversity shows that samples cluster by gut region and diet. Beta diversity was calculated using the Bray-curtis dissimilarity index between relative species abundance of all samples. Samples are colored by diet and gut region. Each point is labeled with the corresponding host ID (1-6). **(C)** Relative abundance of taxa at the family-level at the five gut regions.

Nucleotide diversity along the gut within versus between hosts

We next asked whether sub-species diversity at the single nucleotide level also varies spatially along the gut. To do so, we calculated nucleotide diversity (pi) within and between mice (**Methods**) for each of the 36 most prevalent and abundant species (**Figure 3**).

First, to understand diversity within the guts of individual mice, we computed diversity within gut regions and between gut regions for each species observed in a host. Within a single gut

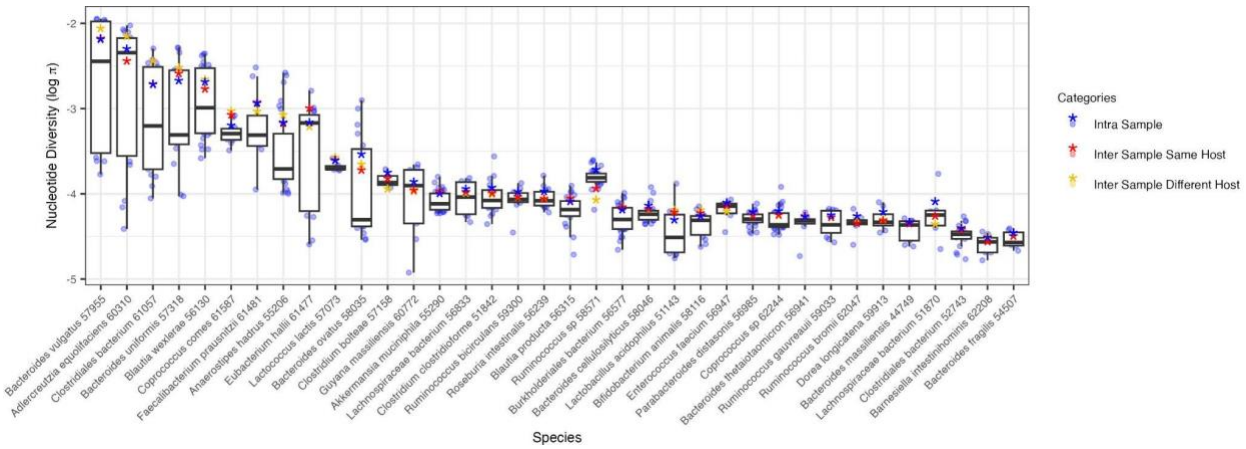
region in a given host, distributions of nucleotide diversity ranged from as small as 1.2 x 10^{-5} /bp in *Guyana massiliensis* (mouse 5's cecum) to as large as 1.1 x 10^{-2} /bp in *Bacteroides vulgatus* (mouse 1's colon).

As previously argued [150] high nucleotide diversity values (>1 x 10^{-3}/bp) are inconsistent with a single strain diversifying within a host over the course of 8 weeks given conservatively high mutation rates (μ~1 x 10^{-9})[190] and generation time estimates (~10 generations/day) [191]. Instead, these high diversity values are most consistent with multiple, genetically distinct strains co-colonizing the host. Previous work has shown that human commensal gut bacteria are well-described by an oligo-colonization model in which some small number (~1 x 10^{-4}) of strains of the same species co-exist [150,192].

Next, to understand whether diversity within a mouse for a given region of the gut resembles overall diversity in the broader community, we computed diversity between pairs of mice. Between host distributions of pairwise nucleotide diversity showed a large range from 1.1 x 10^{-5}/bp in *Clostridiales bacterium* to 1.3 x 10^{-2}/bp in *Adlercreutzia equolifaciens*. Similar to our conclusion that within-host pi > 1 x 10^{-3}/bp is inconsistent with diversity levels arising from a single strain, we reasoned that pairs of samples from different hosts with pi>10^-3/bp is also indicative of multiple strains present across the mouse cohort. We conclude that five of the 36 species—*B. vulgatus*, *A. equilofaciens*, *C. bacterium*, *B. uniformis*, and *B. wexlerae*—have patterns consistent with multiple strains across the mouse cohort. This also implies the presence of multiple strains of these species in the original inoculum.

We hypothesized that mice were likely to be colonized by multiple strains when exposed to multiple strains, which would be reflected by high within-host nucleotide diversity. The average within- and between-host pi values were highly positively correlated (spearman correlation = 0.98), such that species with higher between-host pi values tended to have higher within-host pairwise pi values. The five samples with the highest mean between-host pairwise pi values greater than 1×10^{-3} also had mean within-host pairwise pi values greater than 1×10^{-3}, supporting the idea that oligocolonization is common in the mouse gut when exposure to multiple strains occurs.

We next asked whether pairs of locations along the gut display similar distributions of nucleotide diversity compared to a single location. Indeed, species-level averages of within-host pairwise pi values were highly positively correlated with the corresponding average pi values within a gut segment (spearman correlation = 0.97). All species with mean within-host pairwise pi > 1×10^{-3} also had mean within gut region pi> 1×10^{-3}. This indicated that multiple strains were often able to colonize and persist in the same gut segment. The species with the highest average within host pi was *B. vulgatus* (pi across gut regions = 5.6×10^{-3} ,pi within gut regions = 5.5×10^{-3}), likely reflecting the presence of multiple strains in most hosts and gut regions. Whereas the species with lowest within host pi was *B. intestinihominis* (pi across gut regions = 2.3×10^{-5} ,pi within gut regions = 2.6×10^{-5}), likely reflecting monocolonization across most hosts and gut regions. Importantly, this finding does not exclude the possibility that these strains display finer grain spatial segregation that our sampling could not detect.



13 **Figure 3. Nucleotide diversity within and between pairs of samples.** Nucleotide diversity (pi) computed for 36 species within each tissue for each mouse (see Methods for selection criteria for these species). Inter-sample pi was computed between pairs of samples either from the same host or different host. Asterisks for intra sample calculations represent the mean sample-specific pi, whereas asterisks for inter sample comparisons represent the mean of pi values for multiple intersample calculations. Species appear in descending order according to their between-host pairwise pi.

Strain sharing across gut segments and across hosts given the same inoculum

Given that mice were exposed to the same inoculum, we next hypothesized that different mice would be colonized by the same strains. To test this hypothesis, we estimated rates of strain sharing within hosts between gut regions and between hosts using population average nucleotide identity (popANI) [113]. PopANI quantifies the proportion of fixed differences between populations, with values below 99.999% indicative of at least one strain not shared between two populations [113].

PopANI was calculated for all species between all pairs of samples (**Methods**). We examined the distribution of popANI for the five species with mean pairwise pi values exceeding 1 x 10^{-3}, as these would allow us to explore the dynamics of strain colonization when mice were potentially

exposed to multiple strains. As a control, we also examined the five sufficiently high coverage species (see **Methods**) with the lowest mean inter-host pi values, which we assumed had only one strain across the mouse cohort and would therefore have much simpler colonization dynamics (**Figure S3**).

Inter-segment popANI

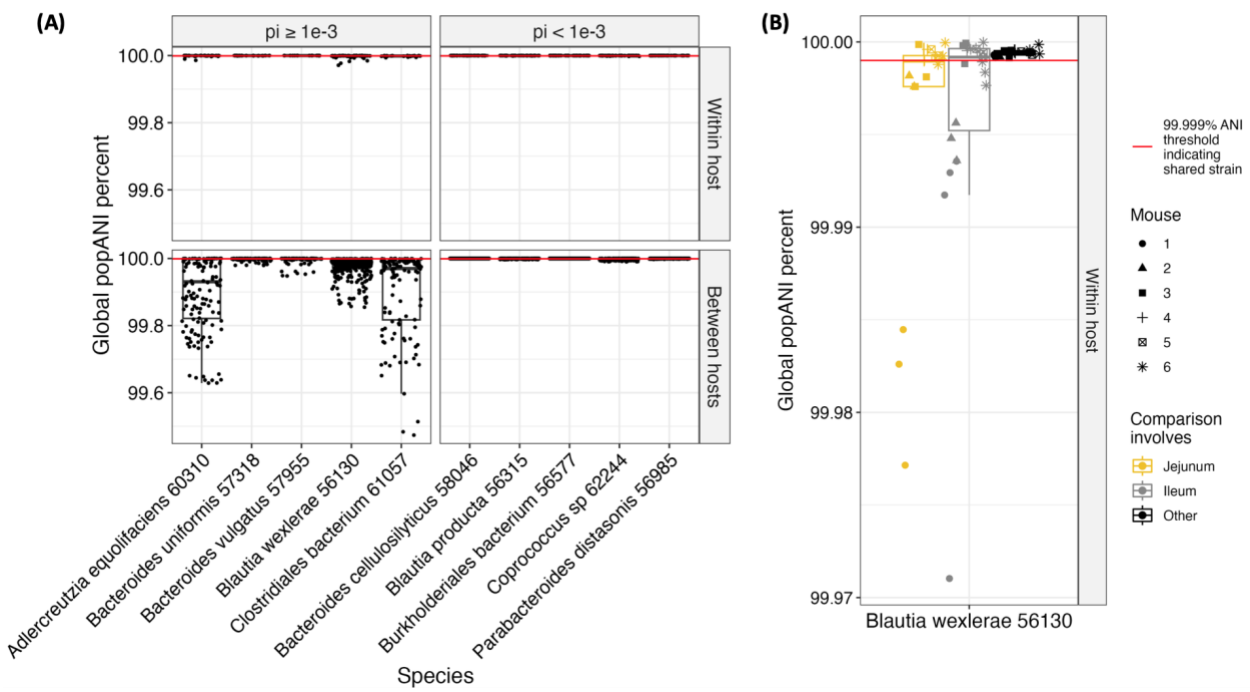
Species with low mean inter-sample pi almost always had popANI values above the 99.999% strain sharing threshold both between segments along the gut and between mice, indicating that all samples for these species shared the same strain. Species with high average pi between segments pi also overwhelmingly had popANI values close to or above the 99.999% strain sharing threshold between segments along the gut (**Figure 4A**). However, unlike the low diversity species, a few instances were observed in which popANI values were below the 99.999% cutoff, indicative of different strain occupancies along the gut.

B. wexlerae shows an excess of intra-host comparisons yielding low popANI values, indicating that the strains of this species could exhibit spatial partitioning along the gut. We found that for this species in particular, within-host comparisons involving the ilea and jejuna of mouse 1 and 2 (and ilea of mouse 3 and 6) had disproportionately low popANI values (**Figure 4B**). This indicates that ileal and jejunal samples harbor strains not present elsewhere in the control mice (mice 1, 2, and 3). *B. wexlerae* is thus an example of a species that displays strain-level spatial variation along the gut. Notably, *B. wexlerae* species is found at relatively high abundances throughout the gut, while the other species specialize in the upper or lower gut, such as *B. uniformis* and *B. vulgatus* which are primarily in lower gut respectively (**Figure S4**).

Inter-host popANI

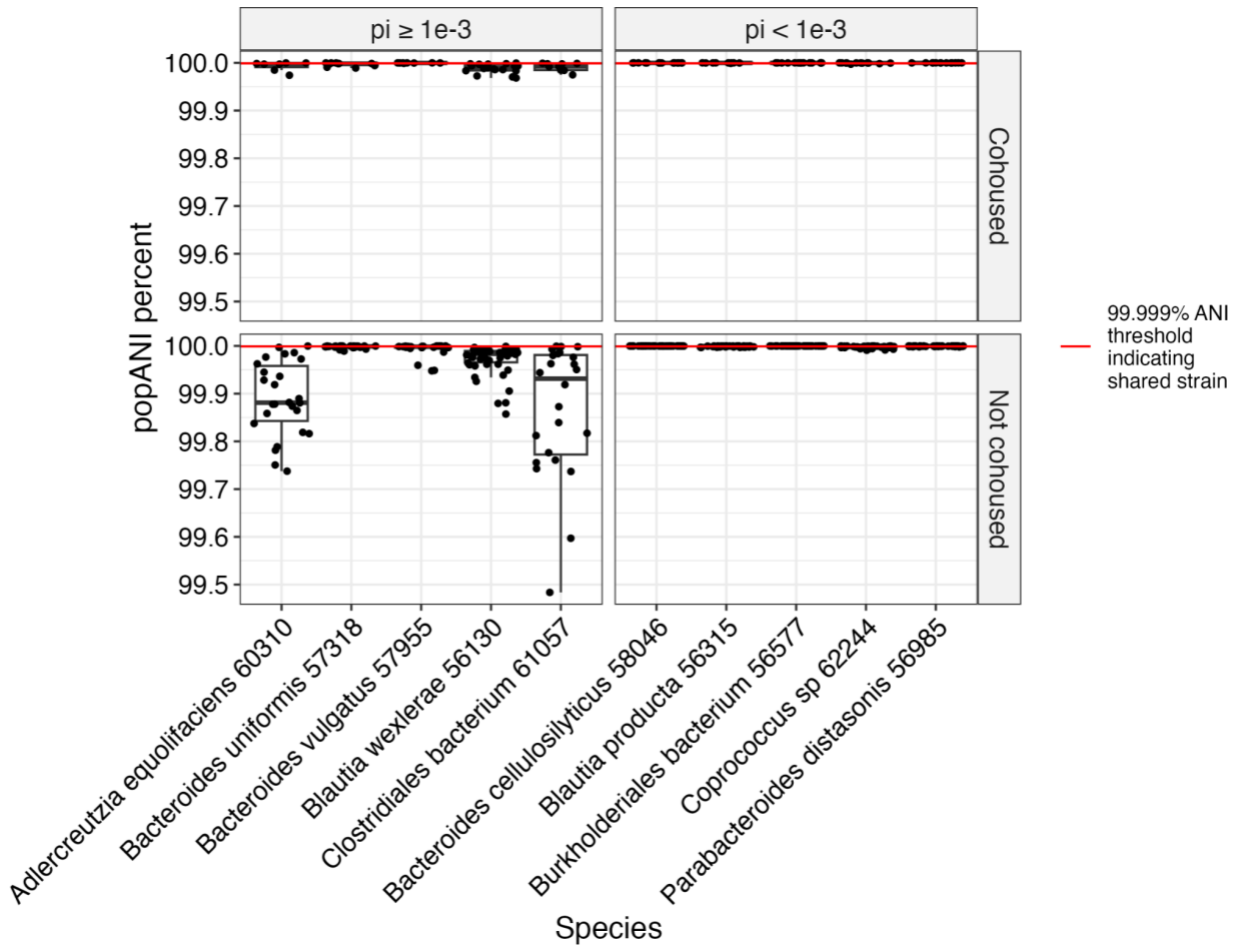
Species with high mean inter-segment pi have distributions of popANI values that extend well below the strain sharing threshold, unlike species with low mean inter-segment pi values. This points to comparatively lower rates of strain sharing between hosts (**Figure 4A**). This indicates that mice are not necessarily colonized by the same strains, even when exposed to an identical inoculum.

Because several mice were co-housed, we asked whether the lower inter-host popANI values were primarily driven by pairs of mice in different cages. Coprophagy has been shown to facilitate transmission of gut microbiota between cohoused mice (citation). This led us to hypothesize that mice in the same cage would share all strains, whereas those not cohoused together would exhibit drastically lowered rates of strain sharing. We plotted popANI values for the same tissue in different mice either co-housed or not co-housed (**Figure 5**). Predictably, the rate of strain sharing was unaffected by cohousing (or diet) among species with a single strain in the metapopulation. However, certain tissue-matched samples from cohoused mice did have popANI values below 99.999% indicating that mice in the same cage did not necessarily harbor the same strains. Mice that were housed separated tended to have comparatively lower rates of strain sharing than those housed together. Given the limits of our dataset, however, it's challenging to disentangle the effects of differential diet from cohousing: all mice except for mouse 6 were cohoused with the other mice on the same diet.



14 **Figure 4. PopANI values within versus between hosts. (A)** Results are partitioned based on mean inter-sample pi values, whereby >10^-3/bp are more consistent with multiple strains present in the innoculum and <10^-3 is more consistent with a single strain present. PopANI was computed between pairs of samples from different gut regions from the same host ("Within host") or between different hosts ("Between hosts"). The 99.999% strain sharing threshold is marked in red. Points below this threshold indicate that at least one strain is not shared between the two populations being compared. **(B)** Within-host popANI values are plotted for *B. wexlerae*. An excess of comparisons involving jejunum (blue) and ileum (yellow) in mice 1 and 2 have popANI values below 99.999%, indicating the presence of a strain not shared elsewhere in the gut.

B. producta, *Coprococcus* sp. 62244, and *C. bacterium*—3 of the 5 species with mean inter-sample pi < 1 x 10^{-3}—exhibit between-host comparison popANI values below 99.999%, although these popANI values are much higher than those observed in species with mean inter-sample pi ≥ 1 x 10^{-3} (**Figure S5**). Two of these species (*B. producta*, *Coprococcus* sp. 62244) also exhibited within-host comparison popANI values below 99.999% in mouse 1. This could represent noise introduced by low coverage (**Figure S4**) or that these species do in fact have multiple strains despite having low inter-sample pi values.



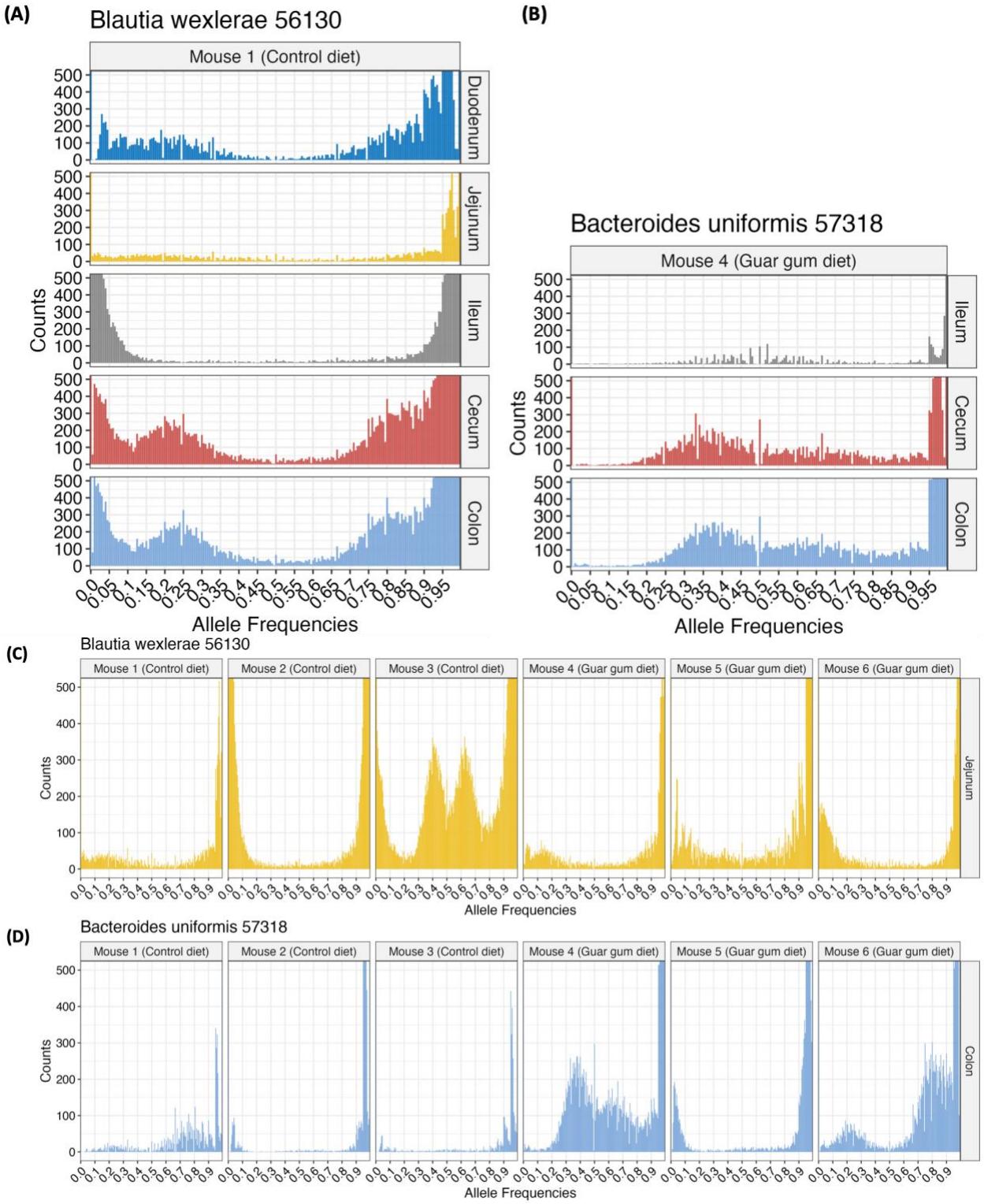
15 **Figure 5**. PopANI between the same gut segments in different hosts either co-housed or not co-housed PopANI between mice for species with either multiple strains (left) or a single strain (right) in the population, further separated by whether the two mice being compared were cohoused or not. The 99.999% strain sharing threshold is marked in red, with points below this line signifying that at least one strain is not shared between the two samples being compared.

