UNIVERSITY OF CALIFORNIA SAN DIEGO

Tracking Microbial Communities Across Human Development in Response to Disturbance and Restoration

A dissertation submitted in partial satisfaction
of the requirements for the Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Cameron Martino

Committee in charge:

Professor Rob Knight, Chair
Professor Pieter C. Dorrestein, Co-Chair
Professor Rachel Dutton
Professor Nathan Lewis
Professor Amir Zarrinpar

2022

The dissertation of Cameron Martino is approved, and it is acceptable in quality and form for publication on microfilm and electronically

University of California San Diego

2022

DEDICATION

To my family, for their love, guidance, and all the sacrifices they made for me.

EPIGRAPH

“I, a universe of atoms, an atom in the universe.” — Richard P. Feynman

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I would like to thank my advisor Rob Knight whose generous support and guidance allowed me to grow as a scientist and as a person. It goes without saying that this work would not have been possible without him.

I would also like to thank my committee members Pieter Dorrestein, Nathan Lewis, Amir Zarrinpar, and Rachel Dutton for their feedback and collaboration during the past four years.

During this time, I would not have been able to achieve anything without the vast experience of the *Knight Laboratory* and their support. Firstly, I would like to thank those that provided the opportunity for everyone in the lab to focus on science by handling the many essential logistics. (Jerry Kennedy, Gail Ackermann, Jeff DeReus, Sarah Adams, Peggy Castaneda, Andrea Iriarte, Yna Villanueva, and Michiko Souza). Secondly, none of this work would be possible without the work of the many of scientists generating data in the lab. (Greg Humphrey, MacKenzie Bryant, Caitlin Tribelhorn, Tara Schwartz, Karenina Sanders, Helena Tubb, and Rebecca Tsai). Finally, to the team of graduate and post-doctoral researchers whose insight both inspired and served to `right-the-ship` on many occasions. (Greg Poore, Gibraan 'Gibs' Rahman, Daniel McDonald, Caitlin Guccione, George Armstrong, Justin Shaffer, Antonio González Peña, Caitlin Guccione, Amanda Hazel Dilmore, Rodolfo Antonio Salido Benítez, Celeste Allaband, Yoshiki Vazquez-Baeza, Pedro Belda-Ferre, Mehrbod Estaki, Clarisse "Lisa" Marotz, Qiyun Zhu, Jeremiah Minich, Qiyun Zhu, Se Jin Song, Lingjing Serene Jiang, Stephany Flores-Ramos, Caitriona Brennan, Kelly Fogelson, Jeffrey Chiu, Victor Cantu, Daniel Hakim).

I reached this point in my education only though constant inspiration and encouragement from others. Starting at West Valley Community College I owe much of the impetus of my academic career in science to Melvin Vaughn and Mike Staskus. Once at UC San Diego as an undergraduate, Karsten Zengler and his lab for challenging, patiently teaching, and allowing me the opportunity to freely explore microbiology. (Lívia Zaramela, Max A-Bassam, Jinu Kim, Daniela Domingos Galzerani). Then, after my undergraduate training, Mallory Embree and Mike Seely for

teaching me how to channel hard work into creativity in a practical manner, all while not letting "perfect become the enemy of the good".

Most of all I am thankful to my family. Firstly, I would like to thank my wife, Shelby for her constant selfless love, support, and encouragement in life and in the pursuit of my dreams. Secondly, my father Ray Martino, grandfather Ray "pop" Martino, and grandmother Rosemary Martino, for being my role models, providing constant support in my life. Finally, to my Mother, Lee Ann Martino who provided a wonderful childhood and foundation for my life, for which I will always be grateful. Any importance and impact of this work should be ultimately be credited to them.

Chapter 1, has been submitted for publication of the material as it may appear in *Nature Reviews Microbiology*, "Healthy microbiota succession throughout life from cradle to the grave." Cameron Martino, Amanda Hazel Dilmore, Zachary M. Burcham, Jessica L. Metcalf, Dilip Jeste, Rob Knight. The dissertation author was the primary investigator and first author of this paper.

Chapter 2, in full, is a reprint of the material as it appears in *mSystems*, "Robust Aitchison PCA reveals microbiome perturbations." Cameron Martino, James T. Morton, Clarisse A. Marotz, Luke R. Thompson, Anupriya Tripathi, Rob Knight, Karsten Zengler. The dissertation author was the primary investigator and first author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in *Nature Biotechnology*, "Context-aware dimensionality reduction deconvolutes gut microbial community dynamics." Cameron Martino, Liat Shenhav, Clarisse Marotz, George Armstrong, Daniel McDonald, Yoshiki Vázquez-Baeza, James T. Morton, Lingjing Jiang, Maria Gloria Dominguez-Bello, Austin D. Swafford, Eran Halperin, Rob Knight. The dissertation author was the primary investigator and first author of this paper.

Chapter 4, in full, is a reprint of the material as it appears in *Med*, "Naturalization of the microbiota developmental trajectory of Cesarean-born neonates after vaginal seeding." Se Jin Song, Jincheng Wang, Cameron Martino, Lingjing Jiang, Wesley K. Thompson, Liat Shenhav,

Daniel McDonald, Clarisse Marotz, Paul R. Harris, Caroll D. Hernandez, Nora Henderson, Elizabeth Ackley, Deanna Nardella, Charles Gillihan, Valentina Montacuti, William Schweizer, Melanie Jay, Joan Combellick, Haipeng Sun, Izaskun Garcia-Mantrana, Fernando Gil Raga, Maria Carmen Collado, Juana I. Rivera-Viñas, Maribel Campos-Rivera, Jean F. Ruiz-Calderon, Rob Knight, Maria Gloria Dominguez-Bello. The dissertation author was the primary investigator and first author of this paper.

VITA

2016 B.S. in Bioengineering, University of California San Diego

2022 Ph. D. in Bioinformatics and Systems Biology, University of California San Diego

PUBLICATIONS

Author names marked with † indicate shared first co-authorship.

Song SJ†, Wang J†, **Martino C**†, Jiang L, Thompson WK, Shenhav L, McDonald D, Marotz C, Harris PR, Hernandez CD, Henderson N, Ackley E, Nardella D, Gillihan C, Montacuti V, William Schweizer, Jay M, Combellick J, Sun H, Garcia-Mantrana I, Raga FG, Collado MC, Rivera-Viñas JI, Campos-Rivera M, Ruiz-Calderon JF, Knight R, Dominguez-Bello MG. 2021. Naturalization of the microbiota developmental trajectory of Cesarean-born neonates after vaginal seeding. Med 2:951–964.e5.

Martino C, Shenhav L, Marotz CA, Armstrong G, McDonald D, Vázquez-Baeza Y, Morton JT, Jiang L, Dominguez-Bello MG, Swafford AD, Halperin E, Knight R. 2021. Context-aware dimensionality reduction deconvolutes gut microbial community dynamics. Nat Biotechnol 39:165–168.

Zaramela LS†, **Martino C**†, Alisson-Silva F†, Rees SD, Diaz SL, Chuzel L, Ganatra MB, Taron CH, Secrest P, Zuñiga C, Huang J, Siegel D, Chang G, Varki A, Zengler K. 2019. Gut bacteria responding to dietary change encode sialidases that exhibit preference for red meat-associated carbohydrates. Nat Microbiol 4:2082–2089.