Strains colonize distinct locations of the gut at similar frequencies

Our results indicated that strains successful in colonizing a host are present throughout that host's small and large intestines. However, previous work has shown that environmental gradients exist along the gut [189](**Figure S1**). Thus, we hypothesized that strains may exist at

different frequencies along the gut, reflecting either stochastic processes or their differential fitness in the context of distinct ecological niches found along the tract of the gut.

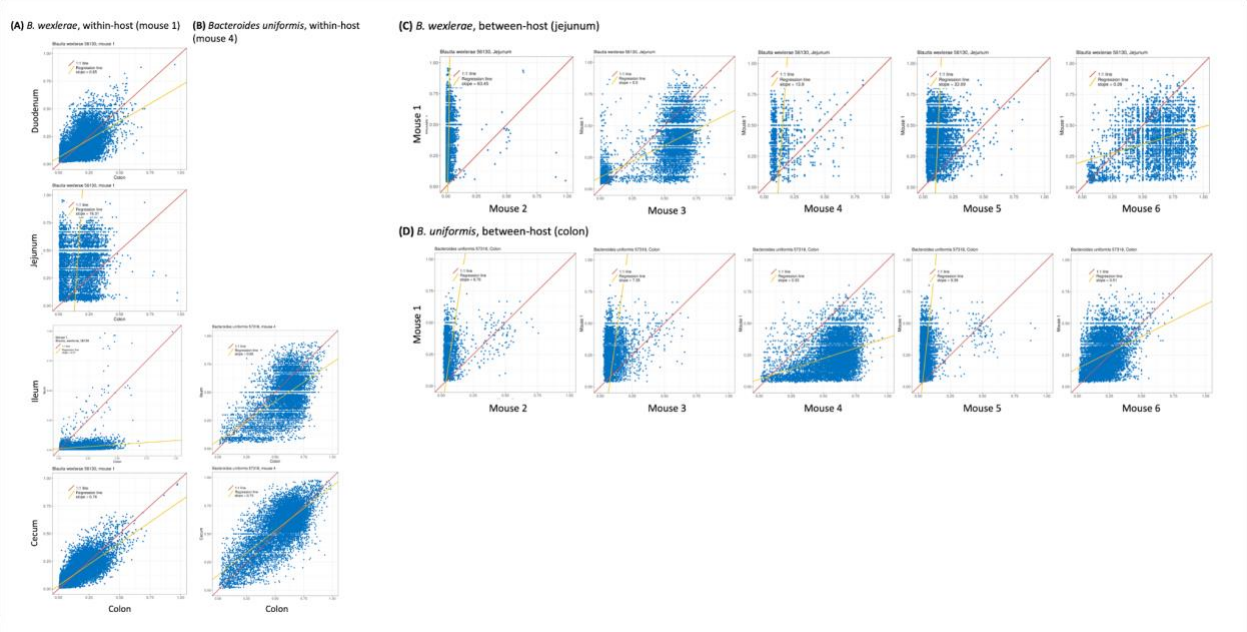
To ascertain whether strains are colonizing the gut at similar or different frequencies , we plotted site frequency spectra (SFS) for *B. wexlerae* in all samples (see **Methods**). An SFS represents a distribution of allele frequencies. An enrichment of loci with alleles at intermediate frequency alleles is inconsistent with diversity arising from a single strain present in the gut, and instead is consistent with the presence of multiple strains, as these represent variant sites that are fixed between the strains [150,192]. Assuming that samples carry the same mixture of strains, differing distributions of allele frequencies can indicate varying strain abundances across samples.We focused on *B. wexlerae* and *B. uniformis*, giving us the opportunity to observe strain dynamics in species with high (*B. wexlerae*) and low (*B. uniformis*) levels of subspecies spatial structure, as inferred from popANI results.



16 **Figure 6. Site frequency spectra for *B. wexlerae* and *B. uniformis*.** **(A)** SFS for *B. wexlerae* in all samples in mouse 1. **(B)** SFS for *B. uniformis* in all samples in Mouse 4. **(C)** SFS for *B. wexlerae* in jejunal samples from all mice (i.e., between). **(D)** SFS for *B. uniformis* in colonic samples from all mice. SFS samples are colored by tissue type.

Qualitatively, allele frequency distributions for *B. wexlerae* and *B. uniformis* are more similar within mice 1 and 4 (**Figure 6A-B**) than they are within the same tissue across mice (**Figure 6C-D**). Note that *B. uniformis* was not found at high abundances in the upper gut (particularly in the duodenum and colon), which prevented high-fidelity SFS from being plotted for those samples. Despite how relatively similar within-host allele frequency distributions could be, the strain-level spatial variation we identified in mouse 1 (**Figure 4B**) is evident in the within-host SFS shown here (**Figure 6A**). Generally, many samples for both species harbor multiple strains simultaneously, as evidenced by the frequent enrichment in intermediate alleles.

To quantify the similarity between strain abundances both within and between mice, we plotted 2D SFS for *B. wexlerae* and *B. uniformis* (see methods). 2D SFS show the corresponding allele frequencies for loci in two samples on the x and y axes, respectively. An enrichment of loci at the same or similar allele frequencies indicates that the same mixture of strains exist in both samples at similar abundances. For each 2D SFS, we fit an OLS linear regression between the allele frequencies in the samples being compared (see **Methods**). Regression slopes closer to 1 provide strong indication for similar strain abundances between samples.



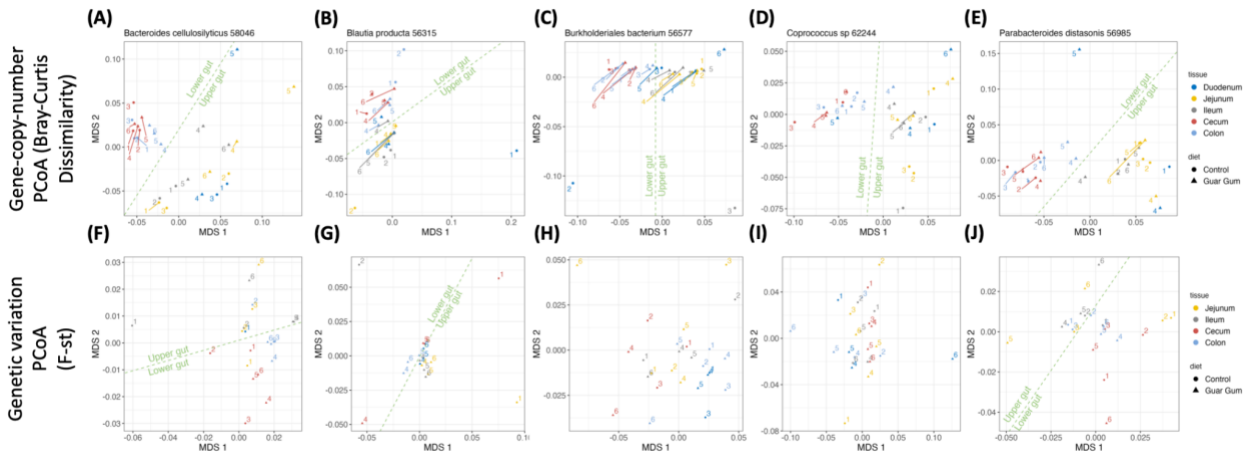
17 **Figure 7. Comparison of allele frequency distributions between pairs of samples within and across mice.** Each plot compares allele frequency distributions of *B. wexlerae* subpopulations in pairs of samples, at select loci (see Methods). **(A)** Pairwise comparisons of *B. wexlerae* allele frequencies between colon versus other tissues in mouse 1 (control diet). **(B)** Pairwise comparisons of *B. uniformis* allele frequencies between colon versus other tissues in mouse 4 (guar gum diet). **(C)** Pairwise comparisons of *B. wexlerae* allele frequencies in jejunal samples of mouse 1 versus other mice. **(D)** Pairwise comparisons of *B. uniformis* allele frequencies in colonic samples of mouse 1 versus other mice.

We hypothesized that allele frequencies would be at similar abundances within hosts, and between hosts that were cohoused. Allele frequencies between samples tended to be more similar within hosts than between them for both *B. wexlerae* and *B. uniformis*, suggesting that strain abundances are relatively more similar within hosts than between them (**Figure 7**). Nonetheless, the strain differentiation between jejunum and ileum in mouse 1 that was detected using popANI is evident here from the incongruent allele frequencies distributions between colon and the two upper gut samples (**Figure 7A**), supporting the notion that ileum and jejunum do not share one or more *B. wexlerae* strains present elsewhere in mouse 1. All three mouse 4 samples in which *B. uniformis* was present with sufficiently high coverage had highly concordant allele frequency distributions, indicating that *B. uniformis* was present at similar abundances throughout the lower

gut of mouse 4 (**Figure 7B**). Contrastingly, allele frequencies were less concordant between samples taken from different mice for both *B. wexlerae* and *B. uniformis* (**Figure 7C-D**), with none reaching the levels of agreement observed between *B. uniformis* subpopulations in mouse 4. Allele frequencies were not more similar between cohoused mice than those housed separately. This suggests that although hosts sharing an environment are likely to share similar strains, the strains are not necessarily present at the same abundances between hosts.

Host diet is a stronger driver of gene-copy-number differentiation between species in different samples than gut location or individual

Given that nucleotide variation tends to not vary significantly along the tract of the gut (**Figure S6**), we hypothesized that that gene content would also be relatively constant along the gut. To control for the effect of strain structure, which could significantly drive gene copy number differences, we limited our analysis to the five aforementioned species with likely a single strain present (or, in other words, had pi < 1 x 10^{-3} /bp). Gene copy numbers were calculated for all genes in the pangenome of these species (see methods). For each species, PCoA was performed using Bray-Curtis dissimilarity indices calculated from these gene copy numbers (**Figure 8A-E,** Bray Curtis values shown in **Figure S7**). PCoA was also performed using genetic distance as represented by F-st values (see **Methods** for Fst calculation, **Figure 8F-J**), another pairwise metric that measures the degree of genetic differentiation between two samples (see **Methods**). Because we selected five species with no strain structure, we hypothesized that genetic variants would not exhibit spatial organization along the gut or between mice.



18 **Figure 8. PCoA on gene copy number and genetic variation** Species-level PCoA was performed on samples using gene copy numbers (A-E) and genetic differentiation (F-J). PCoA plots are shown for **(A, F)** *B. cellulosilyticus*, **(B, G)** *B. producta*, **(C, H)** *B. bacterium*, **(D, I)** *Coporococcus* (sp. 62244) , and **(E, J)** *P. distasonis* all of which are assumed to have a single strain across mice.

Counter to our expectation, gene content did vary along the tract of the gut for all five of the species examined. Specifically, samples clustered based on whether they were from the upper or lower gut, respectively. While it is unclear if these differences in gene copy number have implications for community function, they do suggest that gene duplication and loss events may provide a mechanism for bacterial populations to adapt to the different environmental conditions found in the mammalian gut. We also used genetic distance computed using the fixation index Fst, to perform PCoA on the samples. Three of the five species (*B. cellolosilyticus*, *B. producta*, and *P. distasonis*) exhibited a spatially organized distribution of genetic variation along the gut, with populations of species within a segment of the gut (upper or lower) being more similar to each populations from a different segment. This demonstrates that both genetic variation and gene content can vary spatially along the gut, but that gene content can also vary independent of there being genetic variation. Notably, we rarely detected differences in gene copy number driven by diet/cohousing. When using inter-sample pi for PCoA, *Coprococcus* (sp. 62244)

clustered subtly by cohousing group, suggesting that genetic differentiation could be occurring as a result of physical separation of mouse populations or differing diet regimes.

Source tracking

We sought to quantify sharing across tissues and mice on using inference on the collective set of species and SNVs. Source tracking has been previously used to discern the influence of family members on a host's microbiome [109,193], or the environment on the host's microbiome [127,194]. If we considered the colon a sink for microbes in other parts of the gut, we asked which tissues are the highest contributors to the colon. We applied source tracking with FEAST [109] using two approaches, either species abundance as input or a set of signature SNVs (see **Methods**, [193]). We observe that the species content of tissues mostly resemble the nearest tissue in the gut, particularly the tissue that preceded that tissue within the lower and upper gut (**Figure S8**). For example, jejunum most resembled duodenum, and ileum most resembled jejunum. Cecum and Colon most resembled one another. Using SNVs, we find that multiple neighboring tissues contribute to the tissue of interest with the strongest contributors being the nearest tissues (**Figure S8**).

We next asked how much mice resembled one another. Indeed, mice that cohabitate show the highest source tracking estimates (**Figure S9**). Interestingly, SNVs are able to trace higher estimates of sharing across mice of the same diet but different cages (**Figure S9**).

Discussion

Studies in a wide range of host species have been informative in understanding the gut microbiome of mice and humans. For example, Kokou et al. 2019 found habitat filtering in the gut of seabass was primarily by gut segment and less by diet [198]. Studies in honeybees [199] and plant rhizospheres [200] give further evidence that spatial organization along the "gut" may likely be more a norm than an exception. In a study of humanized mice, Turnbaugh et al. 2009 found that mice were stably colonized along the gut by the microbes from a human donor, and also that diet could rapidly alter the spatial structure of the microbiome [172]. Our observation that the regions of the upper GI tract in mice had similar diversity levels to each other has also been demonstrated canines and emu where the duodenum, jejunum and ileum on average have lower diversity than the cecum or colon [201,202].

Although many of these studies do not investigate composition at a strains level, strain dynamics is likely playing some sort of role in the initial colonization and stability of the microbiome. Vega et al. 2017 found that the heterogeneity in strain colonization is primarily stochastic where bistability was observed in C. elegans that were given dilute inoculums of an even mixture of two identical but differently colored bacterial strains [203].

Understanding the extent to which stochastic versus deterministic forces shape colonization is important for therapeutic design. Stochasticity may be an underlying reason why FMTs and probiotics are variable in efficacy [204–209]. As an example, mother to infant transmission is highly variable, with many maternal strains unsuccessful in colonizing the infant and infant guts showing a high rate of strain turnover [146].

From these studies, it is evident that controlling for genetics, diet and environment is important if we are to discern microbiome colonization at the species and strain levels. In humans, we cannot fully control for genetics even in family studies, and environment and diet will often be variable even with careful study design. Mice are a powerful model organism to understand bacterial strain colonization differences due to the ability to control for genetics and environment. The differences in colonization patterns even with genetically identical hosts are consistent with findings in other studies that found that bacterial strains colonize the gut stochastically [203].

Other forces acting on the gut are operating simultaneously with stochasticity to shape colonization. Gut colonization can be very context dependent, where results differ depending on which other species are co-colonizing [210]and environmental factors like host diet [189,211,212]. Sharing of environments is also important [213–215]. We found that shared environment leads to higher rates of strain sharing between hosts, possibly because strains present in one host have repeated opportunities to colonize other hosts when they live in proximity. Previous research has shown that coprophagy facilitated similar strain composition of barcoded *Escherichia coli* between cohoused mice through migration [216]. Similar migration patterns could explain the high rates of strain sharing we observe between cohoused mice for 5 other species in this study. Additionally, the biochemical environment can favor the presence of certain species. We observed evidence for the colonization of different strains of *B. wexlerae* in the ilea and jejuna of mice, which are also the two segments with the highest pH (**Figure S1**).

We must also consider intrinsic qualities of the microbes that shape the stochasticity. For example, a strain's ability to adhere to the gut impacts its colonization success. Dodge et al. 2023

found that primary colonizers were particularly successful at binding to the foregut of drosophila and could outcompete secondary colonizers [210]. In this case and in others, a strain can preclude another strain from the same species from successfully colonizing.

In certain species (e.g., *B. uniformis*), we observed evidence for the same set of strains being present at similar frequencies everywhere the species was found within the gut. However, we also observed evidence that other species (*B. wexlerae*) can exhibit spatial variation that is replicated across hosts. Our findings raise two questions: why do certain species maintain stable strain abundances across the species' range, while others exhibit strain-level spatial organization along the gut? Notably, *B. wexlerae* was found throughout the gut, whereas *B. uniformis* and the other three species were most abundant in either the upper or lower gut, respectively. This raises the possibility that *B. wexlerae*'s extended ecological range (all along the length of the gut) is due to the fact that *B. wexlerae* strains monopolize different tissues in the gut. Contrastingly, species that exhibited more homogeneous strain mixtures across samples often seemed to be specialized for survival in the upper or lower gut, but not both.

Without high coverage sequencing, it was difficult to definitely confirm the absence of a bacterial strain. Additionally, we lacked stool samples from the mice study. To truly evaluate variation along a spatial axis, more fine-grained sampling along the gut may be needed, particularly because previous studies found that variation in occupancy happened at the micrometer scale [203,210,217]. It is therefore conceivable that strains, like species, display a more small-scale spatial organization than could be assessed by the sampling strategy in this study, warranting further research. Deeper sequencing will also allow us to assess whether the same

strains of species that are highly abundant in one region of the gut are present everywhere in the gut.

Our findings point to a number of future research avenues to more definitely understand colonization. Disentangling the effect of diet and cohousing will be important in future work to understand how these two forces affect strain colonization patterns independently of one another. It may also be useful in a future study to systematically introduce different combinations of strains in a similar fashion to Jones et al. 2022's introduction of different combinations of species [212], which would allow us to infer how species and different strains within a complex community interact to shape strain-level colonization patterns. In addition, horizontal gene transfer between strains of the same species could have affected the strain colonization strain colonization dynamics we observed in our mouse cohort, including through the transfer of adaptive alleles between strain backbones. Thus, developing methods to detect horizontal gene transfer events in this metagenomic dataset and others is warranted. Recently, Dubinsky et al. previously studied strain colonization along the guts of a cohort of humans for which gut endoscopy and stool metagenome data was collected over 5 weeks, but only looked at MAGs in 3 species (*B. fragilis, Ruminococcus gnavus*, and *E. coli*). popANI could be used to identify strain occupancy differences in probiotics and antibiotics patients before and after treatment in this same cohort. Understanding how diet impacts strain and gene composition will require sequencing a larger cohort more deeply, and a cohousing setup that provides more opportunity to disentangle cohousing and diet effects.

Methods

Sequencing

Samples were whole metagenomic shotgun sequenced to a average depth of 30M reads (see **Table S1**).

Alpha diversity of samples

Species richness was computed using the Shannon diversity metric as implemented by the vegan package in R (cite vegan package). Stacked barplots were produced by summarizing species across taxonomic families but excluding species with less than 0.1 % abundance. To compute significance of differences in Shannon diversity, all mice sampled at a given tissue were pooled together and compared against a different tissue using the Wilcoxon signed rank test (paired = TRUE).

Pairwise Bray-Curtis Dissimilarity Index between samples within and across host

Relative species abundances were calculated from single-copy marker gene coverage as part of the MIDAS analysis [110]. For each sample, relative species abundances add to 1. These species abundances were used to calculate Bray-Curtis dissimilarity indices between all samples using the Vegan package in R. These indices were plotted for within-host and between-host comparisons, respectively (**Figure S2**). Bray-Curtis dissimilarity indices were used as the proximity matrix in principal coordinate analysis (PCoA), and samples were visualized on two axes.

For each species, gene copy numbers were estimated from the read counts assigned to a gene relative to the median coverage of 15 known single copy genes [110]. Bray-Curtis dissimilarity indices between samples were calculated using the gene copy numbers of all species deemed present in two or more samples. Species were deemed present if they had a mean genome-wide coverage greater than equal to 3, and dissimilarity indices were only calculated between samples when the species was present in both. Within- and between-host Bray-Curtis dissimilarity distributions were plotted. For five species, PCoA analysis was performed using gene-copy-number-based Bray-Curtis dissimilarity indices as the proximity matrix, and samples were visualized on two axes based on the similarity of their gene composition for each species.

Filtering of genetic loci

A site was considered in Pi and Fst calculations only when there were at least 4 reads. In the presence of variation, an additional requirement was imposed: each allele needed to have at least 2 reads supporting that observation. This was imposed to reduce the effects of singletons from sequencing error. Additionally, sites lying in genes that are known to cause unusual read mapping, often due to multicopy genes, were excluded.

Pi computation

Pi represents the probability of randomly choosing two different alleles at a randomly chosen base pair in the genome. Pi was computed for each sample and pair of samples using the formulas for nucleotide diversity applied in Schloissnig et al. 2013 [117], which accounts for total

read counts in each sample. This formula is an extension of a previously proposed pi estimator on NGS data devised by Begun et al. [195]. Mean pi was computed by summing the pi values across all considered sites and dividing by the number of such sites. Pairwise pi was computed by pooling reads from two samples, calculating the pi values across all considered sites and summing them.

Distributions of pi were plotted for 36 species. These species were selected based on the criteria that they had a minimum coverage of 4 reads in at least 3 samples taken from at least 2 hosts. We used pi to select 5 species with high between-host pairwise pi ($pi \geq 1 \times 10^{-3}$) and low between-host pairwise pi ($pi < 1 \times 10^{-3}$). These two groups would allow us to examine strain colonization dynamics in scenarios where (1) mice are potentially exposed to multiple strains of a species in the original inoculum or (2) mice are exposed to only a single strain in the original inoculum. For the former group, we chose the only 5 species with between-host pairwise pi that exceeded $pi \geq 1 \times 10^{-3}$ (*Adlercreutzia equolifaciens*, *Bacteroides uniformis*, *Bacteroides vulgatus*, *Blautia wexlerae*, and *Clostridiales bacterium*). When selecting 5 species with low pi, we imposed a more stringent filter of requiring species to have a minimum coverage of 20X in at least 2 samples in 3 hosts, as low coverage was found to produce artificially low and high nucleotide diversity estimates (**Figure S10**). From the resulting list of species, we selected the 5 with the lowest between-host pairwise pi values (*Bacteroides cellulosilyticus*, *Blautia producta*, *Burkholderiales bacterium*, *Coprococcus* sp. 62244, and *Parabacteroides_distasonis_56985*).

Pairwise Fst between samples within and across hosts

For each pair of samples of interest, a summary Hudson F-st was computed using the methods of Bhatia et al. 2013 [196] and Hudson 1992 [197]. Each sample was considered a "population". The Hudson's F-st was computing by taking the divergence between the two samples minus the average diversity within each sample and dividing this value by the divergence between the two samples. To get a mean F-st across variants, the numerator was added up across all considered sites (see *filtering of genetic loci* above) and then divided by the denominator added up across all considered sites. Computation was done using the scikit-allel package.

PopANI analysis

popANI is a pairwise metric developed as a part of the inStrain pipeline (inStrain reference). As input, inStrain takes FASTA files and BAM files, the latter of which were produced using samtools mpileup (samtools reference) as a part of the MIDAS workflow. inStrain produced counts of the number of loci per scaffold that were fixed for different alleles between samples (i.e., "substitutions"). By default, inStrain only considered loci if they had coverage of 5 or greater in both samples being compared and passed a false discovery threshold of 1×10^{-6} predetermined by inStrain developers. A global popANI value was generated by aggregating scaffolds by species and dividing the pooled substitution count by the total number of genomic positions considered in the popANI calculations. Based on a benchmarking analysis, inStrain developers determined that popANI values exceeding 99.999% indicate that identical strains are shared between samples, which is the threshold used in this study for determining strain sharing. Here, comparisons were retained only if the number of loci across all scaffolds for a given

species that passed popANI filters was greater than or equal to 5×10^5. Samples with only 5×10^5. comparable sites between them would need to have at least 5 fixed substitutions between them to detect a strain or strains that are not shared between the two samples.

Distributions of popANI values were plotted for the 5 species with mean inter-host, inter-sample pi values exceeding 1×10^{-3}) and 5 species with mean inter-host, inter-sample pi values below 1×10^{-3}. popANI distributions were plotted for within-host and and between-host comparisons, as well as between-host comparisons across mice that were and were not cohoused together. In addition, we take a closer look at the within-host distributions of popANI for *B. wexlerae*.

1- and 2-dimensional Site Frequency Spectra

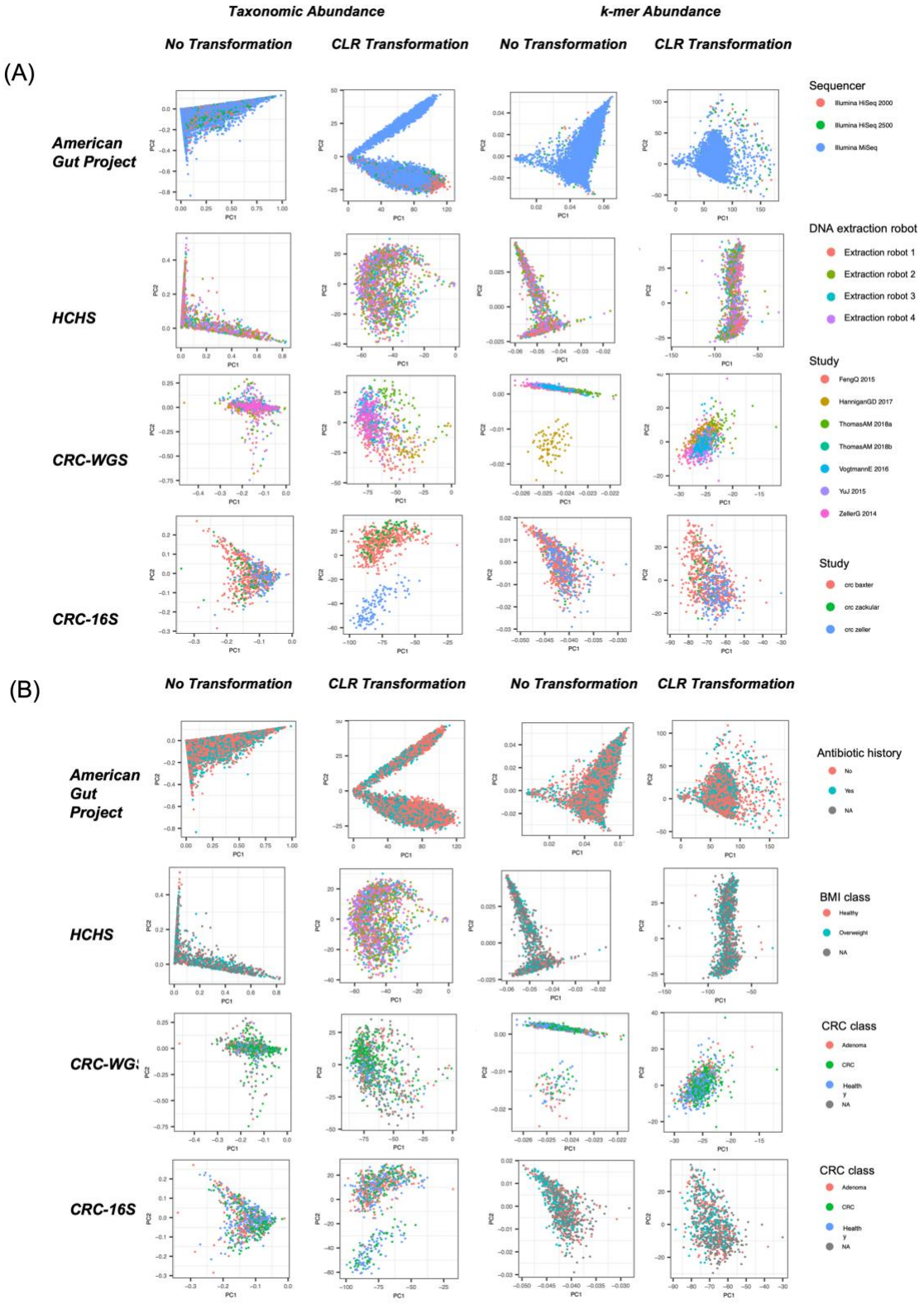
1-dimensional site frequency spectra (1D SFS) were produced for *B. wexlerae*, *B. uniformis*, and other species by counting the number of loci with at least 20X coverage that fell into each of 200 equally sized allele frequency bins ranging from 0 to 1.

2-dimensional site frequency spectra (1D SFS) were produced for the same species by polarizing alleles such that the allele that is the minority allele in over 50% of samples was designated the alternate allele. After polarization, polymorphic loci in each sample were extracted. Loci were fully removed from the analysis if they failed to be polymorphic in at least a quarter of samples. Finally, loci filtered out if they did not have at least 10X coverage in a given sample. For a given species, the 2D SFS was produced by selecting pairs of samples and plotting the allele frequencies of a locus in both samples on the x and y axes, respectively.

Source Tracking

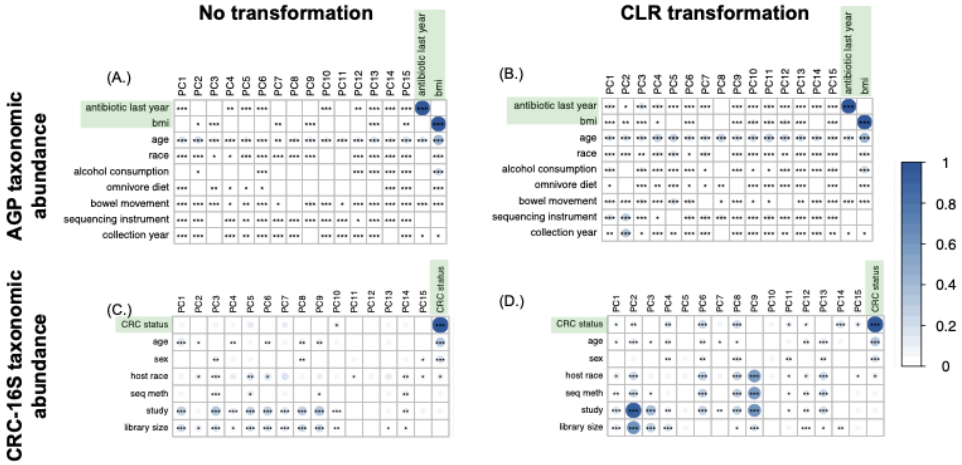
For source tracking with species, the relative abundance computed from marker gene abundance as computed by MIDAS was input into the source tracking software FEAST [109]. For source tracking with SNVs, the software Signature SNVs [193] was applied to determine a signature SNV set, with the parameters for minimum reads per site set at 5. The alternative and reference allele counts from the resulting signature SNVs were input into FEAST.

Supplementary Material 1: Evaluating supervised and unsupervised background noise correction in human gut microbiome data

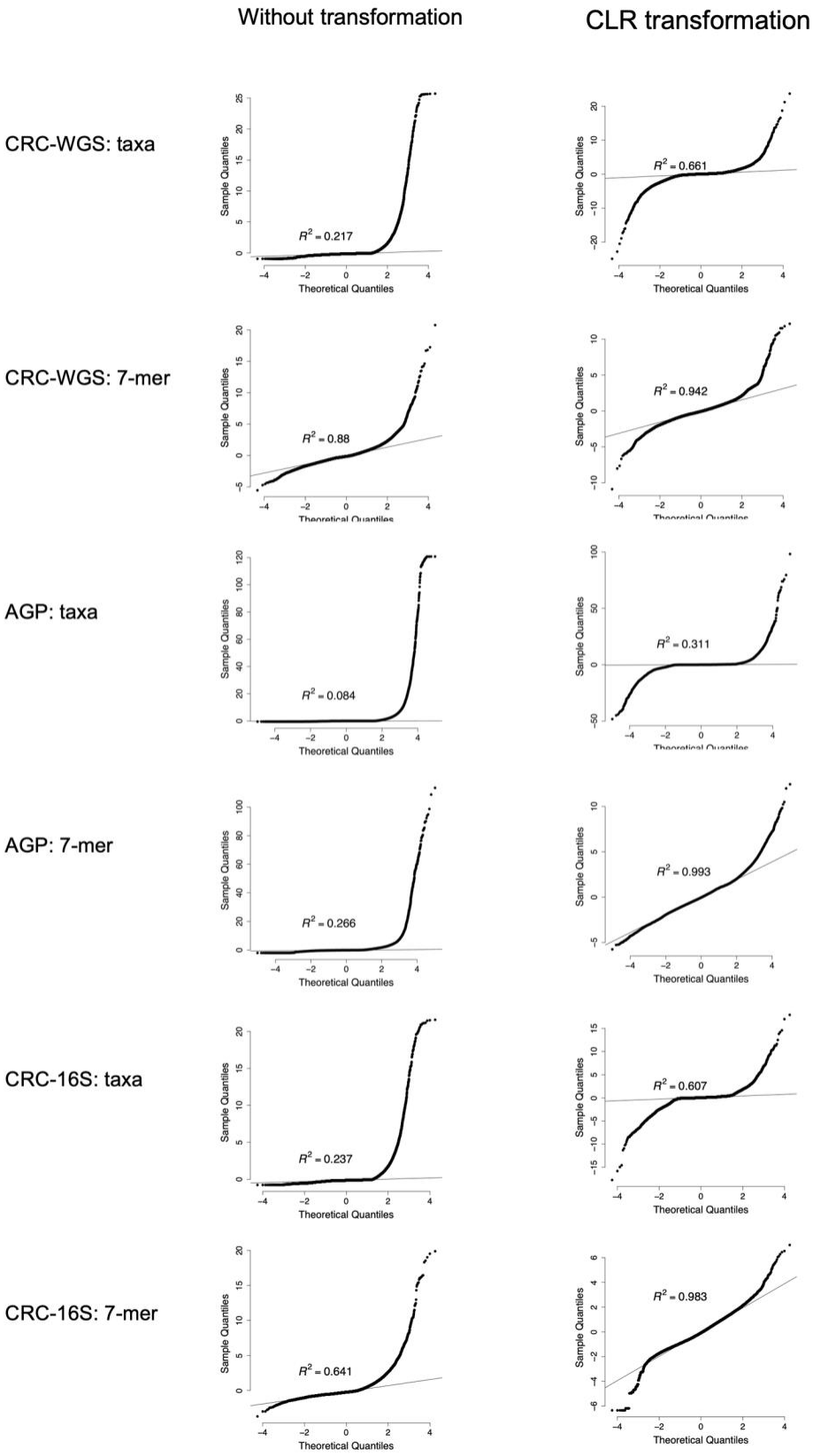


19 **Fig. S1. First two principal components from microbiome dataset studied.** PCA was applied to taxonomic abundance profiles and 6-mer data from the AGP, CRC-WGS merged dataset, CRC-16S merge

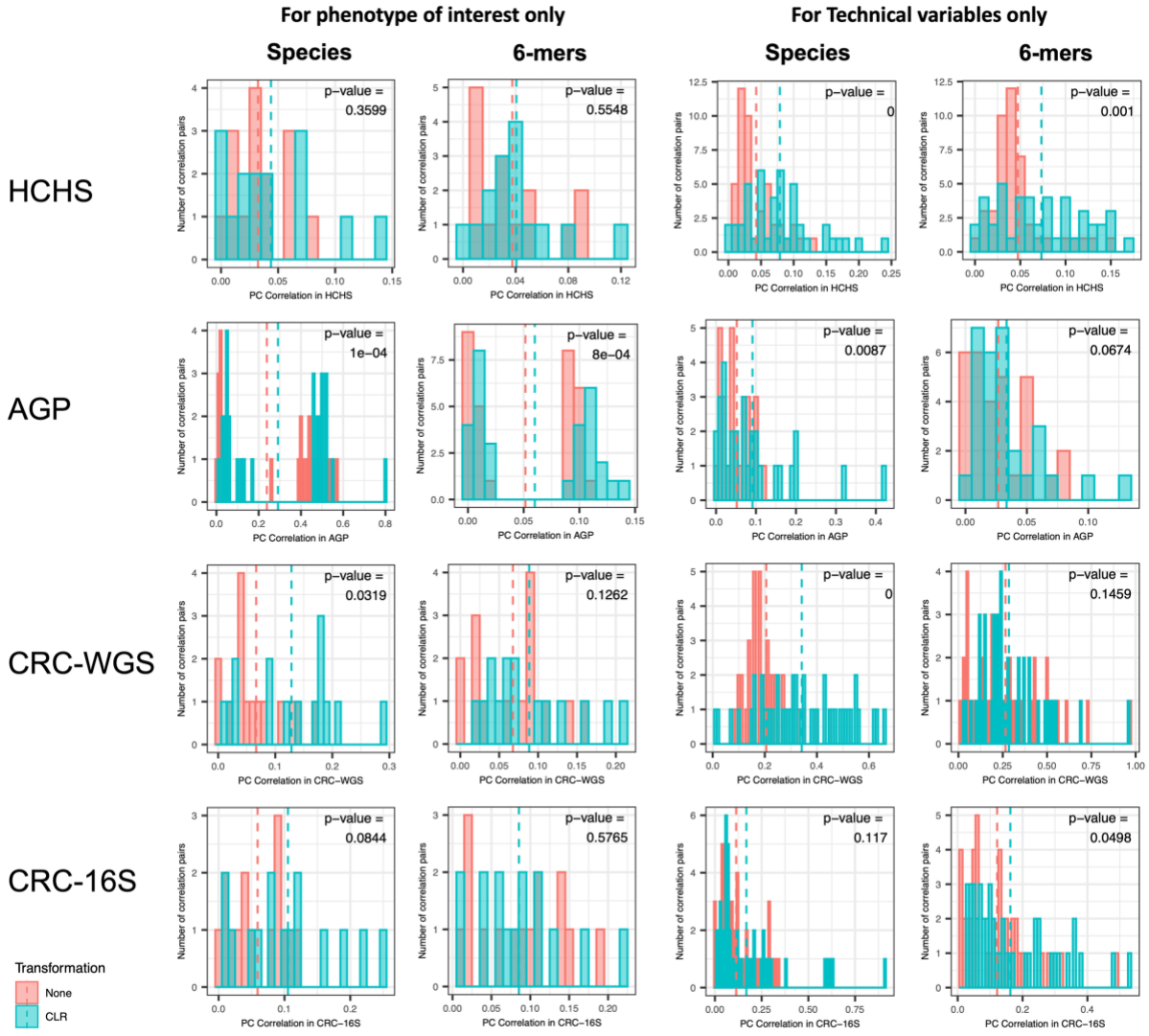
datasets, and Hispanic Community Health Cohort. Samples were plotted along the first 2 PCs with colors indicating (A) dataset or batch membership and (B) phenotype label.



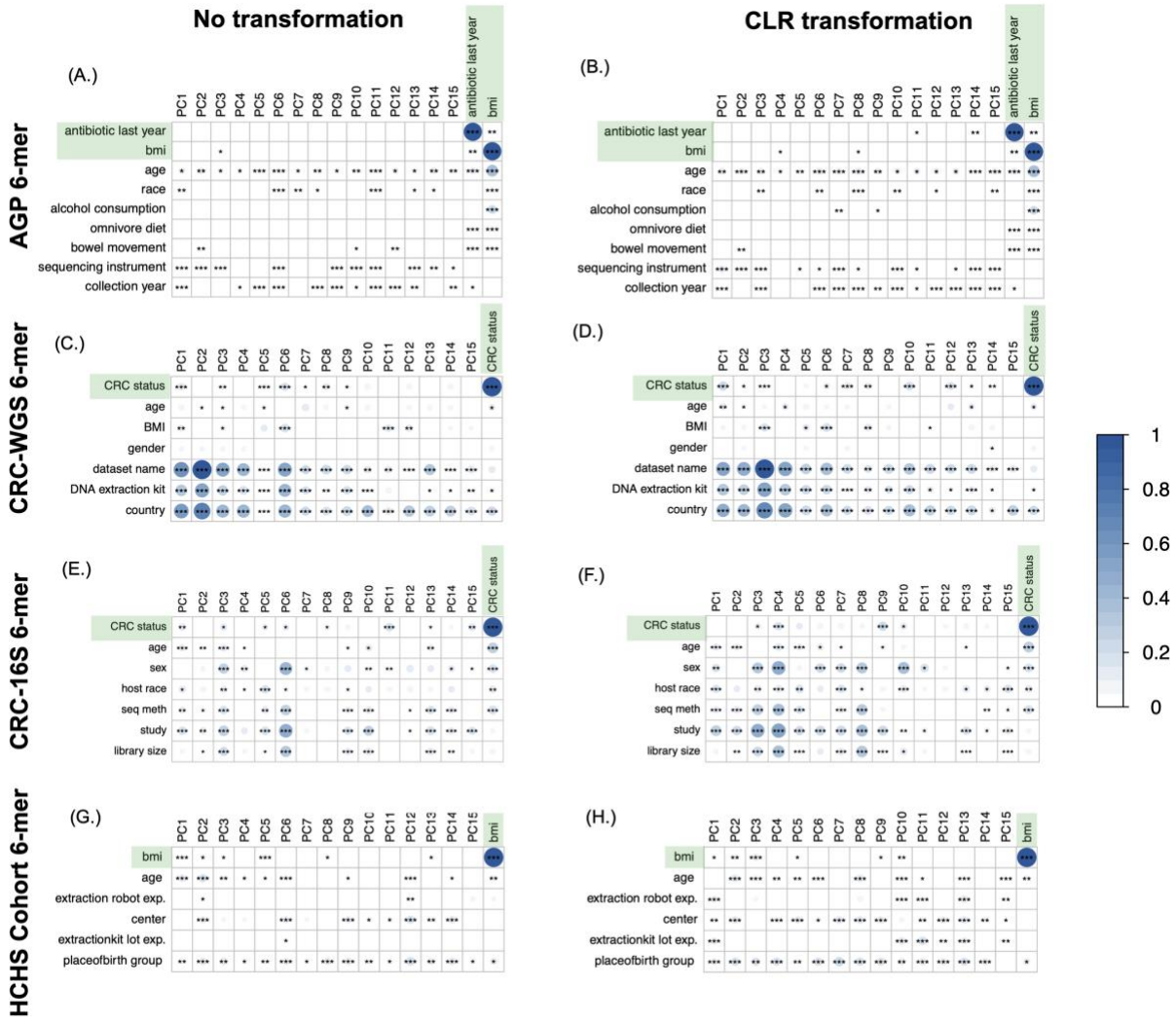
20 **Fig. S2. Top principal components from the CRC-16S dataset correlate with technical and biological covariates.** The first 15 PCs in the CRC-16S taxonomic abundance joined datasets are correlated with variables measured in each of the studies, including phenotype, sex, age, race, dataset label, sequencing method, library size and several others in (A, B) AGP, (C, D) CRC-16S. The size and color of the circles in each cell indicate the magnitude of correlation while black asterisks indicate the significance of the Pearson correlation of the PCs with each of the variables. The color bar at right of each plot represents the range of correlations observed across all datasets. [*,**,*** indicate p-values as follows: $10^{-2} < p < 0.05$, $10^{-3} < p < 10^{-2}$, $p < 10^{-3}$].