Martino C, Morton JT, Marotz CA, Thompson LR, Tripathi A, Knight R, Zengler K. 2019. A Novel Sparse Compositional Technique Reveals Microbial Perturbations. mSystems 4.

Armstrong G†, **Martino C**†, Morris J, Khaleghi B, Kang J, DeReus J, Zhu Q, Roush D, McDonald D, Gonazlez A, Shaffer JP, Carpenter C, Estaki M, Wandro S, Eilert S, Akel A, Eno J, Curewitz K, Swafford AD, Moshiri N, Rosing T, Knight R. 2022. Swapping Metagenomics Preprocessing Pipeline Components Offers Speed and Sensitivity Increases. mSystems e0137821.

The following publications were not included as part of this dissertation, but were also significant byproducts of my doctoral training.

Armstrong GW, Rahman G, **Martino C**, Mcdonald D, Gonzalez A, Mishne G, Knight R. Applications and comparison of dimensionality reduction methods for microbiome data. Frontiers inBioinformatics 18.

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

Hendrickson R, Urbaniak C, Minich JJ, Aronson HS, **Martino C**, Stepanauskas R, Knight R,

Venkateswaran K. 2021. Clean room microbiome complexity impacts planetary protection bioburden. Microbiome 9:238.

Allaband C, Lingaraju A, **Martino C,** Russell B, Tripathi A, Poulsen O, Machado ACD, Zhou D, Xue J, Elijah E, Malhotra A, Dorrestein PC, Knight R, Haddad GG, Zarrinpar A. 2021. Intermittent Hypoxia and Hypercapnia Alter Diurnal Rhythms of Luminal Gut Microbiome and Metabolome. mSystems https://doi.org/10.1128/msystems.00116-21.

Clausen TM, Sandoval DR, Spliid CB, Pihl J, Perrett HR, Painter CD, Narayanan A, Majowicz SA, Kwong EM, McVicar RN, Thacker BE, Glass CA, Yang Z, Torres JL, Golden GJ, Bartels PL, Porell RN, Garretson AF, Laubach L, Feldman J, Yin X, Pu Y, Hauser BM, Caradonna TM, Kellman BP, **Martino C**, Gordts PLSM, Chanda SK, Schmidt AG, Godula K, Leibel SL, Jose J, Corbett KD, Ward AB, Carlin AF, Esko JD. 2020. SARS-CoV-2 Infection Depends on Cellular Heparan Sulfate and ACE2. Cell 183:1043–1057.e15.

Marotz C, Belda-Ferre P, Ali F, Das P, Huang S, Cantrell K, Jiang L, **Martino C**, Diner RE, Rahman G, McDonald D, Armstrong G, Kodera S, Donato S, Ecklu-Mensah G, Gottel N, Salas Garcia MC, Chiang LY, Salido RA, Shaffer JP, Bryant MK, Sanders K, Humphrey G, Ackermann G, Haiminen N, Beck KL, Kim H-C, Carrieri AP, Parida L, Vázquez-Baeza Y, Torriani FJ, Knight R, Gilbert J, Sweeney DA, Allard SM. 2021. SARS-CoV-2 detection status associates with bacterial community composition in patients and the hospital environment. Microbiome 9:132.

Cantrell K, Fedarko MW, Rahman G, McDonald D, Yang Y, Zaw T, Gonzalez A, Janssen S, Estaki M, Haiminen N, Beck KL, Zhu Q, Sayyari E, Morton JT, Armstrong G, Tripathi A, Gauglitz JM, Marotz C, Matteson NL, **Martino C**, Sanders JG, Carrieri AP, Song SJ, Swafford AD, Dorrestein PC, Andersen KG, Parida L, Kim H-C, Vázquez-Baeza Y, Knight R. 2021. EMPress Enables Tree-Guided, Interactive, and Exploratory Analyses of Multi-omic Data Sets. mSystems https://doi.org/10.1128/msystems.01216-20.

Mu A, McDonald D, Jarmusch AK, **Martino C**, Brennan C, Bryant M, Humphrey GC, Toronczak J, Schwartz T, Nguyen D, Ackermann G, D'Onofrio A, Strathdee SA, Schooley RT, Dorrestein PC, Knight R, Aslam S. 2021. Assessment of the microbiome during bacteriophage therapy in combination with systemic antibiotics to treat a case of staphylococcal device infection. Microbiome 9:92.

Shaffer JP, Marotz C, Belda-Ferre P, **Martino C**, Wandro S, Estaki M, Salido RA, Carpenter CS, Zaramela LS, Minich JJ, Bryant M, Sanders K, Fraraccio S, Ackermann G, Humphrey G, Swafford AD, Miller-Montgomery S, Knight R. 2021. A comparison of DNA/RNA extraction protocols for high-throughput sequencing of microbial communities. Biotechniques 70:149–159.

Xue J, Allaband C, Zhou D, Poulsen O, **Martino C**, Jiang L, Tripathi A, Elijah E, Dorrestein PC, Knight R, Zarrinpar A, Haddad GG. 2021. Influence of Intermittent Hypoxia/Hypercapnia on Atherosclerosis, Gut Microbiome, and Metabolome. Front Physiol 12:663950.

Huey SL, Jiang L, Fedarko MW, McDonald D, **Martino C**, Ali F, Russell DG, Udipi SA, Thorat A, Thakker V, Ghugre P, Potdar RD, Chopra H, Rajagopalan K, Haas JD, Finkelstein JL, Knight R, Mehta S. 2020. Nutrition and the Gut Microbiota in 10- to 18-Month-Old Children Living in Urban Slums of Mumbai, India. mSphere 5.

Fedarko MW, **Martino C**, Morton JT, González A, Rahman G, Marotz CA, Minich JJ, Allen EE, Knight R. 2020. Visualizing'omic feature rankings and log-ratios using Qurro. NAR genomics and

bioinformatics 2:lqaa023.

Estaki M, Jiang L, Bokulich NA, McDonald D, González A, Kosciolek T, **Martino C**, Zhu Q, Birmingham A, Vázquez-Baeza Y, Dillon MR, Bolyen E, Caporaso JG, Knight R. 2020. QIIME 2 Enables Comprehensive End-to-End Analysis of Diverse Microbiome Data and Comparative Studies with Publicly Available Data. Curr Protoc Bioinformatics 70:e00010.

Sanders JG, Nurk S, Salido RA, Minich J, Xu ZZ, Zhu Q, **Martino C**, Fedarko M, Arthur TD, Chen F, Boland BS, Humphrey GC, Brennan C, Sanders K, Gaffney J, Jepsen K, Khosroheidari M, Green C, Liyanage M, Dang JW, Phelan VV, Quinn RA, Bankevich A, Chang JT, Rana TM, Conrad DJ, Sandborn WJ, Smarr L, Dorrestein PC, Pevzner PA, Knight R. 2019. Optimizing sequencing protocols for leaderboard metagenomics by combining long and short reads. Genome Biol 20:226.

Taylor BC, Lejzerowicz F, Poirel M, Shaffer JP, Jiang L, Aksenov A, Litwin N, Humphrey G, **Martino C**, Miller-Montgomery S, Dorrestein PC, Veiga P, Song SJ, McDonald D, Derrien M, Knight R. 2020. Consumption of Fermented Foods Is Associated with Systematic Differences in the Gut Microbiome and Metabolome. mSystems 5.