Without transformation
CLR transformation
CRC-WGS: taxa
CRC-WGS: 7-mer
AGP: taxa
AGP: 7-mer
CRC-16S: taxa
CRC-16S: 7-mer
Sample Quantiles
Theoretical Quantiles
R^2 = 0.217
R^2 = 0.661
R^2 = 0.88
R^2 = 0.942
R^2 = 0.084
R^2 = 0.311
R^2 = 0.266
R^2 = 0.993
R^2 = 0.237
R^2 = 0.607
R^2 = 0.641
R^2 = 0.983

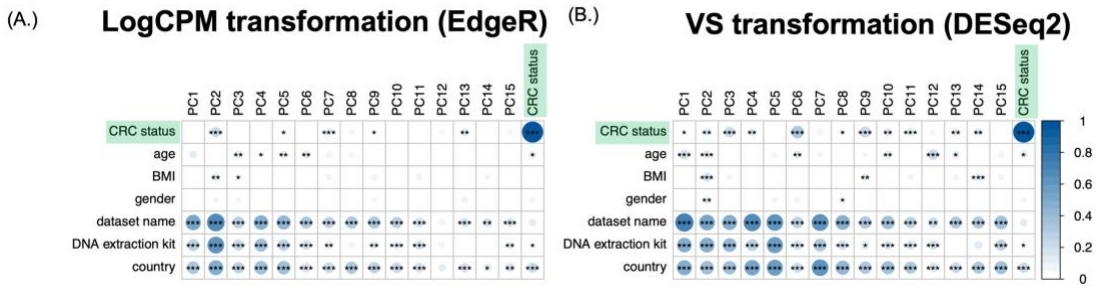
21 **Fig. S3. Quantile-Quantiles plot for AGP, CRC-WGS, and CRC-16S before and after the CLR-transformation.** The quantiles of 100 randomly-selected taxonomic features or *k*-mers, that were converted to z-scores, ranked against the expected quantiles from a normal distribution of mean 0 and variance 1. The R-squared values are reported in the annotated text.



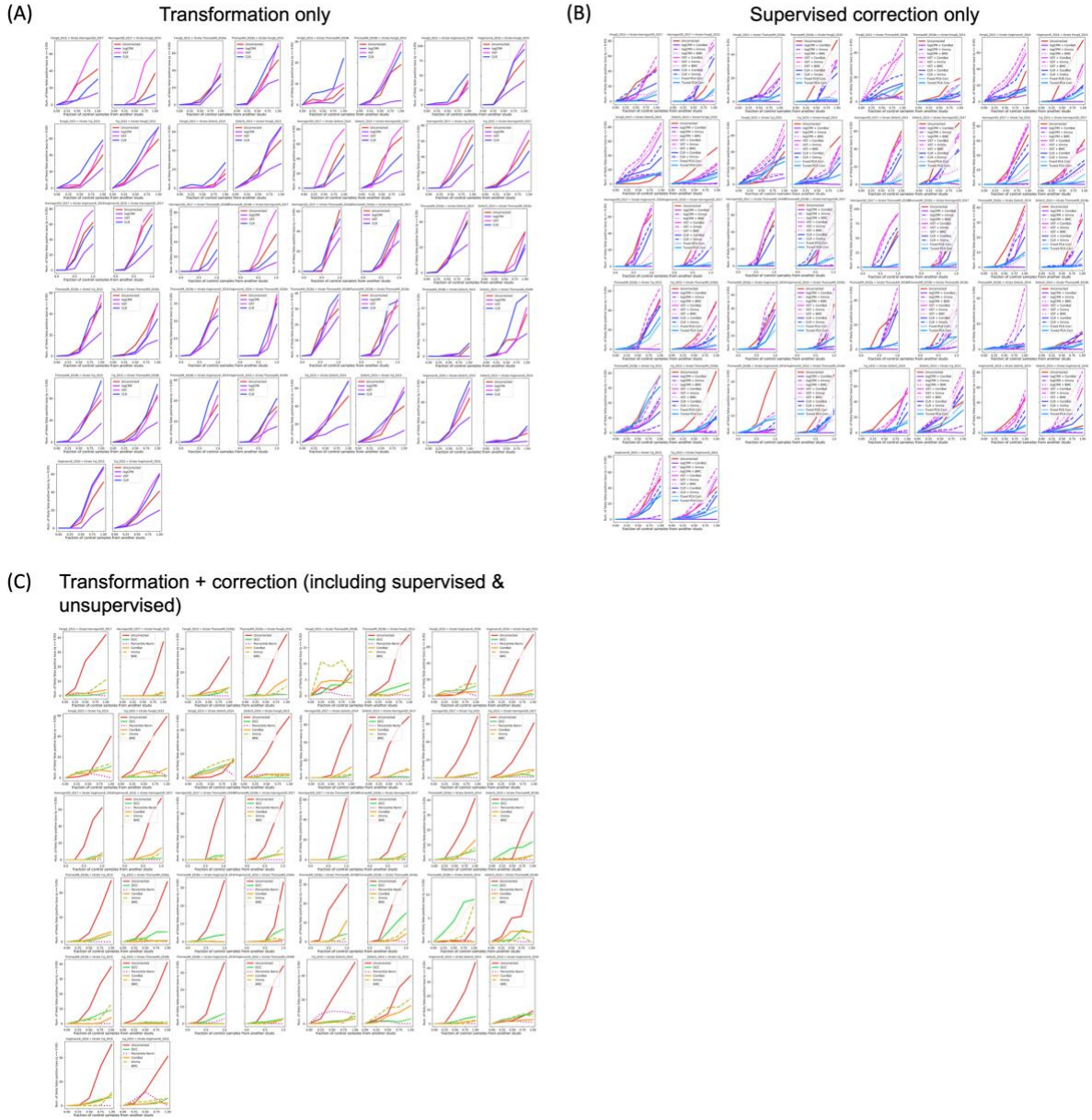
22 **Fig. S4. Histogram of correlation between top 15 PCs and various measured variables.** Histograms show the distribution of correlation values computed between the top 15 PCs of taxonomic features in each dataset and the phenotype covariates and technical covariates. Shown in black text are the Kolmogorov-Smirnov test p-values for the test of the null hypothesis that the distribution of correlations in the non-transformed data is no different from the correlations in the CLR-transformed data. HCHS is the only dataset with significant increase in correlation in the technical covariates but not the phenotype of interest.



23 **Fig. S5. Top principal components from 6-mers correlate with technical and biological covariates.** The first 15 PCs before (a, c, e, and g) and after (b, d, f, and h) the CLR-transformation are correlated with variables measured in each of the studies, including dataset label, library size, DNA extraction kit used, country of origin, age, body mass index (BMI), sex, and colorectal cancer status (CRC). The size and color of the circles in each cell indicate the magnitude of correlation while black asterisks indicate the significance of the Pearson correlation of the PCs with each of the variables. The color bar at right of each plot represents the range of correlations observed across all datasets. [*,**,*** indicate *p*-values as follows: $10^{-2} < p < 0.05$, $10^{-3} < p < 10^{-2}$, $p < 10^{-3}$].



24 **Fig. S6.** Top principal components from LogCPM and VST transformed taxonomic abundance correlate with technical and biological covariates. The first 15 PCs from data transformed with the (A) EdgeR log counts per million (LogCPM) transformation[46] and (B) DESeq2 Variance Stabilizing (VS) transformation are correlated with variables measured in each of the studies, including dataset label, library size, DNA extraction kit used, country of origin, age, body mass index (BMI), sex, and colorectal cancer status (CRC). The size and color of the circles in each cell indicate the magnitude of correlation while black asterisks indicate the significance of the Pearson correlation of the PCs with each of the variables. The color bar at right of each plot represents the range of correlations observed across all datasets. [*,**,*** indicate *p*-values as follows: $10^{-2} < p < 0.05$, $10^{-3} < p < 10^{-2}$, $p < 10^{-3}$].



25 **Fig. S7. Titration analysis for new false positive associations.** For each study in CRC-WGS, an equal number of cases and controls were drawn to determine significant taxa associated with CRC. Then, at proportions of 25%, 50% and 100%, control samples were replaced with controls from a second study. This experiment was repeated after applying (A) transformations, (B) corrections, or (C) a combination of both (including unsupervised methods) to compare the extent to which new false positive associations arise with increasing confounding between CRC and study label.

		Proportion of controls from second study				
		0	**0.25**	**0.5**	**0.75**	**1**
Data Transformation	**Uncorrected**	0	1	10	26	42
	logCPM	0	1	4	12	20
	VST	0	1	12	32	52
	CLR	0	2	10	28	44
Supervised Correction	**DCC**	0	1	2	4	5
	Percentile normalization	0	1	1	1	0
	ComBat	0	1	1	3	5
	limma	0	1	2	3	6
	BMC	0	1	2	3	5
Transformation + Correction (including unsupervised)	**logCPM + ComBat**	0	0	1	1	1
	logCPM + limma	0	1	1	1	2
	logCPM + BMC	0	1	1	1	2
	VST + ComBat	0	2	7	25	44
	VST + limma	0	3	10	32	55
	VST + BMC	0	3	5	11	25
	CLR + ComBat	0	2	4	13	26
	CLR + limma	0	3	6	20	35
	CLR + BMC	0	16	36	94	173
	Fixed PCA correction	0	1	3	7	14

	Tuned PCA correction	0	1	2	6	11

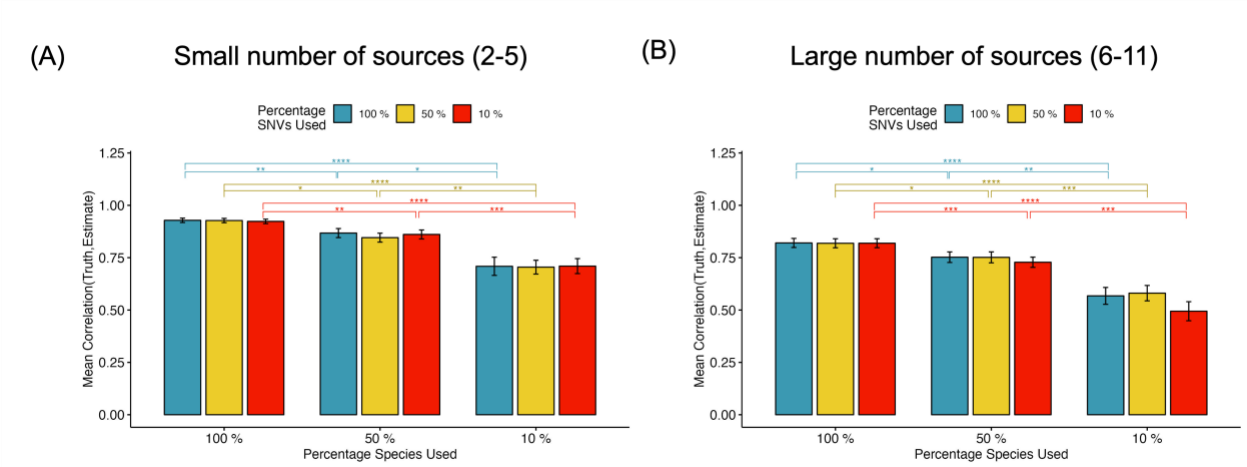
T 3 **Table S1. Mean number of new associations in titration experiment.** Shown is the mean number of likely false positive associations with respect to the original study 1 case and controls before adding control samples from study two, across all pairs of studies within CRC-WGS and across all five-fold replicates of titration at each mixing proportion of 0 %, 25%, 50%, 75%, and 100% controls from study two.

Supplementary Material 2: SNV-FEAST: microbial source tracking with single nucleotide variants

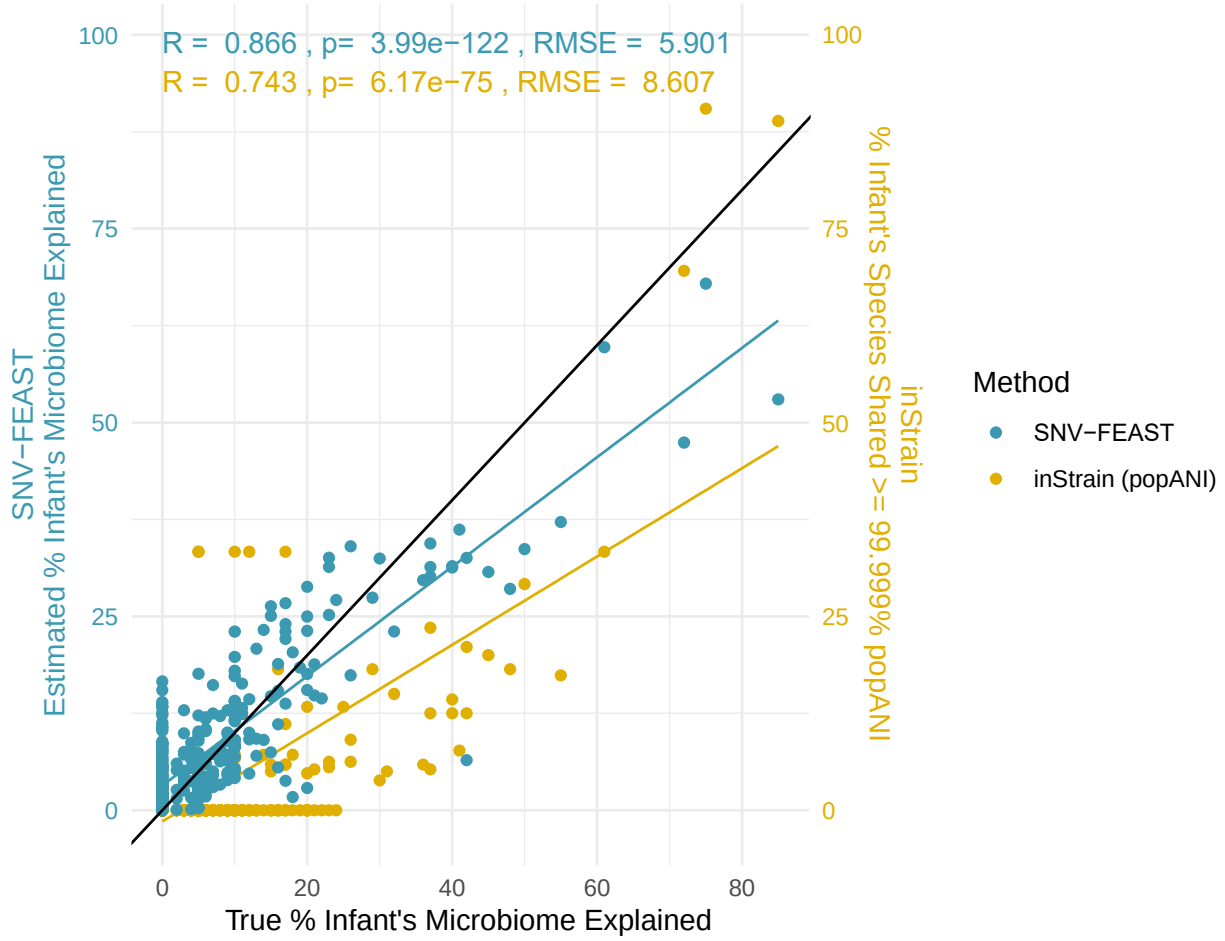
		Source 1	Source 2	Source 3	Source 4	Source 5	Source 6	Source 7	Source 8	Source 9	Source 10	Unknown
Complex Sink	Trial 1	0.2	0.15	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.25
	Trial v2	0.3	0.1	0.08	0.08	0.08	0.06	0.05	0.05	0.05	0.05	0.1
	Trial 3	0.16	0.12	0.07	0.05	0.04	0.04	0.03	0.03	0.02	0	0.44
	Trial 4	0.21	0.18	0.11	0.1	0.1	0.05	0.05	0.05	0.05	0	0.1
	Trial 5	0.23	0.07	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0	0.35
	Trial 6	0.17	0.13	0.05	0.05	0.05	0.05	0.05	0.05	0	0	0.4
	Trial 7	0.24	0.16	0.13	0.1	0.05	0.05	0.05	0.05	0	0	0.17
	Trial 8	0.2	0.17	0.06	0.06	0.06	0.05	0.05	0	0	0	0.35
	Trial 9	0.4	0.16	0.06	0.06	0.06	0.06	0.05	0	0	0	0.15
	Trial 10	0.09	0.03	0.03	0.02	0.02	0.02	0	0	0	0	0.79
	Trial 11	0.17	0.15	0.12	0.1	0.1	0.05	0	0	0	0	0.31
	Trial 12	0.17	0.15	0.12	0.1	0.1	0.05	0	0	0	0	0.31
	Trial 13	0.2	0.17	0.1	0.1	0.1	0.05	0	0	0	0	0.28
	Trial 14	0.25	0.15	0.11	0.06	0.05	0.05	0	0	0	0	0.33
	Trial 15	0.37	0.12	0.1	0.09	0.09	0.07	0	0	0	0	0.16
	Trial 16	0.42	0.11	0.1	0.09	0.09	0.07	0	0	0	0	0.12
	Trial 17	0.42	0.11	0.1	0.07	0.06	0.06	0	0	0	0	0.18
	Trial 18	0.45	0.1	0.1	0.09	0.09	0.07	0	0	0	0	0.1
	Trial 19	0.22	0.2	0.11	0.08	0.05	0	0	0	0	0	0.34
	Trial 20	0.23	0.05	0.04	0.03	0.02	0	0	0	0	0	0.63
	Trial 21	0.29	0.1	0.04	0.04	0.03	0	0	0	0	0	0.5
	Trial 22	0.37	0.03	0.03	0.03	0.03	0	0	0	0	0	0.51
	Trial 23	0.41	0.26	0.1	0.05	0.05	0	0	0	0	0	0.13
Simple Sink	Trial 24	0.2	0.18	0.09	0.09	0	0	0	0	0	0	0.44
	Trial 25	0.37	0.26	0.21	0.05	0	0	0	0	0	0	0.11
	Trial 26	0.19	0.06	0.04	0	0	0	0	0	0	0	0.71
	Trial 27	0.32	0.2	0.07	0	0	0	0	0	0	0	0.41
	Trial 28	0.55	0.14	0.08	0	0	0	0	0	0	0	0.23
	Trial 29	0.75	0.15	0.05	0	0	0	0	0	0	0	0.05
	Trial 30	0.85	0.05	0.05	0	0	0	0	0	0	0	0.05
	Trial 31	0.06	0.04	0	0	0	0	0	0	0	0	0.9
	Trial 32	0.13	0.1	0	0	0	0	0	0	0	0	0.77
	Trial 33	0.16	0.04	0	0	0	0	0	0	0	0	0.8
	Trial 34	0.36	0.23	0	0	0	0	0	0	0	0	0.41
	Trial 35	0.72	0.14	0	0	0	0	0	0	0	0	0.14
	Trial 36	0.1	0	0	0	0	0	0	0	0	0	0.9
	Trial 37	0.17	0	0	0	0	0	0	0	0	0	0.83
	Trial 38	0.31	0	0	0	0	0	0	0	0	0	0.69
	Trial 39	0.4	0	0	0	0	0	0	0	0	0	0.6
	Trial 40	0.48	0	0	0	0	0	0	0	0	0	0.52
	Trial 41	0.5	0	0	0	0	0	0	0	0	0	0.5
	Trial 42	0.61	0	0	0	0	0	0	0	0	0	0.39

Unknown
0-30%
30-70%
70-90%

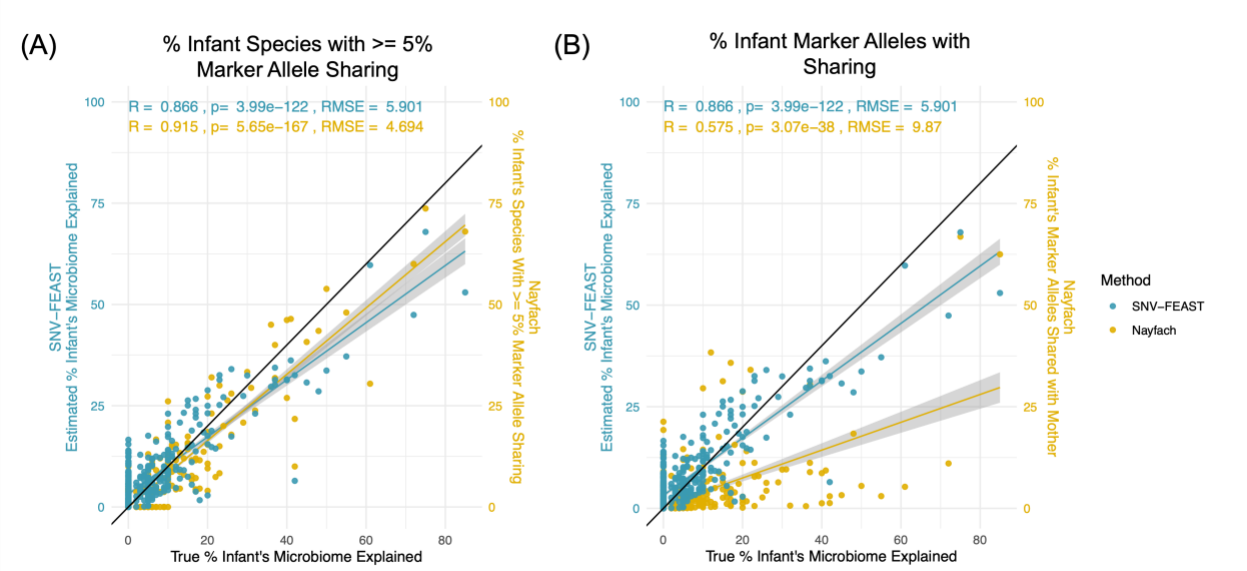
T 4 **Table S1. Mixing proportions for simulated infants.** To simulate complex (N sources > 5) and simple (N sources <= 5) sinks, we mixed varying proportions of reads from the FASTA files of real adult mothers extracted from the Backhed et al. 2015 dataset. Proportions shown represent the proportion of 10 million reads in infants that are taken from each source.



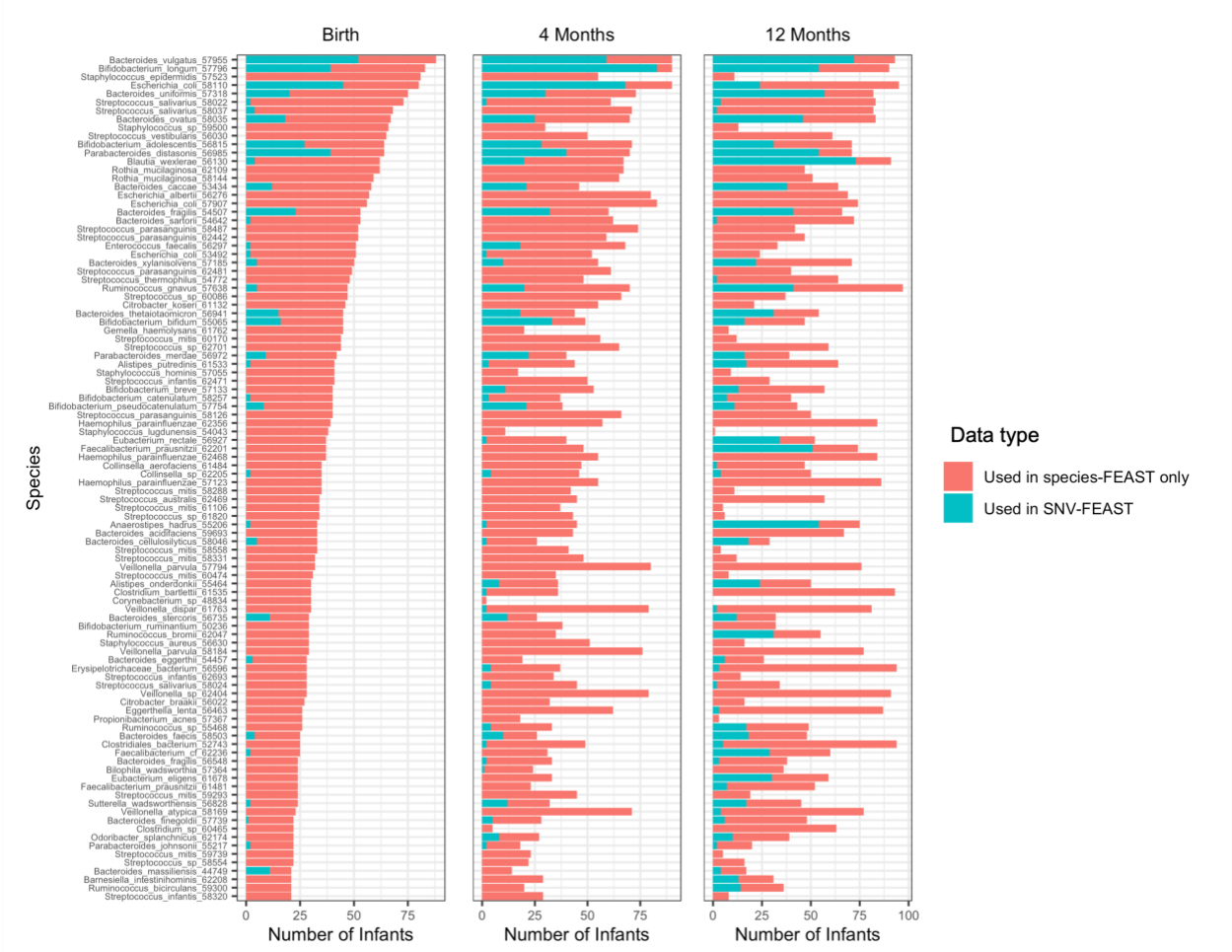
26 **Figure S1: Performance of SNV-FEAST as a function of fraction of species and SNVs included for analysis.** To assess whether all species and all signatures SNVs in the sink are needed for accurate source tracking with SNV-FEAST, we varied the proportion of species (from 10%, 50% or 100%) and SNVs (from 10%, 50% or 100%) included as inputs to the algorithm. The y-axis values are Pearson Correlations between the estimated and true source tracking proportions. The errors bars represent the standard error of the mean. (A) Simulations with small number of contribution sources. (B) Simulations with a large number of contributing sources.



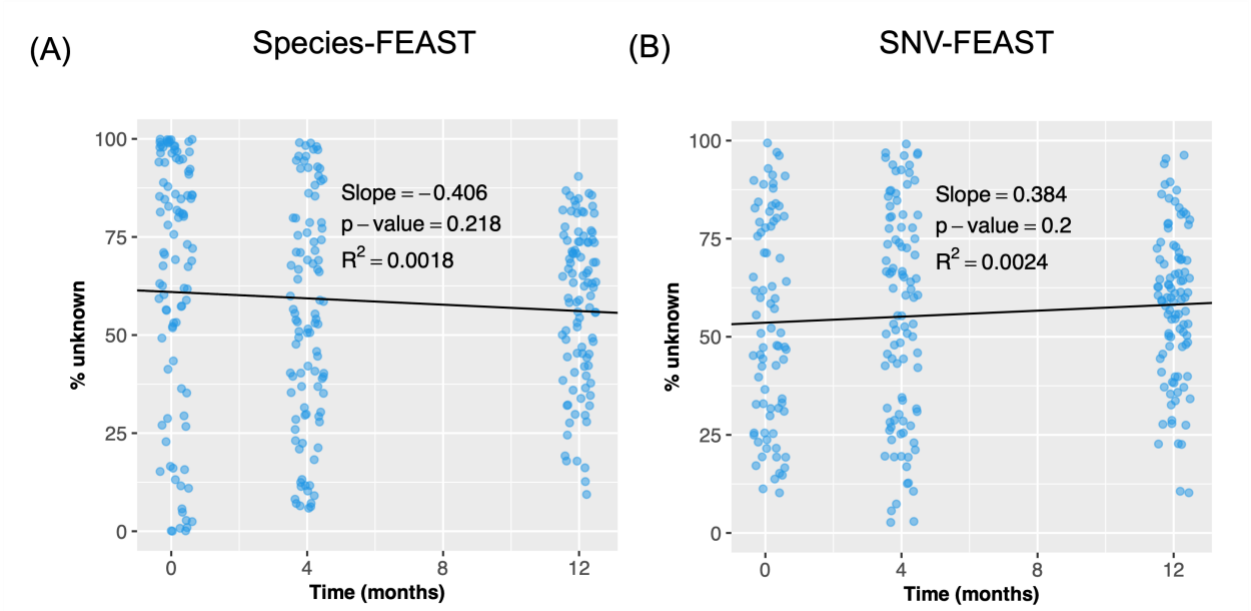
27 **Figure S2: Comparison of SNV-FEAST with inStrain.** Application of SNV-FEAST and inStrain on simulated infant gut microbiomes in which the number of contributing sources was varied from 2 to 11 and the percentage of those contributing sources was varied from 1% to 90%. The x-axis represents the true proportion of the infant seeded by the source. Each point represents an infant-source pair. In the case of SNV-FEAST, the y-value represents the source tracking estimate. In the case of inStrain, the y-value represents the fraction of species in the infant that have at least 99.999% popANI with the source. Shown in the inset text is Pearson correlation and corresponding p-value and RMSE for both approaches.



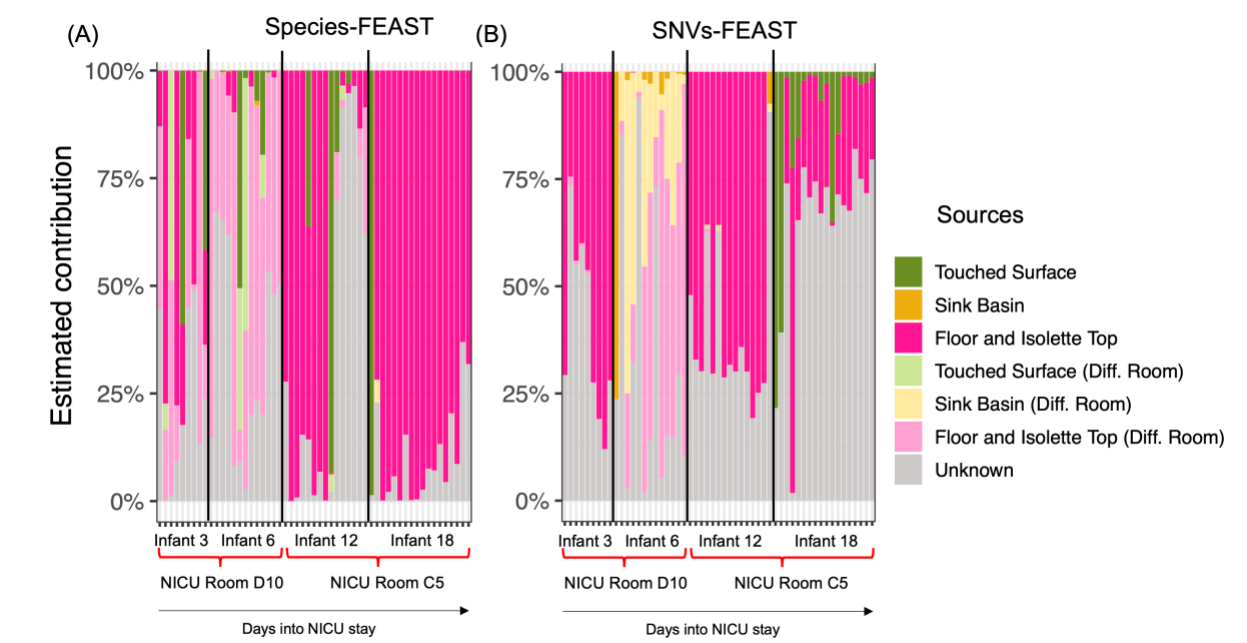
28 **Figure S3: Comparison of SNV-FEAST with the strain tracking approach in Nayfach et al. 2015.** Application of SNV-FEAST and Nayfach et al. 2016 on simulated infant gut microbiomes in which the number of contributing sources was varied from 2 to 11 and the percentage of those contributing sources was varied from 1% to 90%. The x-axis represents the true proportion of the infant seeded by the source. Each point represents an infant-source pair. In the case of SNV-FEAST, the y-value represents the source tracking estimate. In the case of Nayfach et al. 2016, the y-value in (A) represents the fraction of species in the infant have at least 5% marker allele sharing while the y-value in (B) represents the fraction of all marker alleles in the infant that are shared with a given mother. Shown in the inset text is Pearson correlation and corresponding p-value and RMSE for both approaches.



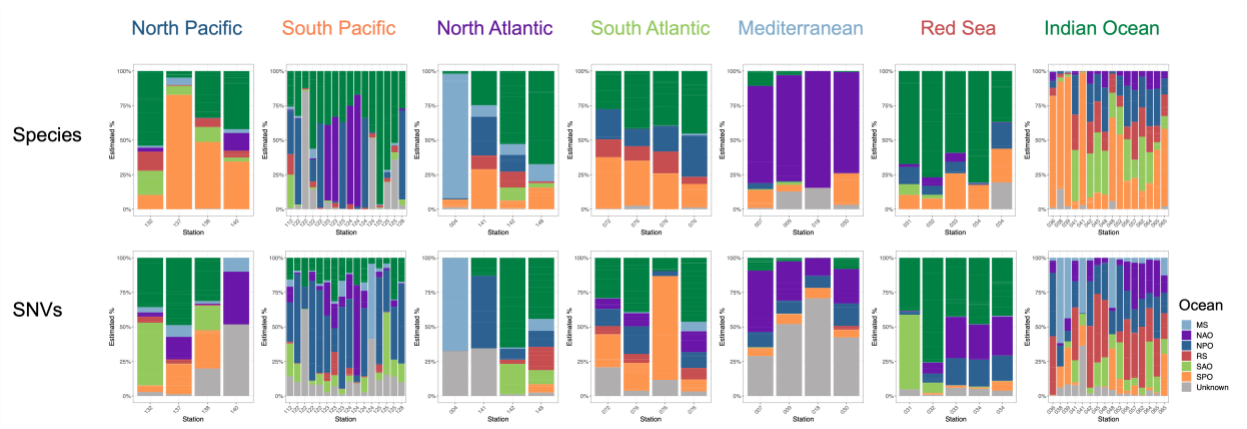
29 **Figure S4: Species with signature SNVs.** Number of infants in which certain species are detected in microbiome samples (whole bar) and in the signature SNV set obtained from those samples (teal bar) while the remained represents infants in which the species was only utilized in species-FEAST (salmon bar). Displayed are the 100 most prevalent species based on samples obtained from infants at birth.



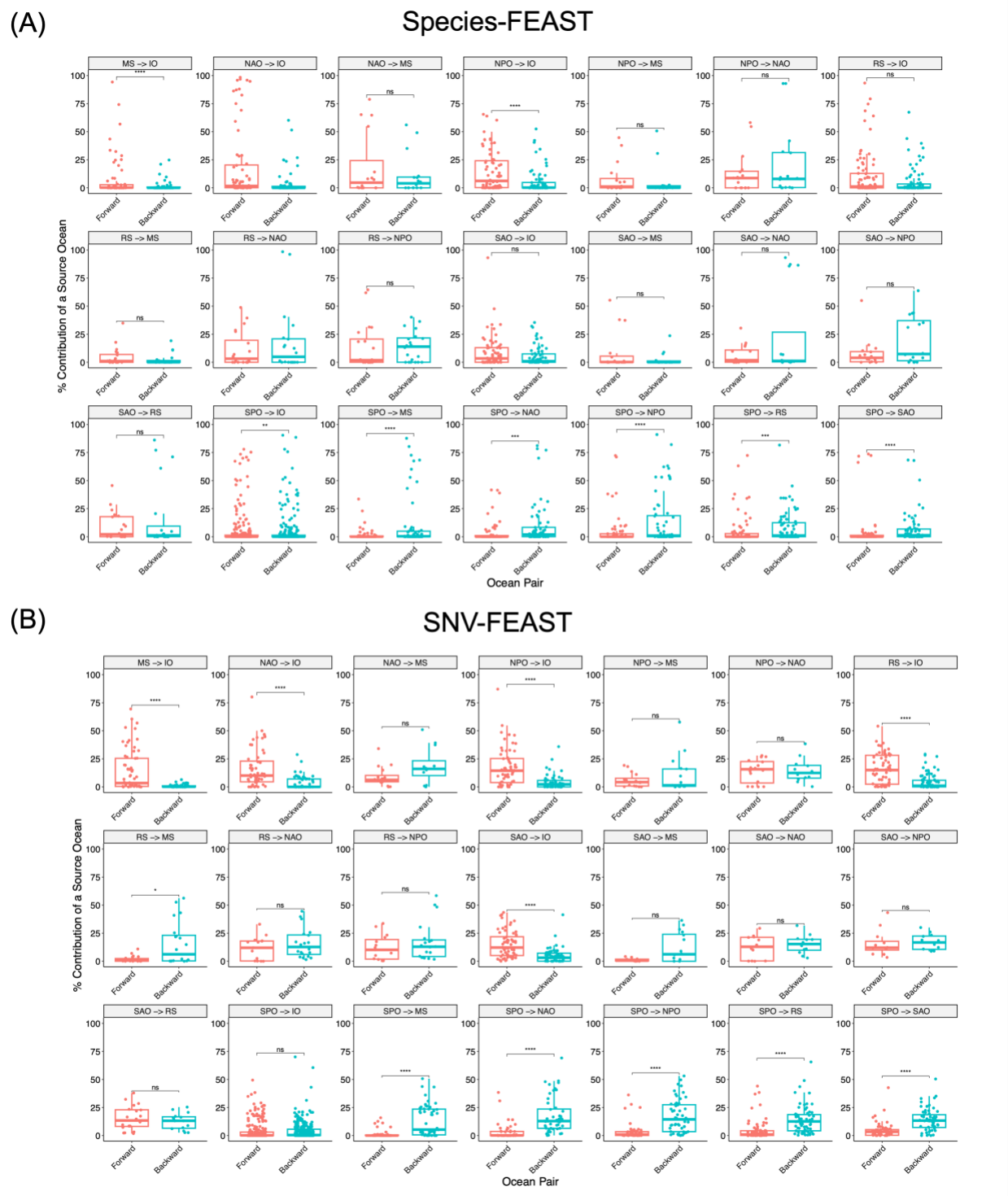
30 **Figure S5: Unknown component in microbial source tracking with infants in the first year of life.** Contribution of only unknown sources to the infant's gut microbiome at birth, four and 12 months when previous time points of the infant are excluded as sources. Note this is a different experiment from the one shown in Figure 3.



31 **Figure S6: Microbial source tracking with infants in the NICU and their built environment.** Contribution of samples from either the infant's own NICU room or a different room from the study estimated using (A) species-FEAST and (B) SNV-FEAST. This is the same data that is plotted in **Figure 4A**, except all potential sources are stacked. This permits visualization of proportion unknown.



32 **Figure S7.** Microbial source tracking in the Tara Oceans dataset with SNV and species-FEAST. Source tracking estimates for the contribution of different oceans are depicted with vertical bars for the North Pacific (n=4), South Pacific (n=16), North Atlan



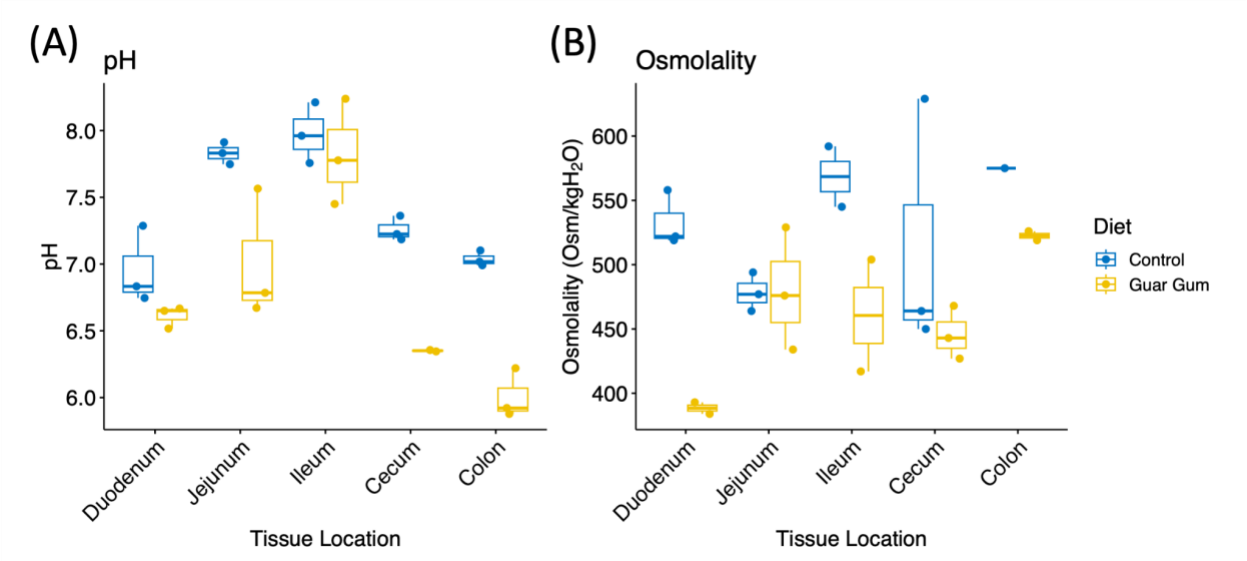
33 **Figure S8. Flipped source tracking for all ocean pairs** Shown are (A) species-FEAST and (B) SNV-FEAST estimates for contribution of one ocean to another. Each dot represents the contributions of each samples from the source ocean to the sink ocean of interest.