Fouladi F, Bailey MJ, Patterson WB, Sioda M, Blakley IC, Fodor AA, Jones RB, Chen Z, Kim JS, Lurmann F, **Martino C**, Knight R, Gilliland FD, Alderete TL. 2020. Air pollution exposure is associated with the gut microbiome as revealed by shotgun metagenomic sequencing. Environ Int 138:105604.

Bluemel S, Wang L, **Martino C**, Lee S, Wang Y, Williams B, Horvath A, Stadlbauer V, Zengler K, Schnabl B. 2018. The Role of Intestinal C-type Regenerating Islet Derived-3 Lectins for Nonalcoholic Steatohepatitis. Hepatology Communications 2:393–406.

ABSTRACT OF THE DISSERTATION

Tracking Microbial Communities Across Human Development in Response to Disturbance and Restoration

by

Cameron Martino

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2022

Professor Rob Knight, Chair

Professor Pieter Dorrestein, Co-Chair

Microbial communities play a crucial role in human health. The study of these host resident microbial communities is limited by the current inability to efficiently cultivate each member. Therefore, high throughput sequencing has been widely adopted to describe which microbial strains are present and in some cases their activity, directly from the sampled environment. The central aim of many microbiome studies is to understand how these communities form, perpetuate, and can be altered across time for the benefit of the host. However, several technical limitations in the analysis of these data have left many of these questions only partially revealed.

First, in Chapter 1 of this thesis, we review the current understanding of the human microbiome across age, from birth until death. We also highlight many of the open gaps in our knowledge of how human microbiomes are formed and sustained, in addition to the current methodologies for exploring these questions. Chapter 2, introduces the current microbiome-based dimensionality reductions, which take high-dimensional microbiome data and reduces it into a few human interpretable dimensions. In this chapter we describe the limitations of these methods in microbiome data including sparsity, nonnormality, and compositionality. We address these problems with a novel method for dimensionality reduction which uniquely handles the inherent challenges of this data type. However, microbiomes are also highly individualized with each person containing a unique set of microbial communities. To overcome this, longitudinal studies of the microbiome are growing in popularity. In chapter 3, we describe the challenges to analyze these valid study designs, in addition to the lack of methods existing to properly account for the structure. In the chapter we address these challenges through tensor-based factorization, which accounts for the structure of the study. Through this method we are able to better understand microbial community development in infants, in particular those altered through C-section rather than vaginal birth. Finally, in chapter 4, we utilize these methods to explore a method for naturalizing C-section birth through seeding of the infant with the mother's vaginal microbiome at birth. From this we found that engraftment of the mother's vaginal microbiota at birth successfully naturalizes the microbiome development.

Chapter 1. Healthy microbiota succession throughout life from cradle to the grave

Abstract

Associations between age and the human microbiome are robust and reproducible. The microbial composition at several body sites can even be used to reveal human chronological age accurately. Although it is largely unknown why specific microbes are more abundant at certain ages, human microbiome research has elucidated a series of microbial community transformations that take place between birth and death. In this review, we explore microbial succession in the healthy human microbiome from the cradle to the grave. We discuss the stages from primary succession at birth, to disruptions by disease or antibiotic use, to microbial expansion at death. We address how these successions differ by body site and by domain (bacteria, fungi, or virus). We also review experimental and analytical tools that microbiome researchers use to conduct this work. Finally, we discuss future directions for studying the microbiome's relationship with age, including integrated experimental design across studies, more robust statistical analyses, and improved characterization of non-bacterial microbes.

Themes: the microbiota genera (bacteria, fungi, and virus) succession across life

1.1. Introduction

Human-associated microbiota are communities of bacteria, fungi, and viruses (often referred to as the bacteriome, mycobiome, and virome respectively) that live on and/or inside the human body. The amount of information known about how the community structure of bacteria changes across age groups far outweighs that about the fungi and viruses, but does not

necessarily translate to bacteria being of disproportionate importance. Microbial communities exist on every mucosal surface in the human body, and each body site within a person contains a unique ecology [1–3]. Each individual's human-associated microbial community is unique compared to that of all other humans [4]. Human-resident microbes encode an estimated 2 to 20 million genes, while the human genome encodes an estimated 20 to 25 thousand; therefore, microbiota represent 99.9% of the genetic capacity in the human body [5]. During each stage of life from birth to death and decomposition, microbial communities act as a dynamic organ of the body, and have revolutionized our understanding of human biology. However, these natural and induced changes in our microbiota still harbor many mysteries waiting to be discovered and fully understood.

Microbial succession is defined as a change in the presence, relative abundance, or absolute abundance of one or more organisms within a microbial community. Microbial succession processes can be deterministic or stochastic. Factors that drive *deterministic succession* fall into three categories: abiotic (e.g. pH/redox potential [6]), biotic (e.g. cross-feeding [7], diet [8], travel [9]), and host factors (e.g. innate and adaptive immunity (reviewed in [10])). *Stochastic succession* is defined as microbial community changes that are not the consequence of environmentally determined fitness (also called *ecological drift*) [11,12]. Whether microbial succession is more deterministic or stochastic is driven by several factors in the formation of the community, including birth mode, diet (i.e. human breast milk), and antibiotics [13–15]. There are three main stages of microbial succession that naturally occur across human life during normal or healthy aging.

The first stage, ***Primary succession,*** begins at birth when pioneer species first establish the community and is followed by rapid changes in the microbial community. These changes decrease in their rate of change from birth until childhood, and many intermediate species exist

between birth and late childhood [13–15] (**Figure 1.1A**). Primary succession ends at the formation of a ***climax community***, thought to be achieved by adolescence and sustained through adulthood; this community is characterized by its relative stability [16,17] (**Figure 1.1B**). Although the microbiome is more stable in adulthood than childhood, there is still variability, fueling the debate over the existence of a climax community in the human microbiome [18]. Natural variation in the adult microbiota exists on the time scale of hours (circadian rhythms [19]) to years (aging), but microbiota are generally stable except in the presence of a disturbance such as change in diet or medications. The next stage, ***Secondary succession,*** occurs when some or all of a pre-existing stable community is altered or removed, followed by regeneration of the community to either the same or a different state. This can be done either deliberately, through medical treatments such as antibiotics [20,21] or spontaneously, through diseases such as *Vibrio cholerae* infection [22]. Secondary succession in humans is characterized by at least some period of stochastic process dominance. In induced conditions, such as a single course of antibiotics, the community follows a process similar to primary succession, where parts of the existing microbial community act as "microbial memory" and help guide back to a similar community that existed before. This process is thought to be driven by keystone community members [23,24], rather than the pioneer microbes that drive primary succession (**Figure 1.1C**). Third, ***Final succession*** is part of the natural host senescence and death. During old age, the microbial community again succeeds to a community composed of fewer total members, dominated by Proteobacteria [25], and the rate of change increases, almost in an inverse relationship to primary succession [26] (**Figure 1.1D**).