Supplementary Material 3: Effects of diet, spatial location, and shared environment on microbiome diversity along the mammalian gut

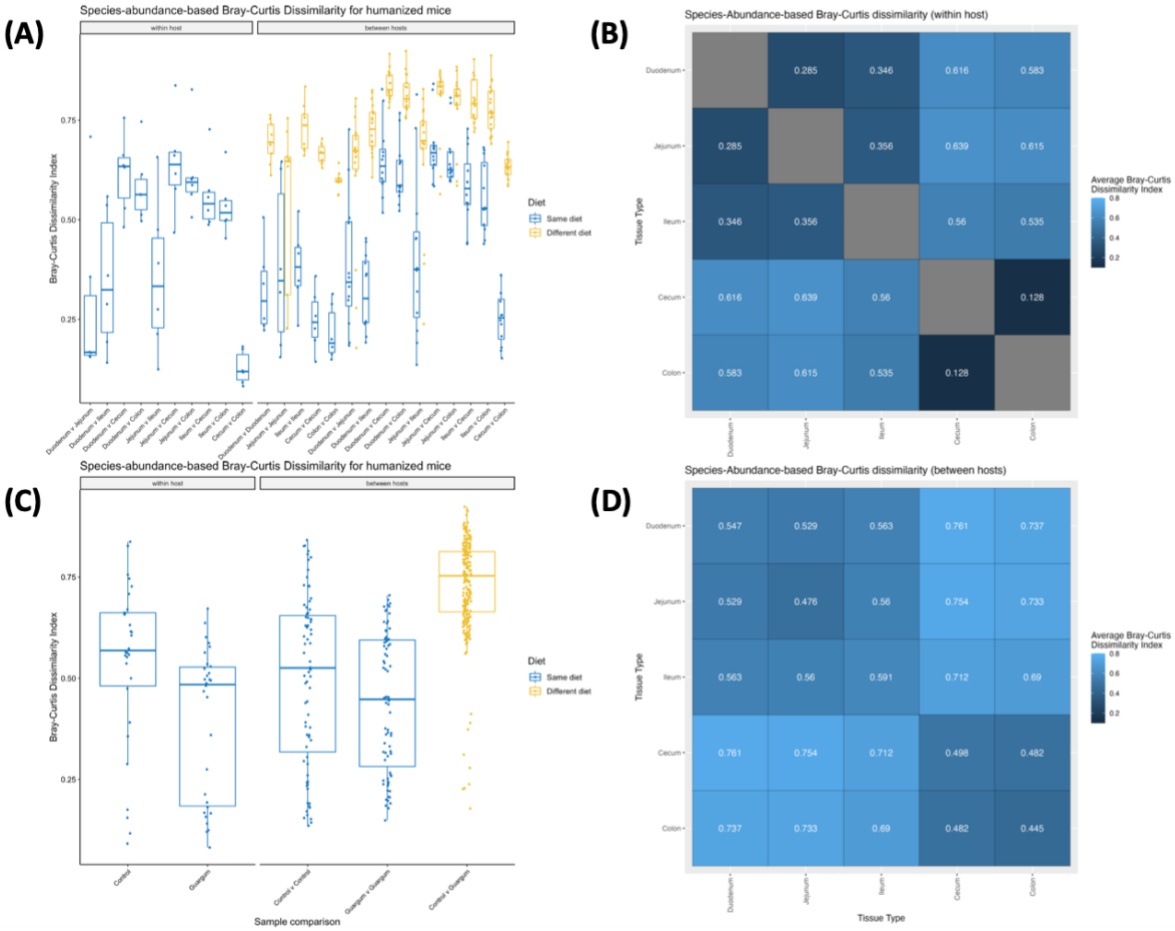
v	Tissue	Paired-end reads per sample
Mouse 1	Duodenum	34,126,263
	Jejunum	34,729,946
	Ileum	45,325,529
	Cecum	33,057,592
	Colon	32,374,899
Mouse 2	Duodenum	30,911,961
	Jejunum	34,871,836
	Ileum	34,789,365
	Cecum	32,920,833
	Colon	29,564,608
Mouse 3	Duodenum	36,008,731
	Jejunum	44,174,370
	Ileum	31,686,124
	Cecum	32,002,846
	Colon	30,546,791
Mouse 4	Duodenum	23,398,036
	Jejunum	30,787,783

	Ileum	33,305,802
	Cecum	29,389,586
	Colon	32,240,930
Mouse 5	Duodenum	33,612,764
	Jejunum	33,446,740
	Ileum	39,828,361
	Cecum	31,864,259
	Colon	30,393,912
Mouse 6	Duodenum	23,235,658
	Jejunum	31,781,992
	Ileum	51,606,729
	Cecum	34,283,821
	Colon	33,540,305
Average		33,660,279

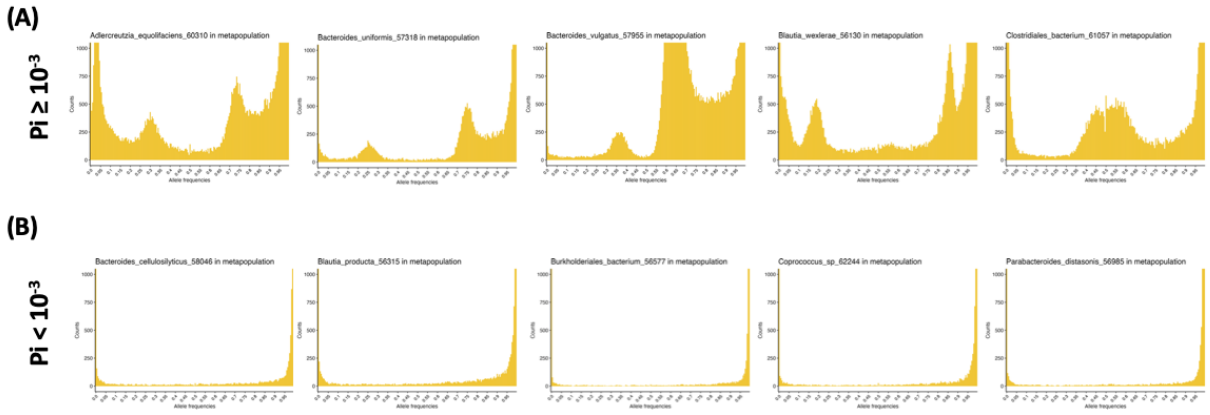
34 **Table S1 Sequencing reads per sample** The total number of raw reads are shown for each gut segment in each of the six mice.



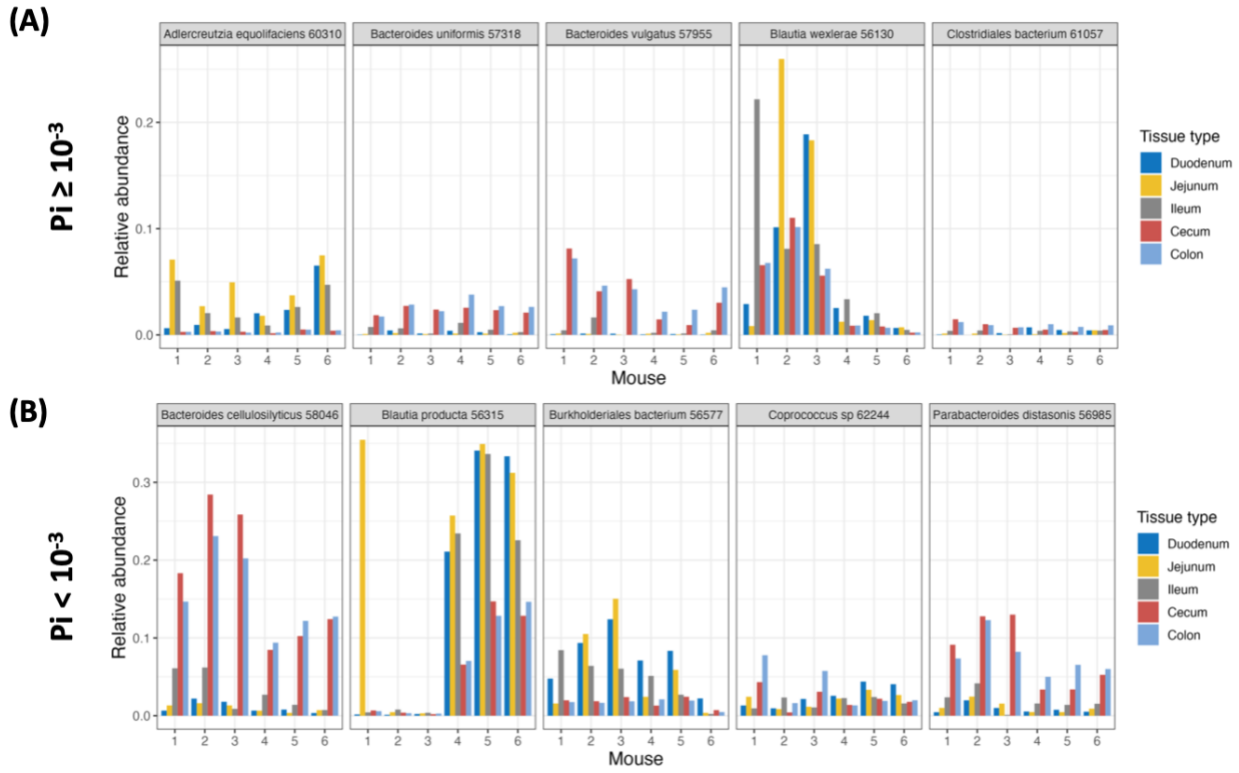
35 **Figure S1. pH and osmolality measurements at each gut segment** Each point represents the measurement at a single segment in a single host for (A) pH and (B) osmolality.



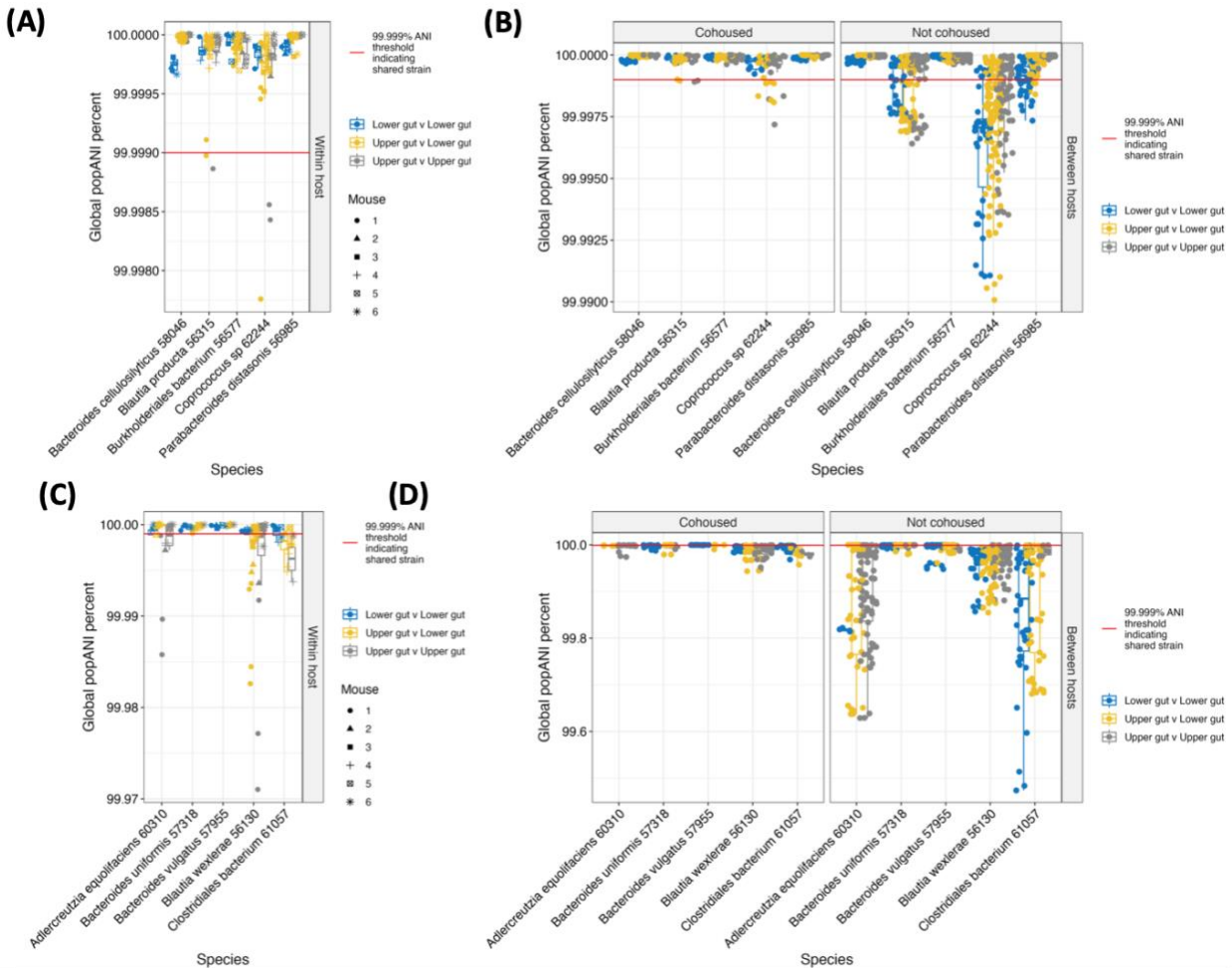
36 **Figure S2. Beta diversity of species abundance between gut segments and between hosts.** Beta diversity (Bray-Curtis dissimilarity index) was calculated between all samples using relative species abundances. In A-D, beta diversity measures are presented for both within- and between-host comparisons.



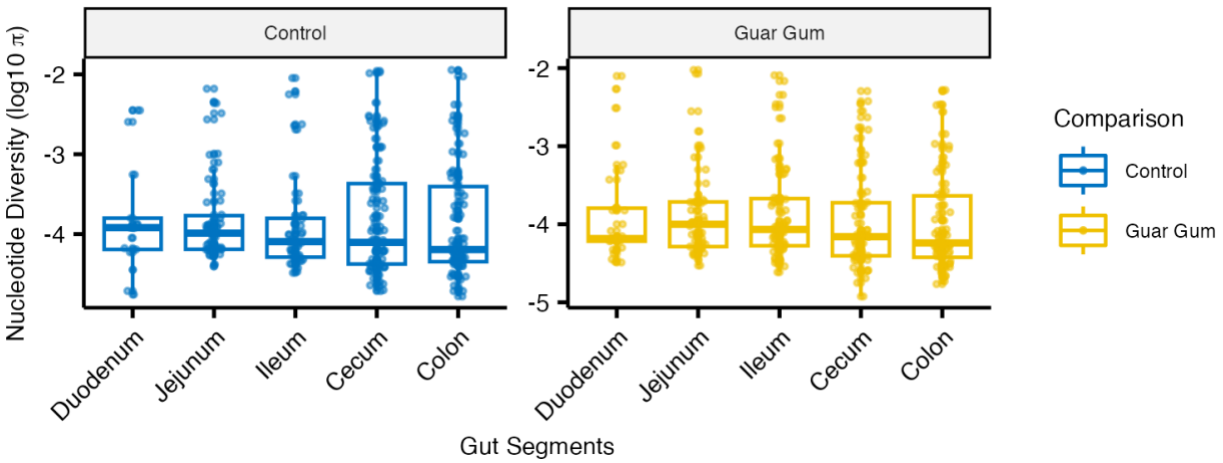
37 **Figure S3. Site frequency spectrum in the metapopulation** Shotgun sequencing data for all samples were pooled and processed in MIDAS. Genetic level data was used to plot unfolded site frequency spectra, whereby the number of sites with that fall in each allele frequency bin (200 bins, each of width 0.005) are counted. Enrichment for very low and high frequency alleles indicates the presence of only a single strain in the metapopulation, while an enrichment in intermediate frequency alleles indicates the presence of multiple strains.



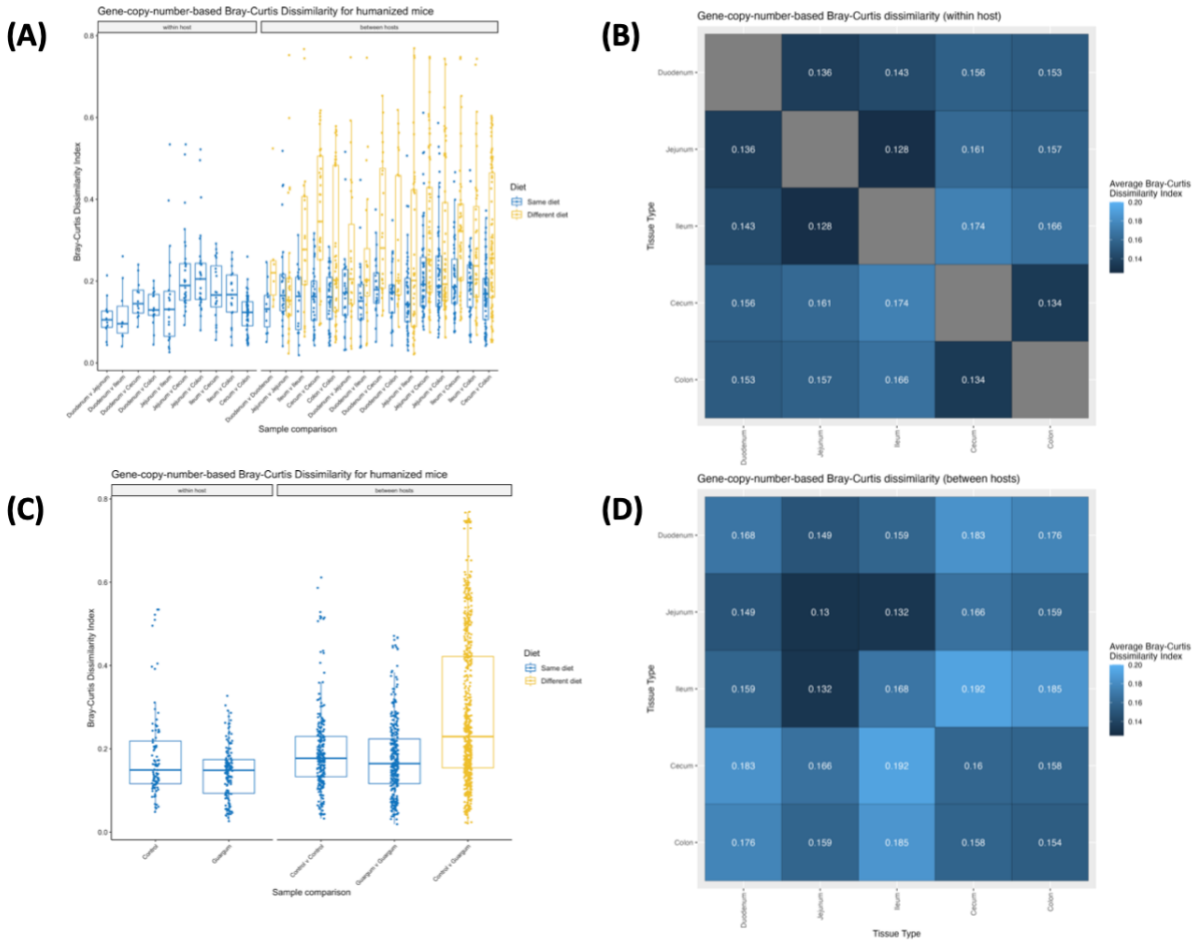
38 **Figure S4 Abundance of select high and low diversity species across hosts** Relative abundances were plotted for **(A)** species with mean between-host pairwise pi > 1×10^{-3} and **(B)** species with mean between-host pairwise pi < 1×10^{-3}.



39 **Figure S5 popANI within and across hosts** popANI distributions are plotted for species with pi $\geq 1 \times 10^{-3}$ for **(A)** within-host and **(B)** between-host comparisons, as well as for species with pi $< 1 \times 10^{-3}$ for (**C**) within-host and **(D)** between-host comparisons.

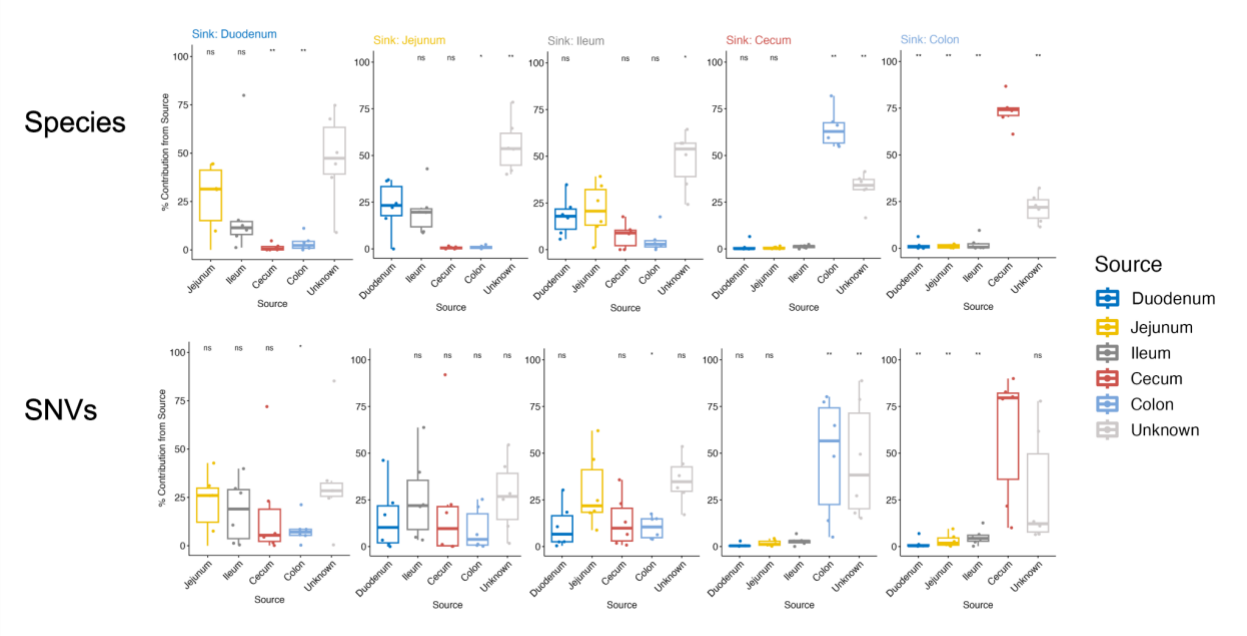


40 **Figure S6 Intra-Sample Nucleotide Diversity Along the Gut** Nucleotide diversity (pi) for each segment in each mouse on either a control or guar gum diet. Each point represents the mean pi for a given species observed in a segment sample.

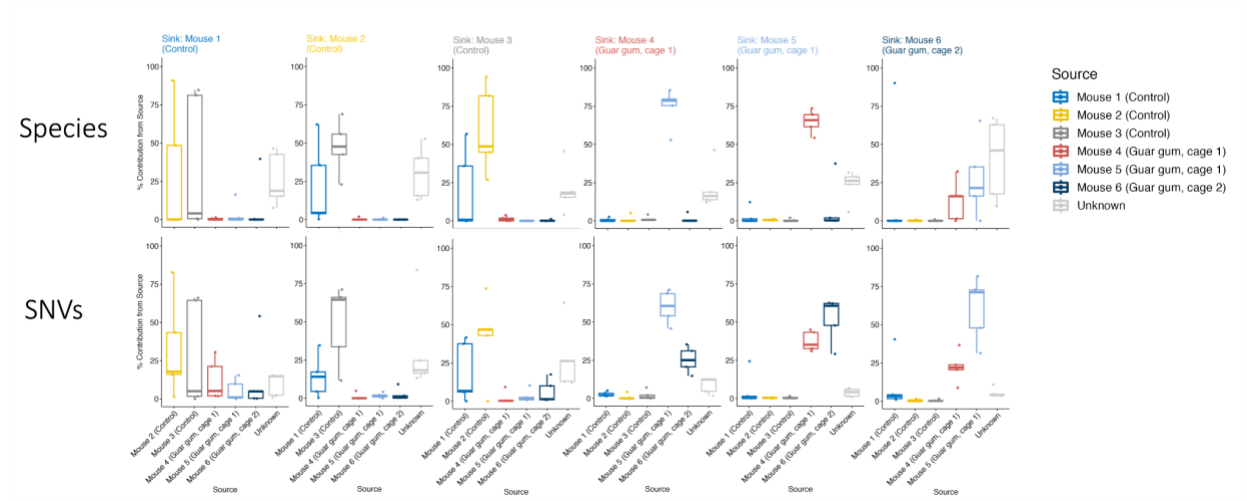


41 **Figure S7. Beta diversity of gene copy number abundance between gut segments and between hosts.** Beta diversity (Bray-Curtis dissimilarity index) was calculated between all samples using gene

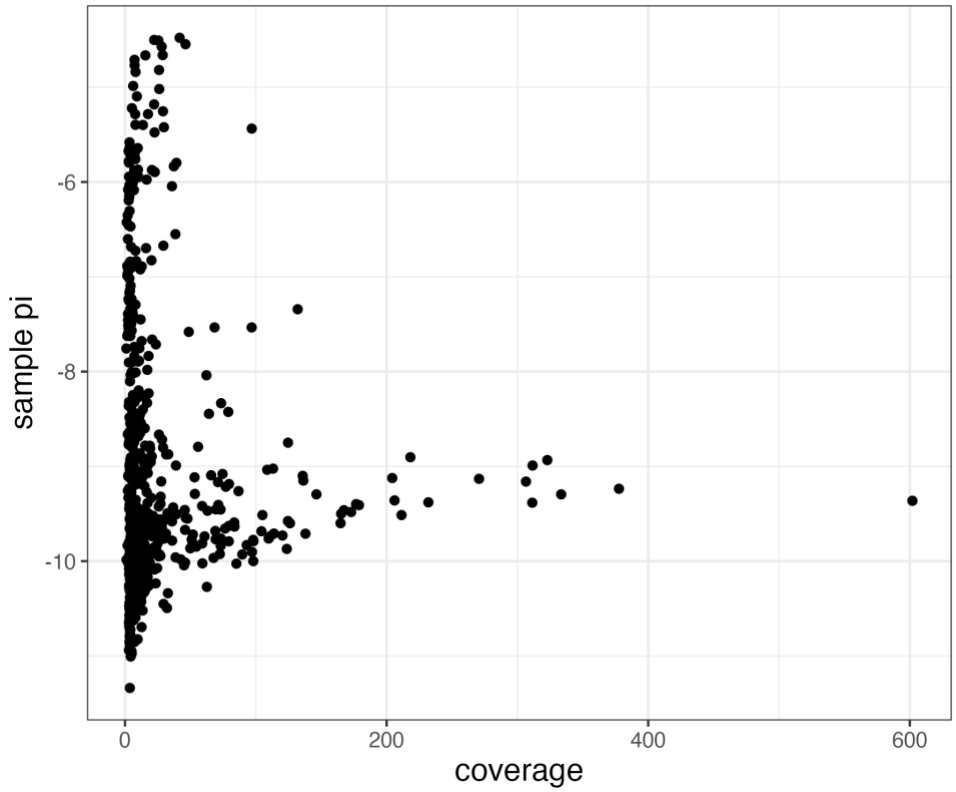
copy numbers. Gene copy number is calculated by dividing the coverage of a gene by the median coverage of 15 universal single copy genes. In A-D, beta diversity measures are presented for both within- and between-host comparisons.



42 **Figure S8 Source tracking for tissues** Within each mouse, we estimated the source contribution of each tissue to a tissue of interest. Each dot represents the source tracking experiment for one of the 6 mice



43 **Figure S9 Source tracking for mice** Each dot represents the source tracking experiment for each of the 5 tissues



44 **Figure S10 Coverage and sample pi** Each dot represents a sample. Sample coverage is plotted on the x axis and log transformed pi is plotted on the y axis.

References

1. Thomas, A. M. *et al.* Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* **25**, 667–678 (2019).

2. Ley, R. E. *et al.* Obesity alters gut microbial ecology. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 11070–11075 (2005).

3. Ley, R. E., Turnbaugh, P. J., Klein, S. & Gordon, J. I. Microbial ecology: Human gut microbes associated with obesity. *Nature* **444**, 1022–1023 (2006).

4. Jakobsson, H. E. *et al.* Short-Term Antibiotic Treatment Has Differing Long-Term Impacts on the Human Throat and Gut Microbiome. *PLoS One* **5**, e9836 (2010).

5. Jernberg, C., Löfmark, S., Edlund, C. & Jansson, J. K. Long-term impacts of antibiotic exposure on the human intestinal microbiota. *Microbiology* vol. 156 3216–3223 (2010).

6. Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A. & Alm, E. J. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat. Commun.* **8**, (2017).

7. Vujkovic-Cvijin, I. *et al.* Host variables confound gut microbiota studies of human disease. *Nature* **587**, 448–454 (2020).

8. Song, J. *et al.* Preservation Methods Differ in Fecal Microbiome Stability, Affecting Suitability for Field Studies Downloaded from. **1**, 21–37 (2016).

9. Amir, A. *et al.* Correcting for Microbial Blooms in Fecal Samples during Room-Temperature Shipping. *mSystems* **2**, (2017).

10. Lauber, C. L., Zhou, N., Gordon, J. I., Knight, R. & Fierer, N. Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples. *FEMS Microbiol. Lett.* **307**, 80–86 (2010).

11. Kim, D. *et al.* Optimizing methods and dodging pitfalls in microbiome research. *Microbiome* vol. 5 (2017).

12. Morgan, J. L., Darling, A. E. & Eisen, J. A. Metagenomic Sequencing of an In Vitro-Simulated Microbial Community. *PLoS One* **5**, e10209 (2010).

13. Costea, P. I. *et al.* Towards standards for human fecal sample processing in metagenomic studies. *Nat. Biotechnol.* **35**, 1069–1076 (2017).

14. Sipos, R. *et al.* Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. *FEMS Microbiol. Ecol.* **60**, 341–350 (2007).

15. D'Amore, R. *et al.* A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. *BMC Genomics* **17**, 55 (2016).

16. Hugerth, L. W. & Andersson, A. F. Analysing microbial community composition through amplicon sequencing: From sampling to hypothesis testing. *Frontiers in Microbiology* vol. 8 1561 (2017).

17. Sinha, R. *et al.* Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. *Nat. Biotechnol.* **35**, 1077–1086 (2017).

18. Pollock, J., Glendinning, L., Wisedchanwet, T. & Watson, M. The madness of microbiome: Attempting to find consensus 'best practice' for 16S microbiome studies. *Applied and Environmental Microbiology* vol. 84 (2018).

19. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).

20. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).

21. Fortin, J. P. *et al.* Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol.* **15**, 1–17 (2014).

22. Sun, Z. *et al.* Batch effect correction for genome-wide methylation data with Illumina Infinium platform. *BMC Med. Genomics* **4**, 1–12 (2011).

23. Xu, L., Paterson, A. D., Turpin, W. & Xu, W. Assessment and selection of competing models for zero-inflated microbiome data. *PLoS One* **10**, e0129606 (2015).

24. Kaul, A., Mandal, S., Davidov, O. & Peddada, S. D. Analysis of microbiome data in the presence of excess zeros. *Front. Microbiol.* **8**, 2114 (2017).

25. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome datasets are compositional: And this is not optional. *Frontiers in Microbiology* vol. 8 (2017).

26. Zakrzewski, M. *et al.* Calypso: a user-friendly web-server for mining and visualizing microbiome–environment interactions. *Bioinformatics* **33**, btw725 (2016).

27. Shaw, L. P. *et al.* Modelling microbiome recovery after antibiotics using a stability landscape framework. *ISME J.* **13**, 1845–1856 (2019).

28. Francino, M. P. Antibiotics and the human gut microbiome: Dysbioses and accumulation of resistances. *Frontiers in Microbiology* vol. 6 1543 (2016).

29. Qin, N. *et al.* Alterations of the human gut microbiome in liver cirrhosis. *Nature* **513**, 59–

64 (2014).

30. Bartolomaeus, T. U. P. *et al.* Quantifying technical confounders in microbiome studies. *Cardiovasc. Res.* (2020) doi:10.1093/cvr/cvaa128.

31. Gaulke, C. A. *et al.* Evaluation of the effects of library preparation procedure and sample characteristics on the accuracy of metagenomic profiles. *bioRxiv* 2021.04.12.439578 (2021) doi:10.1101/2021.04.12.439578.

32. Sacristán-Soriano, O., Banaigs, B., Casamayor, E. O. & Becerro, M. A. Exploring the links between natural products and bacterial assemblages in the sponge aplysina aerophoba. *Appl. Environ. Microbiol.* **77**, 862–870 (2011).

33. McLaren, M. R., Willis, A. D. & Callahan, B. J. Consistent and correctable bias in metagenomic sequencing experiments. *Elife* **8**, (2019).

34. Brooks, J. P. *et al.* The truth about metagenomics: Quantifying and counteracting bias in 16S rRNA studies Ecological and evolutionary microbiology. *BMC Microbiol.* **15**, (2015).

35. Nearing, J. T. *et al.* Microbiome differential abundance methods produce disturbingly different results across 38 datasets. *bioRxiv* 2021.05.10.443486 (2021) doi:10.1101/2021.05.10.443486.

36. Armour, C. R., Nayfach, S., Pollard, K. S. & Sharpton, T. J. A Metagenomic Meta-analysis Reveals Functional Signatures of Health and Disease in the Human Gut Microbiome. *mSystems* **4**, (2019).

37. Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S. & Kyrpides, N. C. New insights from uncultivated genomes of the global human gut microbiome. *Nature* **568**, 505–510 (2019).

38. Gibbons, S. M., Duvallet, C. & Alm, E. J. Correcting for batch effects in case-control microbiome studies. *PLoS Comput. Biol.* **14**, (2018).

39. Su, X. *et al.* Multiple-Disease Detection and Classification across Cohorts via Microbiome Search. *mSystems* **5**, (2020).

40. Wang, Y. & Lê Cao, K.-A. Managing batch effects in microbiome data. *Brief. Bioinform.* **2019**, 1–17 (2019).

41. Pasolli, E., Truong, D. T., Malik, F., Waldron, L. & Segata, N. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. *PLOS Comput. Biol.* **12**, e1004977 (2016).

42. Asnicar, F. *et al.* Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. *Nat. Med.* (2021) doi:10.1038/s41591-020-01183-8.

43. Gibbons, S. M., Duvallet, C. & Alm, E. J. Correcting for batch effects in case-control microbiome studies. *PLOS Comput. Biol.* **14**, e1006102 (2018).

44. McDonald, D. *et al.* American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems* **3**, (2018).

45. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol. 2014 1512* **15**, 1–21 (2014).

46. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–40 (2010).

47. Lloréns-Rico, V., Vieira-Silva, S., Gonçalves, P. J., Falony, G. & Raes, J. Benchmarking microbiome transformations favors experimental quantitative approaches to address compositionality and sampling depth biases. *Nat. Commun. 2021 121* **12**, 1–12 (2021).

48. Pan, A. Y. Statistical analysis of microbiome data: The challenge of sparsity. *Curr. Opin. Endocr. Metab. Res.* **19**, 35–40 (2021).

49. Stoffel, M. A. *et al.* Early sexual dimorphism in the developing gut microbiome of northern elephant seals. *Mol. Ecol.* **29**, 2109–2122 (2020).

50. Espinosa-Gongora, C., Larsen, N., Schønning, K., Fredholm, M. & Guardabassi, L. Differential Analysis of the Nasal Microbiome of Pig Carriers or Non-Carriers of Staphylococcus aureus. *PLoS One* **11**, e0160331 (2016).

51. Kim, K. J., Park, J., Park, S. C. & Won, S. Phylogenetic tree-based microbiome association test. *Bioinformatics* **36**, 1000–1006 (2020).

52. Poore, G. D. *et al.* Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nat. 2020 5797800* **579**, 567–574 (2020).

53. Morgan, X. C. *et al.* Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. *Genome Biol.* **16**, 67 (2015).

54. Pérez-Jaramillo, J. E. *et al.* Linking rhizosphere microbiome composition of wild and domesticated Phaseolus vulgaris to genotypic and root phenotypic traits. *ISME J.* **11**, 2244–2257 (2017).

55. Kim, H. J., Li, H., Collins, J. J. & Ingber, D. E. Contributions of microbiome and mechanical deformation to intestinal bacterial overgrowth and inflammation in a human gut-on-a-chip. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E7–E15 (2016).

56. Nayfach, S. & Pollard, K. S. Toward Accurate and Quantitative Comparative Metagenomics. *Cell* vol. 166 1103–1116 (2016).

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

58. Mallick, H. *et al.* Experimental design and quantitative analysis of microbial community multiomics. *Genome Biology* vol. 18 1–16 (2017).

59. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3**, 1724–1735 (2007).

60. Gagnon-Bartsch, J. A., Jacob, L. & Speed, T. P. *Removing Unwanted Variation from High Dimensional Data with Negative Controls*. (2013).

61. Rahmani, E. *et al.* Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nat. Methods* **13**, 443–445 (2016).

62. Wang, Y. & Cao, K.-A. L. A multivariate method to correct for batch effects in microbiome data. *bioRxiv* 2020.10.27.358283 (2020) doi:10.1101/2020.10.27.358283.

63. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* vol. 11 459–463 (2010).

64. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).

65. Tucker, G., Price, A. L. & Berger, B. Improving the power of GWAS and avoiding confounding from population stratification with PC-select. *Genetics* vol. 197 1045–1049 (2014).

66. Berner, D., Adams, D. C., Grandchamp, A. C. & Hendry, A. P. Natural selection drives patterns of lake-stream divergence in stickleback foraging morphology. *J. Evol. Biol.* **21**, 1653–1665 (2008).

67. Blekhman, R. *et al.* Host genetic variation impacts microbiome composition across human body sites. *Genome Biol.* **16**, 191 (2015).

68. Ghannoum, M. A. *et al.* Characterization of the Oral Fungal Microbiome (Mycobiome) in Healthy Individuals. *PLoS Pathog.* **6**, e1000713 (2010).

69. Shan, X. & Cordero, O. Deconstructing the association between abiotic factors and species assemblages in the global ocean microbiome. *bioRxiv* 2020.03.12.989426 (2020) doi:10.1101/2020.03.12.989426.

70. Sims, A. H. *et al.* The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets – improving meta-analysis and prediction of prognosis. *BMC Med. Genomics* **1**, 42 (2008).

71. Ritchie, M. E. *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

72. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).

73. Fernandes, A. D. *et al.* 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**, 1–13 (2014).

74. Martino, C. *et al.* A Novel Sparse Compositional Technique Reveals Microbial Perturbations. *mSystems* **4**, (2019).

75. Shi, P., Zhang, A. & Li, H. Regression analysis for microbiome compositional data. *Ann. Appl. Stat.* **10**, 1019–1040 (2016).

76. Lê Cao, K.-A. *et al.* MixMC: A Multivariate Statistical Framework to Gain Insight into Microbial Communities. *PLoS One* **11**, e0160169 (2016).

77. van den Boogaart, K. G. & Tolosana-Delgado, R. *Analyzing compositional data with R*. *Analyzing Compositional Data with R* (Springer Berlin Heidelberg, 2013). doi:10.1007/978-3-642-36809-7.

78. Wirbel, J. *et al.* Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* **25**, 679–689 (2019).

79. Kaplan, R. C. *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.* **20**, 219 (2019).

80. Baxter, N. T., Ruffin, M. T., Rogers, M. A. M. & Schloss, P. D. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Med.* **8**, 37 (2016).

81. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10**, 766 (2014).

82. Zackular, J. P., Rogers, M. A. M., Ruffin, M. T. & Schloss, P. D. The human gut microbiome as a screening tool for colorectal cancer. *Cancer Prev. Res.* **7**, 1112–1121 (2014).

83. Feng, Q. *et al.* Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nat. Commun.* **6**, 1–13 (2015).

84. Vogtmann, E. *et al.* Colorectal Cancer and the Human Gut Microbiome: Reproducibility with Whole-Genome Shotgun Sequencing. *PLoS One* **11**, e0155362 (2016).

85. Yu, J. *et al.* Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* **66**, 70–78 (2017).

86. Hannigan, G. D., Duhaime, M. B., Ruffin, M. T., Koumpouras, C. C. & Schloss, P. D. Diagnostic potential and interactive dynamics of the colorectal cancer virome. *MBio* **9**, (2018).

87. Sze, M. A. & Schloss, P. D. Looking for a signal in the noise: Revisiting obesity and the microbiome. *MBio* **7**, (2016).

88. Ross, E. M., Moate, P. J., Marett, L. C., Cocks, B. G. & Hayes, B. J. Metagenomic Predictions: From Microbiome to Complex Health and Environmental Phenotypes in Humans and Cattle. *PLoS One* **8**, e73056 (2013).

89. Liu, W., Fang, X., Zhou, Y., Dou, L. & Dou, T. Machine learning-based investigation of the relationship between gut microbiome and obesity status. *Microbes Infect.* 104892 (2021) doi:10.1016/J.MICINF.2021.104892.

90. Susin, A., Wang, Y., Lê Cao, K.-A. & Calle, M. L. Variable selection in microbiome compositional data analysis. *NAR Genomics Bioinforma.* **2**, (2020).

91. Pawlowsky-Glahn, V. & Buccianti, A. *Compositional Data Analysis Theory and Applications Edited by*. https://books.google.com/books?hl=en&lr=&id=Ggpj3QeDoKQC&oi=fnd&pg=PT17&dq=compositional+data+analysis&ots=cKF9nDohOe&sig=_0YoTuX3_hGjJV003e9Q_In5M-o (2011).

92. Aitchison, J. Principles of compositional data analysis. in 73–81 (Institute of Mathematical Statistics, 1994). doi:10.1214/lnms/1215463786.