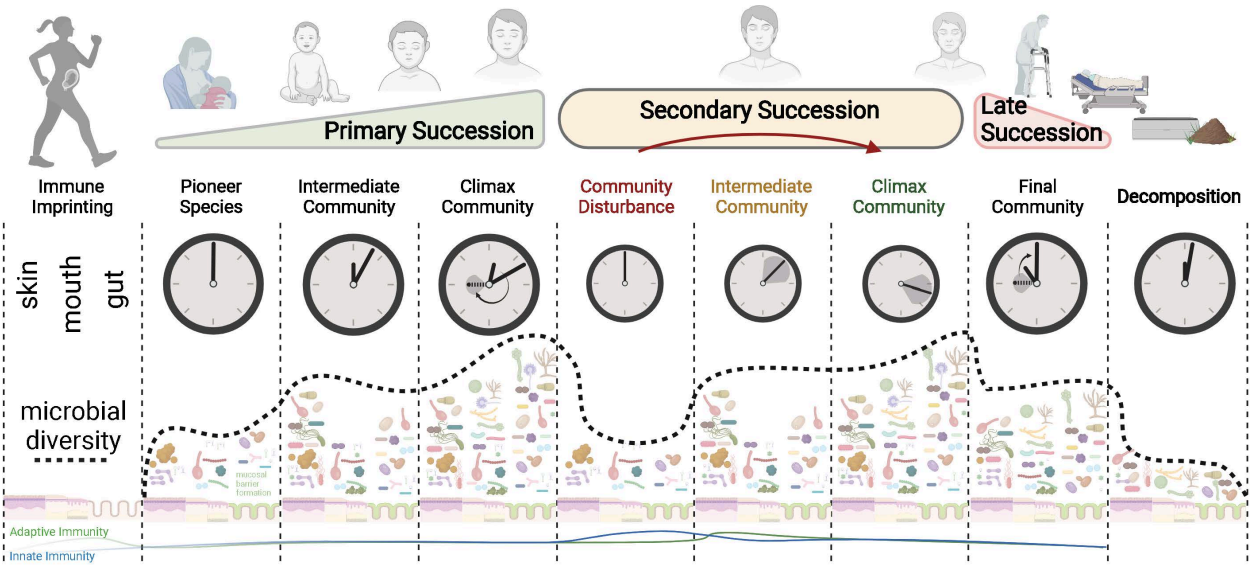

Figure 1.1. The succession of the human microbiota from conception to death. Bacteria, fungi, and viral diversity across human life stages (black dotted line). The analog clock represents the relative time of host age at which each microbial community stage develops. Immune imprinting begins before birth through the mother's microbiota and its metabolites (first column). Initial colonization of pioneer species begins at birth and body site-specific microbial communities emerge (second column). These communities continue to increase in complexity until they reach a stable community structure (third and fourth columns). Secondary successions of these microbial communities can occur from internal and external perturbations (fifth column). Intermediate species of microbes re-establish the initial community and reach a steady-state again (sixth and seventh columns). At late age, the community goes through a final succession and changes as the host nears natural death (eighth column). The last stage of microbial succession occurs at putrefaction and decomposition. During this stage diversity further declines and during the first 24-48 hours many of the human microbiome structures are conserved, but then quickly begin to erode (final column). The relative strength of adaptive (green line) and innate (blue blue) across different stages of life and microbial succession (bottom). Created with BioRender.com.

Unlike the human genome which is encoded at birth, and cannot be altered during life (at least with current technology), each of these unique microbiome changes can be deliberately modified across time. Within a host species and a body site, age has the strongest relationship with the healthy microbiome of any physiological or demographic variable measured to date [16]. Age drives both alpha and beta diversity in human microbiomes (**Figure 1.2**; **see Box 1 for a description of methods for exploring microbial communities**). Studying each stage of succession allows researchers to try to understand how human-associated microbial communities are formed and maintained. By understanding these processes, we may better understand how to manage microbiota as we age and in relation to human health. Although methodology to measure and describe microbial communities is an area of active development, standard practices do exist and are useful for integrating results across cohorts.

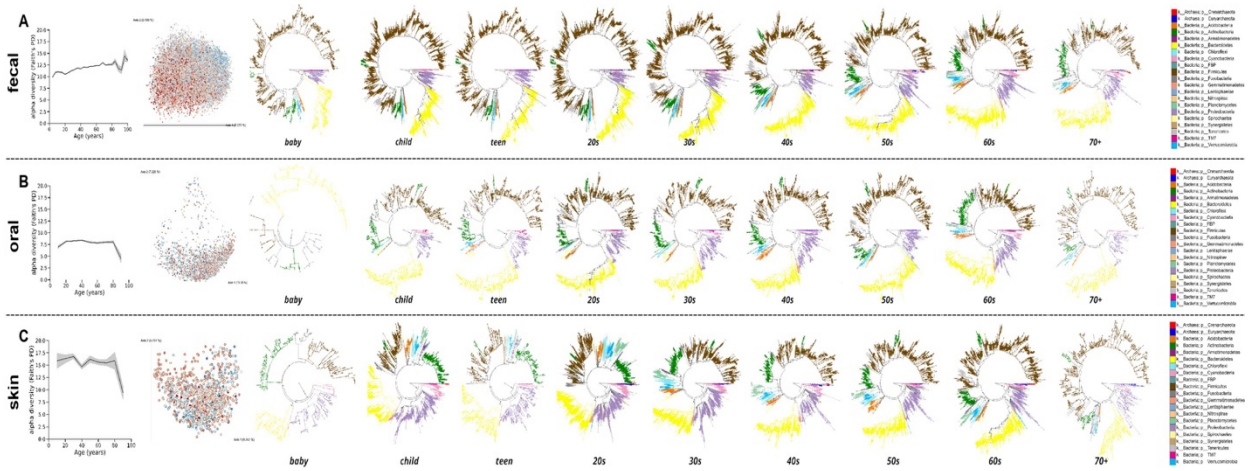

Figure 1.2. Measurements of bacterial diversity across age. The bacterial diversity and phylogenetic history of the human fecal (A), oral (B), and skin (C) microbiomes from birth to old age measured in the American Gut Project dataset, citizen science project containing 21,919 fecal, 1,920 oral, and 998 skin microbiome samples with 16S gene amplicon sequencing [56]. Alpha diversity, a quantitative measure of the number of different types of microbes in a sample, measured through Faith's PD alpha diversity metric across age (first column). The UniFrac beta diversity PCoA, a method for comparing the similarity of microbial communities where spatially close dots are similar samples and spatially distant dots represent dissimilar samples, colored by age (second column). The different microbes found at each stage of life represented by a phylogeny of their predicted evolutionary history, produced through SEPP insertion [201] of the Greengenes phylogeny [202] (third to last columns).

Primary succession (pre-life and early life)

The first factors that shape a human microbiome come from the mother during fetal development. The fetus is exposed to metabolites produced by the mother's microbial community through the placenta, which imprint its immune system and can affect both the normal microbiome and also various aspects of pathology later in life [27]. The composition and transfer of these metabolites to the fetus can be impacted by the mother's health, diet, and use of antibiotics during pregnancy [28–34]. The mothers microbiota play a role in shaping the fetal immune system which plays a role in disease susceptibility later in life. Dietary fiber is fermented by the mothers gut microbiota resulting in short-chain fatty acids (SCFAs) such as acetate which have been observed to be transferred across the placenta. Acetate in the fetal tissue has been observed to impact the epigenetic imprinting linked to the generation of T cells (Tregs) in adults, which is associated with protection from the development of asthma later in life [28]. In addition to microbial metabolites, Aryl hydrocarbon receptor (Ahr) a ligand produced by *E. coli* boosts the number and activity of myeloid cells, as well as of type 3 innate lymphoid cells (ILC3) which helps to shape the neonatal microbial and immune development [29,35–37]. During pregnancy, antibiotic use and gastrointestinal-related diseases such as Inflammatory Bowel Disease (IBD) are also thought to increase the risk of pathology in offspring later in life by imprinting of the fetal immune system [38–41]. However, these links have only been conducted in non-human experiments, such as in Torres et al, where germ-free mice were colonized with pregnant patients of IBD and infant microbiome demonstrated both aberrant microbiota and immune development indicative of IBD [40]. The mother's microbiome and immune system are also altered during pregnancy [42,43]. The mothers vaginal microbiome becomes pluripotent, containing many microbes conventionally found at other body sites [44]. While the immune system during pregnancy forms cooperative interactions with the fetus, forming the fetal immune system with transplacental IgG antibodies (reviewed in [45] and [46]) (**Figure 1.3A**).

The beginning of the human microbial community and the start of primary succession occur at birth with the seeding of the infant from the mother's microbiota. There is some debate as to whether the microbiota obtained at birth originate from both vaginal and fecal sources, through mixing, or if the vaginal microbiome itself is pluripotent at birth and is the sole source of microbial pioneers (i.e. the first species to colonize, setting the stage for other species later in succession) [44,47–49]. Regardless of the exact maternal source, this stage is characterized by pioneer bacterial taxa such as *Lactobacillus*, *Enterobacter*, *Bacteroides*, *Parabacteroides*, and *Prevotella,* which then colonize their conventional body sites: the gut, mouth, and skin [15,16,50,51]. At first, each body site of an infant is relatively undifferentiated, but pioneer microbes quickly begin a cascade of body site-dependent microbial diversity, and at least the bacteria at each site can be easily distinguished by the 4 - 6th week of life [1,13] (**Figure 1.3B**).

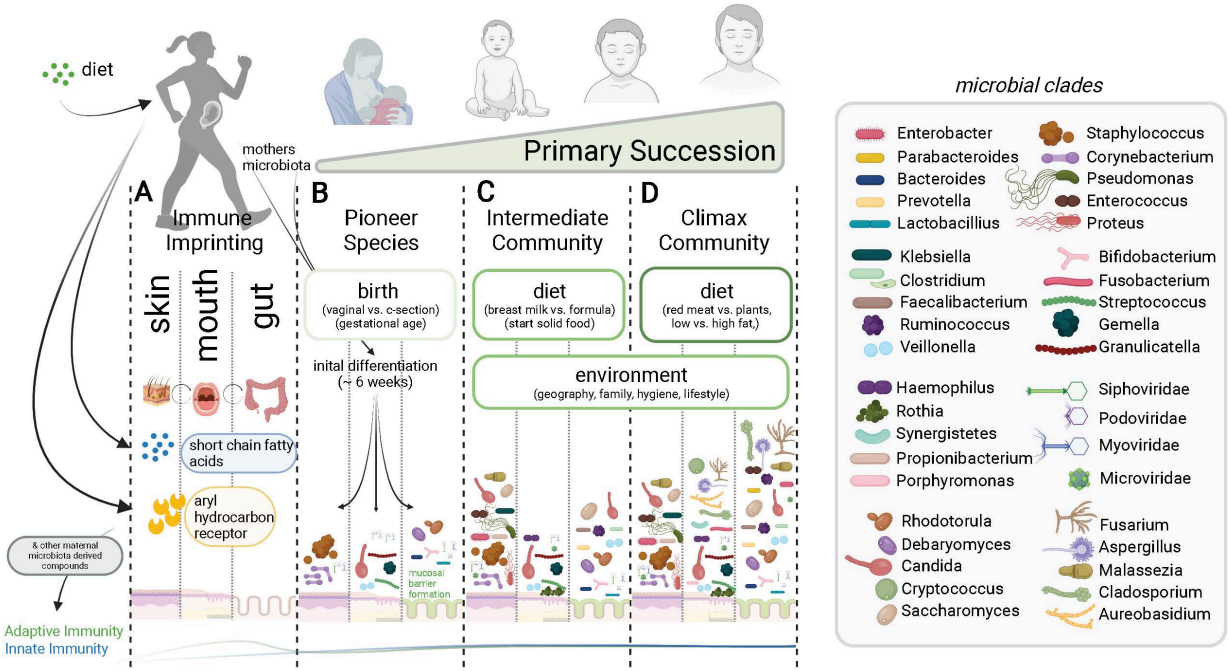

Figure 1.3. Primary succession (pre-life and early life). The future of the yet-to-be colonized fetus is set on an initial community assembly trajectory through the priming of each body site by the mothers imprinting on the immune system. Metabolites such as short chain fatty acids (e.g. Acetate) and other microbial compounds such as the bacterial ligand aryl hydrocarbon receptor can be transferred to the fetus through the placenta and influence immune development. These metabolites are also influenced by the mothers diet and health (**A**). Upon birth, the microbial community quickly differentiates by body site (**B**). During this initial colonization, the pioneer species and the community development of the next four years can be impacted by birth mode and gestation time. The following intermediate community is shaped by diets such as the consumption of breast milk or formula and the environment (**C**). Finally, the stable climax community is again shaped by diet and environment (**D**). Created with BioRender.com.

The development of the bacterial community in the human gut has been well studied (reviewed in [52]). *Bifidobacterium* spp. alone are dominant in the first month, but give way to a combination of *Bifidobacterium*, *Clostridium*, and *Bacteroides* spp. By the end of year one, this is followed by a greater increase in *Bacteroides*, a more diverse set of genera within the phylum Firmicutes (e.g. *Clostridium*, *Faecalibacterium*, *Ruminococcus*, *Veillonella*), and a relative decrease in pioneer species such as *Bifidobacterium* [14,15,47]. *Bifidobacterium* spp. catabolize Human Milk Oligosaccharides (HMOs) from the mother's breast milk, which is believed to begin imprinting the immune systems for life [53–56]. Most recently, Henrick et al. demonstrated that functional links exist between bacteria such as *Bifidobacterium* spp. containing genes required for catabolism of HMOs and the infant immune development. In particular, fecal waters from *Bifidobacterium infantis* EVC001 supplemented infants polarized naive T cells differently to those without, in a manner associated with decreased intestinal inflammation [53]. By about the third year of life, the gut bacterial community converges to the climax community sustained through adulthood. This community of microbes is one of the densest and most diverse ecologies known [57,58]. However, only two bacterial phyla are dominant in an average healthy person during this time: *Firmicutes* and *Bacteroidetes* [59].

The virome and mycobiome are far less explored than the bacteriome during the course of human gut development. The fungal community in the first few days of life is dominated by *Rhodotorula* and *Debaryomyces* spp., followed in the next month by *Candida*, *Cryptococcus*, and *Saccharomyces* spp. [2,60]. By adulthood, the dominant fungal genera are *Aspergillus*, *Candida*, and *Saccharomyces* [61–63]. The viral phage community is thought to be highly populated in the first week of life [64]. Phage families including Siphoviridae, Podoviridae, and Myoviridae are prevalent immediately after birth, primarily in lysogenic form (integrated into the bacterial genome) [65,66]. By the fourth month of life, the Caudovirales family of phages grow in abundance and are more often lytic (infectious phage particles or actively replicating phage) [67–69]. In adults, Caudovirales and Microviridae dominate the gut phage community but the phage gut virome is highly host-specific,

and much is still unknown about their succession (reviewed in [70]). Unlike phages, the gut virome of eukaryote-infecting viruses is mostly associated with pathology both in children (e.g. gastroenteritis reviewed in [71]) and in adults (reviewed in [72]). Recently, some eukaryote-infecting viruses have also been observed in low abundance both in children and adults, but their timing and prevalence are unknown [65,66,73] (**Figure 1.3B-D gut columns**).