93. Skums, P. *et al.* Efficient error correction for next-generation sequencing of viral amplicons. *BMC Bioinformatics* **13 Suppl 10**, S6 (2012).

94. Martin, J. *et al.* Rnnotator: An automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC Genomics* **11**, 663 (2010).

95. Carvalho, A. B., Dupim, E. G. & Goldstein, G. Improved assembly of noisy long reads by k-mer validation. *Genome Res.* **26**, 1710–1720 (2016).

96. Ross, E. M., Moate, P. J., Marett, L. C., Cocks, B. G. & Hayes, B. J. Metagenomic Predictions: From Microbiome to Complex Health and Environmental Phenotypes in Humans and Cattle. *PLoS One* **8**, e73056 (2013).

97. Garud, N. R. & Pollard, K. S. Population Genetics in the Human Microbiome. *Trends in Genetics* vol. 36 53–67 (2020).

98. Wang, J. & Jia, H. Metagenome-wide association studies: Fine-mining the microbiome. *Nature Reviews Microbiology* vol. 14 508–522 (2016).

99. Chen, P. E. & Shapiro, B. J. The advent of genome-wide association studies for bacteria. *Current Opinion in Microbiology* vol. 25 17–24 (2015).

100. Martín-Fernández, J. A., Pawlowsky-Glahn, V., Egozcue, J. J. & Tolosona-Delgado, R. Advances in Principal Balances for Compositional Data. *Math. Geosci. 2017 503* **50**, 273–298 (2017).

101. Filzmoser, P. & Hron, K. Correlation Analysis for Compositional Data. *Math Geosci* **41**, 905–919 (2009).

102. Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods* vol. 12 902–903 (2015).

103. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).

104. Hoffman, G. E. & Schadt, E. E. variancePartition: Interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics* **17**, 483 (2016).

105. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

106. Loomba, R. *et al.* Gut Microbiome-Based Metagenomic Signature for Non-invasive Detection of Advanced Fibrosis in Human Nonalcoholic Fatty Liver Disease. *Cell Metab.* **25**, 1054-1062.e5 (2017).

107. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).

108. Knights, D. *et al.* Bayesian community-wide culture-independent microbial source tracking. *Nat. Methods* **8**, 761–765 (2011).

109. Shenhav, L. *et al.* FEAST: fast expectation-maximization for microbial source tracking. *Nat. Methods* **16**, (2019).

110. Nayfach, S., Rodriguez-Mueller, B., Garud, N. & Pollard, K. S. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res.* **26**, 1612–1625 (2016).

111. Asnicar, F. *et al.* Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling. *mSystems* **2**, (2017).

112. Ferretti, P. *et al.* Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. *Cell Host Microbe* **24**, 133-145.e5 (2018).

113. Olm, M. R. *et al.* inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat. Biotechnol. 2021 396* **39**, 727–736 (2021).

114. Korpela, K. *et al.* Selective maternal seeding and environment shape the human gut microbiome. *Genome Res.* **28**, 561–568 (2018).

115. Li, S. S. *et al.* Durable coexistence of donor and recipient strains after fecal microbiota transplantation. *Science (80-.).* **352**, 586–589 (2016).

116. Schmidt, T. S. B. *et al.* Extensive transmission of microbes along the gastrointestinal tract. *Elife* **8**, (2019).

117. Schloissnig, S. *et al.* Genomic variation landscape of the human gut microbiome. *Nature* (2013) doi:10.1038/nature11711.

118. Bäckhed, F. *et al.* Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host Microbe* **17**, 690–703 (2015).

119. Yassour, M. *et al.* Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life. *Cell Host Microbe* **24**, 146-154.e4 (2018).

120. Sprockett, D. D. *et al.* Microbiota assembly, structure, and dynamics among Tsimane horticulturalists of the Bolivian Amazon. *Nat. Commun. 2020 111* **11**, 1–14 (2020).

121. Sloan, W. T. *et al.* Quantifying the roles of immigration and chance in shaping prokaryote community structure. *Environ. Microbiol.* **8**, 732–740 (2006).

122. Sloan, W. T., Woodcock, S., Lunn, M., Head, I. M. & Curtis, T. P. Modeling taxa-abundance distributions in microbial communities using environmental sequence data. *Microb. Ecol.* **53**, 443–455 (2007).

123. Chen, E. Z. & Li, H. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics* **32**, 2611–2617 (2016).

124. Martin, B. D., Witten, D. & Willis, A. D. MODELING MICROBIAL ABUNDANCES AND DYSBIOSIS WITH BETA-BINOMIAL REGRESSION. *Ann. Appl. Stat.* **14**, 94 (2020).

125. Consortium, T. H. M. Human Microbiome Project. https://aws.amazon.com/datasets/human-microbiome-project/ (2013).

126. Lloyd-Price, J. *et al.* Strains, functions and dynamics in the expanded Human Microbiome Project. *Nat. 2017 5507674* **550**, 61–66 (2017).

127. Brooks, B. *et al.* Strain-resolved analysis of hospital rooms and infants reveals overlap between the human and room microbiome. *Nat. Commun.* **8**, 1–7 (2017).

128. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science (80-.).* **348**, (2015).

129. Ladau, J. *et al.* Global marine bacterial diversity peaks at high latitudes in winter. *ISME J. 2013 79* **7**, 1669–1677 (2013).

130. Cavalli-Sforza, L. L. & Feldman, M. W. The application of molecular genetic approaches to the study of human evolution. *Nat. Genet. 2003 333* **33**, 266–275 (2003).

131. DeGiorgio, M. & Rosenberg, N. A. Geographic Sampling Scheme as a Determinant of the

Major Axis of Genetic Variation in Principal Components Analysis. *Mol. Biol. Evol.* **30**, 480–488 (2013).

132. Golani, D. Distribution of Lessepsian migrant fish in the Mediterranean. *http://dx.doi.org/10.1080/11250009809386801* **65**, 95–99 (2009).

133. Bentur, Y. *et al.* Lessepsian migration and tetrodotoxin poisoning due to Lagocephalus sceleratus in the eastern Mediterranean. *Toxicon* **52**, 964–968 (2008).

134. Bianchi, C. N. & Morri, C. Global sea warming and "tropicalization" of the Mediterranean Sea: biogeographic and ecological aspects. *Biogeogr. – J. Integr. Biogeogr.* **24**, (2003).

135. Elsaeed, E., Fahmy, N., Hanora, A. & Enany, S. Bacterial Taxa Migrating from the Mediterranean Sea into the Red Sea Revealed a Higher Prevalence of Anti-Lessepsian Migrations. *Omi. A J. Integr. Biol.* **25**, 60–71 (2021).

136. Antunes, A., Ngugi, D. K. & Stingl, U. Microbiology of the Red Sea (and other) deep-sea anoxic brine lakes. *Environ. Microbiol. Rep.* **3**, 416–433 (2011).

137. Flores, G. E. *et al.* Microbial Biogeography of Public Restroom Surfaces. *PLoS One* **6**, e28132 (2011).

138. McGhee, J. J. *et al.* Meta-SourceTracker: application of Bayesian source tracking to shotgun metagenomics. *PeerJ* **8**, e8783 (2020).

139. Austin, G. I. *et al.* Contamination source modeling with SCRuB improves cancer phenotype prediction from microbiome data. *Nat. Biotechnol. 2023* **4**, 1–9 (2023).

140. Dlugosch, L. *et al.* Significance of gene variants for the functional biogeography of the near-surface Atlantic Ocean microbiome. *Nat. Commun. 2022 131* **13**, 1–11 (2022).

141. Delmont, T. O. *et al.* Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade. *Elife* **8**, (2019).

142. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).

143. Chiu, A. M., Molloy, E. K., Tan, Z., Talwalkar, A. & Sankararaman, S. Inferring population structure in biobank-scale genomic data. *Am. J. Hum. Genet.* (2022) doi:10.1016/J.AJHG.2022.02.015.

144. Siranosian, B. A., Tamburini, F. B., Sherlock, G. & Bhatt, A. S. Acquisition, transmission and strain diversity of human gut-colonizing crAss-like phages. *Nat. Commun. 2020 111* **11**, 1–11 (2020).

145. Vatanen, T., Jabbar, K. S., Vlamakis, H., Knip, M. & Correspondence, R. J. X. Mobile genetic elements from the maternal microbiome shape infant gut microbial assembly and metabolism. *Cell* **185**, 4921-4936.e15 (2022).

146. Chen, D. W. & Garud, N. R. Rapid evolution and strain turnover in the infant gut microbiome. *Genome Res.* **32**, 1124–1136 (2022).

147. Katsanevakis, S. *et al.* Invading the Mediterranean Sea: Biodiversity patterns shaped by human activities. *Front. Mar. Sci.* **1**, 32 (2014).

148. Shi, Z. J., Dimitrov, B., Zhao, C., Nayfach, S. & Pollard, K. S. Fast and accurate metagenotyping of the human gut microbiome with GT-Pro. *Nat. Biotechnol. 2021 404* **40**, 507–516 (2021).

149. Consortium, T. H. M. A framework for human microbiome research. *Nat. 2012 4867402* **486**, 215–221 (2012).

150. Garud, N. R., Good, B. H., Hallatschek, O. & Pollard, K. S. Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. *PLoS Biol.* **17**, e3000102 (2019).

151. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods 2012 94* **9**, 357–359 (2012).

152. Hijmans, R. J. *et al.* Package 'geosphere'. (2021) doi:10.1007/s00190012.

153. CRAN - Package ggpubr. https://cran.r-project.org/web/packages/ggpubr/index.html.

154. Briscoe, Leah; Halperin, Eran; Garud, N. Signature-SNVs. *PyPi* (2023).

155. Briscoe, Leah; Halperin, Eran; Garud, N. Signature-SNVs. *Github* https://github.com/garudlab/Signature-SNVs (2023).

156. Briscoe, Leah; Halperin, Eran; Garud, N. Signature-SNVs. *Zenodo* (2023) doi:10.5281/zenodo.7515044.

157. Brooks, B. *et al.* Microbes in the neonatal intensive care unit resemble those found in the gut of premature infants. *Microbiome* **2**, 1–16 (2014).

158. Nishihara, Y. *et al.* Mucosa-associated gut microbiota reflects clinical course of ulcerative colitis. *Sci. Reports* | **11**, 13743 (123AD).

159. Ghazi, A. R. *et al.* High-sensitivity pattern discovery in large, paired multiomic datasets. *Bioinformatics* **38**, i378–i385 (2022).

160. Scher, J. U. *et al.* Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. *Elife* **2013**, (2013).

161. Rinninella, E. *et al.* What is the Healthy Gut Microbiota Composition? A Changing Ecosystem across Age, Environment, Diet, and Diseases. *Microorg. 2019, Vol. 7, Page 14* **7**, 14 (2019).

162. Miyauchi, E., Shimokawa, C., Steimle, A., Desai, M. S. & Ohno, H. The impact of the gut microbiome on extra-intestinal autoimmune diseases. *Nat. Rev. Immunol.* **23**, 9–23 (2023).

163. Li, Q. *et al.* Implication of the gut microbiome composition of type 2 diabetic patients from northern China. *Sci. Reports 2020 101* **10**, 1–8 (2020).

164. Jie, Z. *et al.* The gut microbiome in atherosclerotic cardiovascular disease. *Nat. Commun. 2017 81* **8**, 1–12 (2017).

165. Wang, R. *et al.* Gut microbiome, liver immunology, and liver diseases. *Cell. Mol. Immunol. 2020 181* **18**, 4–17 (2020).

166. Rinninella, E. *et al.* Food Components and Dietary Habits: Keys for a Healthy Gut Microbiota Composition. *Nutr. 2019, Vol. 11, Page 2393* **11**, 2393 (2019).

167. Jones, R. C. *et al.* The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science (80-.).* **376**, (2022).

168. James, K. R. *et al.* Distinct microbial and immune niches of the human colon. *Nat. Immunol.* **21**, 343–353 (2020).

169. Vaga, S. *et al.* Compositional and functional differences of the mucosal microbiota along the intestine of healthy individuals. *Sci. Reports 2020 101* **10**, 1–12 (2020).

170. Montassier, E. *et al.* Probiotics impact the antibiotic resistance gene reservoir along the human GI tract in a person-specific and antibiotic-dependent manner. *Nat. Microbiol.* **6**, 1043–1054 (2021).

171. Tropini, C., Earle, K. A., Huang, K. C. & Sonnenburg, J. L. The Gut Microbiome: Connecting Spatial Organization to Function. *Cell Host Microbe* **21**, 433–442 (2017).

172. Turnbaugh, P. J. *et al.* The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Sci. Transl. Med.* **1**, (2009).

173. Anders, J. L. *et al.* Comparing the gut microbiome along the gastrointestinal tract of three sympatric species of wild rodents. *Sci. Reports 2021 111* **11**, 1–12 (2021).

174. Dutch, R., Tell, L. A., Bandivadekar, R. & Vannette, R. L. Microbiome composition of Anna's hummingbirds differs among regions of the gastrointestinal tract. *J. Avian Biol.* **2022**, e02856 (2022).

175. Schlomann, B. H. & Parthasarathy, R. Gut bacterial aggregates as living gels. *Elife* **10**, (2021).

176. Shkoporov, A. N. *et al.* Viral biogeography of the mammalian gut and parenchymal organs. *Nat. Microbiol. 2022 78* **7**, 1301–1311 (2022).

177. Kordahi, M. C. *et al.* Genomic and functional characterization of a mucosal symbiont involved in early-stage colorectal cancer. *Cell Host Microbe* **29**, 1589-1598.e6 (2021).

178. Amos, G. C. A. *et al.* Exploring how microbiome signatures change across inflammatory

bowel disease conditions and disease locations. *Sci. Reports 2021 111* **11**, 1–9 (2021).

179. Atreya, R. & Siegmund, B. Location is important: differentiation between ileal and colonic Crohn's disease. *Nat. Rev. Gastroenterol. Hepatol. 2021 188* **18**, 544–558 (2021).

180. Ryan, F. J. *et al.* Colonic microbiota is associated with inflammation and host epigenomic alterations in inflammatory bowel disease. *Nat. Commun.* **11**, (2020).

181. Mas-Lloret, J. *et al.* Gut microbiome diversity detected by high-coverage 16S and shotgun sequencing of paired stool and colon sample. *Sci. Data 2020 71* **7**, 1–13 (2020).

182. Zmora, N. *et al.* Personalized Gut Mucosal Colonization Resistance to Empiric Probiotics Is Associated with Unique Host and Microbiome Features. *Cell* **174**, 1388-1405.e21 (2018).

183. Schmidt, F. *et al.* Noninvasive assessment of gut function using transcriptional recording sentinel cells. *Science (80-.).* **376**, (2022).

184. Yang, Y. *et al.* Within-host evolution of a gut pathobiont facilitates liver translocation. *Nat. 2022 6077919* **607**, 563–570 (2022).

185. Shalon, D. *et al.* Profiling the human intestinal environment under physiological conditions. doi:10.1038/s41586-023-05989-7.

186. Kashyap, P. C. *et al.* Complex Interactions Among Diet, Gastrointestinal Transit, and Gut Microbiota in Humanized Mice. *Gastroenterology* **144**, 967 (2013).

187. Fettig, N. M. *et al.* Inhibition of Th1 activation and differentiation by dietary guar gum ameliorates experimental autoimmune encephalomyelitis. doi:10.1016/j.celrep.2022.111328.

188. Ohashi, Y. *et al.* Consumption of partially hydrolysed guar gum stimulates Bifidobacteria and butyrate-producing bacteria in the human large intestine. *https://doi.org/10.3920/BM2014.0118* **6**, 451–455 (2015).

189. Ng, K. M. *et al.* Single-strain behavior predicts responses to environmental pH and osmolality in the gut microbiota. *bioRxiv* 2022.08.31.505752 (2022) doi:10.1101/2022.08.31.505752.

190. Sung, W., Ackerman, M. S., Miller, S. F., Doak, T. G. & Lynch, M. Drift-barrier hypothesis and mutation-rate evolution. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 18488–18492 (2012).

191. Poulsen, L. K., Licht, T. R., Rang, C., Krogfelt, K. A. & Molin, S. Physiological state of Escherichia coli BJ4 growing in the large intestines of streptomycin-treated mice. *J. Bacteriol.* **177**, 5840–5845 (1995).

192. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level

population structure and genetic diversity from metagenomes. *Genome Res.* **27**, 626–638 (2017).

193. Briscoe, L., Halperin, E. & Garud, N. R. SNV-FEAST: microbial source tracking with single nucleotide variants. *Genome Biol. 2023 241* **24**, 1–23 (2023).

194. Lax, S. *et al.* Forensic analysis of the microbiome of phones and shoes. *Microbiome* **3**, 21 (2015).

195. Begun, D. J. *et al.* Population Genomics: Whole-Genome Analysis of Polymorphism and Divergence in Drosophila simulans. *PLOS Biol.* **5**, e310 (2007).

196. Bhatia, G., Patterson, N., Sankararaman, S. & Price, A. L. Estimating and interpreting FST: The impact of rare variants. *Genome Res.* **23**, 1514–1521 (2013).

197. Hudson, R. R., Slatkin, M. & Maddison, W. P. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**, 583–589 (1992).

198. Kokou, F. *et al.* Core gut microbial communities are maintained by beneficial interactions and strain variability in fish. *Nat. Microbiol. 2019 412* **4**, 2456–2465 (2019).

199. Hammer, T. J., Le, E., Martin, A. N. & Moran, N. A. The gut microbiota of bumblebees. *Insectes Sociaux 2021 684* **68**, 287–301 (2021).

200. Nuccio, E. E. *et al.* Niche differentiation is spatially and temporally regulated in the rhizosphere. *ISME J.* **14**, 999–1014 (2020).

201. Kim, J. E., Tun, H. M., Bennett, D. C., Leung, F. C. & Cheng, K. M. Microbial diversity and metabolic function in duodenum, jejunum and ileum of emu (Dromaius novaehollandiae). *Sci. Reports 2023 131* **13**, 1–18 (2023).

202. Suchodolski, J. S. Analysis of the gut microbiome in dogs and cats. *Vet. Clin. Pathol.* **50**, 6–17 (2022).

203. Vega, N. M. & Gore, J. Stochastic assembly produces heterogeneous communities in the Caenorhabditis elegans intestine. *PLoS Biol.* **15**, 1–20 (2017).

204. Danne, C., Rolhion, N. & Sokol, H. Recipient factors in faecal microbiota transplantation: one stool does not fit all. *Nat. Rev. Gastroenterol. Hepatol.* **18**, 503–513 (2021).

205. Wilson, B. C., Vatanen, T., Cutfield, W. S. & O'Sullivan, J. M. The Super-Donor Phenomenon in Fecal Microbiota Transplantation. *Front. Cell. Infect. Microbiol.* **9**, (2019).

206. Kim, S. G. *et al.* Microbiota-derived lantibiotic restores resistance against vancomycin-resistant Enterococcus. *Nat. 2019 5727771* **572**, 665–669 (2019).

207. Choi, H. H. & Cho, Y. S. Fecal Microbiota Transplantation: Current Applications,

Effectiveness, and Future Perspectives. *Clin. Endosc.* **49**, 257–265 (2016).

208. Clark, L. C. & Hodgkin, J. Commensals, probiotics and pathogens in the Caenorhabditis elegans model. *Cell. Microbiol.* **16**, 27–38 (2014).

209. Kim, Y. & Mylonakis, E. Caenorhabditis elegans immune conditioning with the probiotic bacterium Lactobacillus acidophilus strain ncfm enhances gram-positive immune responses. *Infect. Immun.* **80**, 2500–2508 (2012).

210. Dodge, R. *et al.* A symbiotic physical niche in Drosophila melanogaster regulates stable association of a multi-species gut microbiota. *Nat. Commun. 2023 141* **14**, 1–13 (2023).

211. Bittleston, L. S., Gralka, M., Leventhal, G. E., Mizrahi, I. & Cordero, O. X. Context-dependent dynamics lead to the assembly of functionally distinct microbial communities. *Nat. Commun. 2020 111* **11**, 1–10 (2020).

212. Jones, E. W., Carlson, J. M., Sivak, D. A. & Ludington, W. B. Stochastic microbiome assembly depends on context. *Proc. Natl. Acad. Sci. U. S. A.* **119**, (2022).

213. McCafferty, J. *et al.* Stochastic changes over time and not founder effects drive cage effects in microbial community assembly in a mouse model. *ISME J.* **7**, 2116–2125 (2013).

214. Valles-Colomer, M. *et al.* The person-to-person transmission landscape of the gut and oral microbiomes. *Nature* **614**, 125–135 (2023).

215. Van Rossum, T., Ferretti, P., Maistrenko, O. M. & Bork, P. Diversity within species: interpreting strains in microbiomes. *Nat. Rev. Microbiol.* **18**, 491–506 (2020).

216. Vasquez, K. S. *et al.* Quantifying rapid bacterial evolution and transmission within the mouse intestine. *Cell Host Microbe* **29**, 1454 (2021).

217. Mondragon-Palomino, O. *et al.* Three-dimensional imaging for the quantification of spatial patterns in microbiota of the intestinal mucosa. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2118483119 (2022).