The oral bacteriome is dominated by members of the genera *Streptococcus*, *Gemella*, *Granulicatella*, and *Veillonella* at birth [74]. In the following months, the genera *Lactobacillus* and *Fusobacterium* also become prevalent. *Staphylococcus* peaks around 3 months of life then steadily decreases, giving way to *Gemella*, *Granulicatella*, *Haemophilus*, and *Rothia* spp. [75]. After the formation of teeth, the oral microbiome shifts again, being dominated by the phyla Fusobacteriota, Synergistetes, Tenericutes, TM7, and SR1 into adulthood [76–79]. The oral mycobiome is believed to harbor less fungal diversity than the skin and gut [2]. *Candida* spp. are the first fungal colonizers of the oral cavity, on the first day of life [80,81]. Very little is known of the intermediate oral fungal community, but by adulthood it is known that *Candida*, *Cladosporium*, *Aureobasidium*, *Aspergillus*, *Fusarium*, and *Cryptococcus* spp. are in high prevalence [82]. To the authors' knowledge, not much is currently known about the colonization of the oral virome in human infants. In adults, similar to the gut, the most common phage group is Caudovirales [4,83,84]. The eukaryotic oral viral community is generally viewed as pathological in nature (e.g. Coxsackie A virus, *Morbillivirus*, *Rubulavirus*, and human papillomavirus), and there are no longitudinal studies of viral community composition [85]. However, many eukaryotic viral taxa have also been observed in asymptomatic and otherwise healthy adult subjects [86] (**Figure 1.3B-D oral columns**).

The skin bacterial community is dominated by the mother's vaginal *Lactobacillus* at birth [13,44]. By week 4-5 the infant skin microbiota resembles the adult skin microbiota, but continues to become more site-specific into adolescence, with dominant genera such as *Staphylococcus* and *Corynebacterium* across sites and *Pseudomonas*, *Enterobacter*, *Enterococcus*, *Proteus*, and *Klebsiella* at specific sites (e.g. armpit vs. forearm) [1,87]. In the skin mycobiome in the first 30 days

of life, it has been observed that species of the *Malassezia*, *Candida*, and *Saccharomyces* genera are most prevalent [2,88,89]. Little is known about the exact compositions of the intermediate community, but the adult mycobiome is dominated by *Malassezia* species, with estimates ranging from 75-90 % of the total fungal community composition [2,90]. Unlike the gut and oral cavity, the healthy skin microbiome harbors relatively little-known viral diversity and little study has been devoted to it, likely due to the technical limitations associated with low biomass samples [91]. However, it is known that there is some naturally residing viral population on the skin [92] (**Figure 1.3B-D skin columns**).

Several factors shape and differentiate microbial community development in the first few years of life. First, birth mode and antibiotic use are among the best-studied and clearest factors that influence the human microbial community. The process of natural microbial community establishment can be disrupted, in all body sites, through cesarean section and perinatal and neonatal antibiotic exposure [13–15,93–95]. Two of the best-sampled infant development studies, commonly abbreviated as DIABIMMUNE [14] and ECAM [15] followed infants for the first 2 and 3 years of life respectively, and focused on the impacts of antibiotic usage or birth mode. In both DIABIMMUNE and ECAM there were observed diseases in abundance of *Bacteroides spp.* in the development of those infants born by c-section compared to vaginal birth. The lack of natural pioneer microbiota to establish the microbial community results in a more variable community composition thought to be driven more by a stochastic than deterministic process, with the effects of birth mode on microbial community composition still observable until the fourth year of life [96,97]. This alteration in the natural development of the infant microbiome is associated with increased risks of infections, immune diseases, obesity, and neuroendocrine abnormalities [93,98–108]. Second, breastfeeding has been shown to have a large effect on microbiota development compared to other factors [94]. Similar to the impacts of cesarean section, the use of formula compared to breastfeeding leads to a higher diversity and less deterministic microbial community[109]. For example, given the natural dominance of Bifidobacteriaceae in the gut at birth, the lack of HMOs

as a primary nutrient source can lead to instability in the initial colonization [53]. However, much of the multi-omics integration of the microbiome, milk metabolome, and immune systems development is an area of active and rapidly advancing research [110]. As previously mentioned, one of the primary constituents of breast milk that affect the developing microbiota is a class of glycans referred to as HMOs, which are fermented by beneficial pioneer microbiota such as *Bifidobacterium* spp. and reduce pathogens through competitive binding to bacterial receptors over the gut mucosa in addition to the immune influences previously described [111–117]. In addition to HMOs, breast milk also contains immune-modulatory compounds such as lipopolysaccharide (LPS), secretory IgA (sIgA), innate immune factors, antimicrobial peptides, and prebiotic factors [29,118–121]. Finally, all of these factors impact human immune development. Microorganism-associated molecular pattern (MAMP)-based pattern recognition receptors (PRR) (e.g. Toll-like receptors (TLR) [122] & NOD-like receptors (NLR)[123]) interact with microbiota-derived molecular (e.g., LPS) and metabolites (e.g. SCFA, which interact with GPR43/GPR41/GPR109 [28], and secondary bile acids, which interact with FXR [124]) impacting immune development directly (reviewed in [125]). Some microbiota also rely on the immune system for colonization, such as *B. fragilis,* which depends on immunoglobulin A (IgA) [126]. Together, many of these factors contribute to the development of a unique, relatively stable microbial community of bacteria, fungi, and viruses that persists for a large part of the human lifespan.

Secondary succession in adolescence and adult life

Although the adult microbial community is largely stable compared to the large changes that occur during primary succession in infancy, the community can be perturbed away from the climax community state. The understanding of the microbiome during health and disease is a deepening and disease-specific research field (review in gut [127], skin [89], and oral microbiota [128]). There are also natural short-term changes that occur in the adult microbiome at timescales of a day to months or years. One of the best characterized examples of short-term changes is the

circadian rhythm in microbial community composition. Human gene expression and immune activation are known to be linked to the circadian rhythm[129], and the abundance and composition of bacteria within the microbiome also follow this pattern[19]. Bacterial families known to show a diurnal cycle in mice include Ruminococcaceae, Lachnospiraceae, S24-7, and Verrucomicrobiaceae, but little is known about equivalent cycles in humans because they produce feces less frequently than do mice[130]. A well-studied example of changes that occur on the scale of weeks to years is diet-driven alteration of the gut microbiome. Diet is known to have a large effect on microbial communities and can include natural and reversible changes in the community (reviewed in [131]). For example, the Hadza tribe of Tanzania, who eat a diet rich in meat and tubers in the dry season but a diet rich in honey and berries during the wet season, exhibit large seasonal fluctuations in genera such as *Bacteroides* that break down carbohydrates in meat [8,131,132] (**Figure 1.4A**). The large influence of diet in shaping the microbiome may also play a role in human health (reviewed in [133]), and much work is being dedicated to understanding how specific dietary components, and how dietary patterns overall, influence the microbiome and the impact in health. For example, western diets high levels of Red meat consumption have been linked to all-cause mortality [134]. The gut microbiota can act in a deleterious manner to convert L-carnitine, which is rich in red meat, to trimethylamine (TMA) and the liver converts TMA into trimethylamine N-oxide (TMAO) which is known to lead to atherosclerosis [135]. The gut microbiota can also act in a protective manner, for example by cleaving carcinogenic molecules from red meat before they are absorbed in the gut, acting as a protection from inflammation [132]. Besides diet many other factors help to shape the adult microbiome including genetics, geography, host factors such as metabolic disease or medicine (reviewed in [136]).

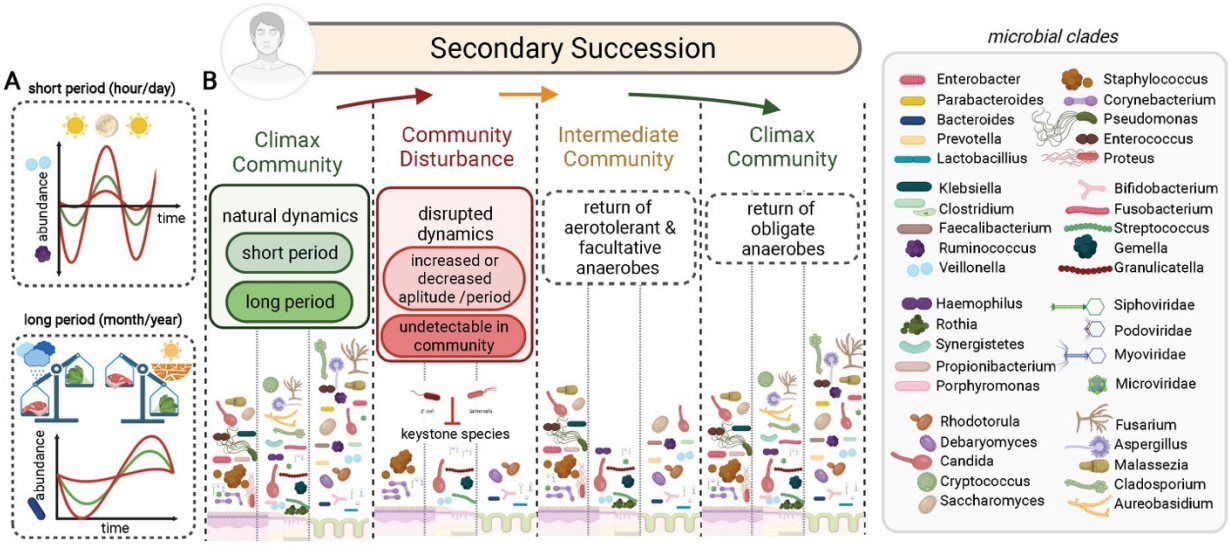

Figure 1.4. Secondary succession (adolescence and adult life).Compared to microbial community assembly in primary succession and human development, the microbial community in adulthood is relatively stable. There are natural dynamics and changes to this community such as microbial oscillations that correlate with host circadian rhythms by day/night and changes during diets or seasons (**A**). Secondary succession occurs when there is a disturbance, of which there are many possible avenues of impact, with antibiotics being one of the clearest examples. This disruption can cause microbial community members to be lost or fall below the level of detection and large changes in microbial dynamics such as the amplitude or periodicity. During this stage keystone species, similar to pioneer species, are thought to play a key role in preventing the overgrowth of opportunistic pathogens. Soon after the intermediate community forms dominated by the return of aerotolerant and facultative anaerobes. Finally, the community resembles the initial community with the return of obligate anaerobes (**B**). Created with BioRender.com.

Secondary successions that occur due to a disruption in the microbial community have been studied and reviewed extensively. Of the many factors that disrupt the microbiome, antibiotics are among the strongest, often with slow and subject-specific recovery after treatment [20,21]. The ability of the microbial community to rebound after antibiotic treatment is thought to depend on specific community members such as *Bacteroides thetaiotaomicron* and *Bifidobacterium adolescentis* [23]. Many species associated with post-antibiotic microbiome recovery are known keystone species [24]. Disease itself can also disrupt the microbiome, whether the change is initiated within the microbial community (overgrowth of a pathogen), from the host, or some combination of factors (reviewed in [10]). In some cases, such as cystic fibrosis (CF), the community experiences a series of secondary successions, which often can be overcome only through extreme measures such as surgery (reviewed in [137]) (**Figure 1.4B**). Many other diseases such as IBD disrupt the microbial community but are not observed to reach a new stable community composition, but rather continue to be chronically unstable in the absence of intervention (reviewed in [138]).

Challenges in microbial community recovery after a disturbance have led many researchers to explore the possibility of interventions for targeted restoration of the microbiota. Microbial community restoration involves directed reseeding or enrichment/depletion of certain species, with the intent to induce recovery to a microbial community close to that from before the disturbance. This can be attempted through probiotics, prebiotics, antibiotics or other drugs, transplantation of the complete microbial community from a healthy subject, or a combination of these. Although these therapies can be highly effective for restoring a healthy microbial community [139,140] they are often limited by lack of mechanistic knowledge of their interaction with the existing community [141], or by their ability to engraft only transiently [142]. To address the mechanisms, researchers have focused on two areas. The first area involves gaining a better understanding of how communities are assembled. The study of human development helps identify modifiable factors later in life; naturalizing microbial successions through seeding infants

with the mother's vaginal community at birth may prevent the need for intervention later in life [13,44,47]. Second, new methods for determining mechanism by exploring microbial community interactions both computationally [143] and experimentally, including high-throughput co-culturing [144] and genome editing of microbial communities, are being developed [145]. To address transience, two main approaches have been applied. First, the transient and individualized impact of microbial community therapeutics is driven by the individual nature of each person's microbiome [146]. Therefore, precision medicine, where community alteration is targeted to each person's unique microbiome, holds great promise. For example, personalized nutrition based on microbial community compositions effectively modified postprandial blood glucose in a blinded randomized controlled intervention [147]. Second, going beyond the bacteriome to explore the virome and mycobiome, and their inter-kingdom interactions, holds great promise. For example, phage therapy has already been employed in severe cases of drug-resistant bacterial infection [148], and is highly specific to the target bacterial strains [149].

Late succession (approaching the end of life)

Aging due to both biological programming and accumulation of damage throughout life impacts every aspect of cellular function, and the microbiome is no exception [150]. With advanced age, the gut microbiota alpha diversity decreases and the beta diversity (variation between individuals) increases [17,26,63,151,152]. Much is still unknown about the microbiota in old age, and the literature has been somewhat contradictory (e.g., Claesson et al. [153] reports increased *Bacteroides* in older adults, contradicting other studies), and most research has focused on gut bacteria. Generally, the community succession observed in the gut is bacteria is a decrease in genera dominant and prevalent in younger adults, such as *Bifidobacteria*, *Bacteroides*, and *Lactobacillus,* with a characteristic decrease in the ability to fend off blooms of opportunistic bacteria such as Enterobacteriacaea and *Clostridium spp.* [59,153,154]. Skin bacteria of the genus *Cutibacterium* (formerly, *Propionibacterium*) and *Staphylococcus* decrease in older age with a

greater abundance of *Corynebacterium* being observed [155]. In the oral body site, *Rothia* and *Streptococcus* spp. have been reported as dominating the core oral bacterial community, with consistent decreases in *Porphyromonas*, *Treponema*, and *Faecalibacterium* [156,157]. The gut mycobiome in old age is characterized by an increased dominance of *Penicillium*, *Candida*, *Aspergillus*, and *Saccharomyces* [61,63,158]. In the skin and oral body sites, very few studies exist, but old age is characterized by a decreased abundance of *Malassezia* in the skin and *Candida* in the oral cavity [159]. In phages, Siphoviridae dominance in adulthood gives way to Microviridae, Podoviridae, and crAssphages in old age [17]. Contrary to gut bacteria, fungi, and bacteriophage populations, eukaryotic viral diversity stays constant after childhood throughout the rest of life [17] (**Figure 1.5A**)

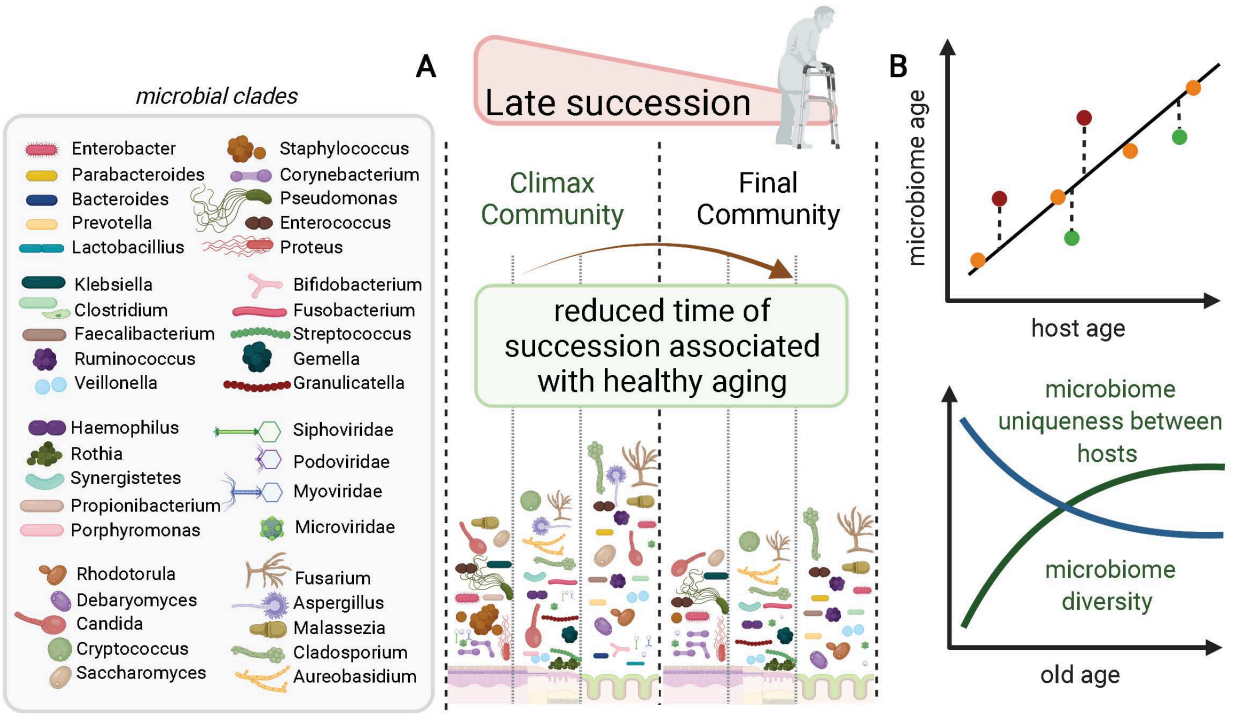

Figure 1.5. Late succession (approaching end of life). The final transition from the adult table microbial community to the final community in old age (A). Healthy aging is generally associated with a delayed transition to the final community (B, top). The final community characteristics are lower alpha diversity and increase uniqueness compared across different people of the same age (B, bottom). Created with BioRender.com.

Due to the high variability between individuals, the focus of research into microbial succession in old age has primarily been in the comparison of healthy and unhealthy aging. It remains unclear if the microbiome plays a mechanistic role in healthy aging or is just a strong indicator being influenced by the host (reviewed in Nagpal et al. [160]). However, in those who live longer and healthily, commonalities can be observed in sustained retention of those taxa highly prevalent in healthy adults such as Bacteroides [161]. This has led to defining a "microbiome age" which is based on the average microbial composition at a given host age [162]. The difference between the microbiome age and the true age has been an effective measure for human development [50] and similar approaches are being utilized in the microbiome in old age (**Figure 1.5B**). However, centenarians exhibit a wholly unique microbiome with increased alpha and beta diversity, complicating many of these comparisons [26,151,152]. Although promising, this area of research is still underpowered and an exciting area of current research.

The microbiome after death

Microbial succession does not end with the death of an individual, and in fact host death can be primarily viewed as an ecological disturbance to the microbiome. Immediately following cessation of the heart, tissues begin to break down due to the lack of oxygen [163,164]. Cellular functions continue until all the remaining oxygen is depleted and carbon dioxide is no longer able to be transported from the tissue [163,164]. The intracellular build-up of carbon dioxide creates a hypoxic, acidic environment leading to cell rupture [164,165]. Cellular components, such as enzymes (e.g., lipases), leak into the surrounding where they further facilitate tissue breakdown in a process called autolysis [165]. Autolysis triggers a cascade of microbial processes responsible for tissue breakdown (i.e., putrefaction) by eliminating the immune system, loosening cellular junctions, and providing nutrients to the microbiota [164–166]. During the first few days to weeks of decomposition, putrefaction is dominated by bacteria, but fungi have an increased role as

decomposition progresses [167–170] (**Figure 1.6A**). However, little is known about the virome succession and functional role during this process.

The human microbiome is relatively stable during the first 24-48 hours after death with distinct body site microbial ecologies [170], alpha diversity patterns by age [170], and identifiable personalized skin microbiome signatures [171]. Afterwards, the cascade of environmental changes facilitates a microbial succession that alters the human body and microbiome in a way that no longer resembles a living individual (unless the body is cooled or frozen). Microorganisms, released from the environmental constraints during host life, allow for both rapid changes in the relative abundance of microbes [168,170,172] as well as movement across body sites [166,173,174]. Migrating bacterial groups become pioneer species that translocate from the intestinal tract to extraintestinal sites taking part in either primary or secondary succession depending on the body site [174,175]. As the post-mortem interval of the host increases, alpha diversity of communities generally decreases (as is expected with nutrient pulses) and community composition (beta diversity) becomes more similar across body sites [168,176,177]. The gut and the skin are the two most well-studied human post-mortem microbial ecologies. Interestingly, the post-mortem community succession in the gut follows trends also detected in host's old age, with decreases in relative abundance of *Bacteroides* and *Lactobacillus* and an increase in relative abundance of *Clostridium* and taxa in the family Enterococcaceae [169,172,176,178]. The post-mortem composition and succession of skin microbial communities depends on the external environment. For example, if exposed to soil, most post-mortem microbes, including eukaryotes such as nematodes, appear to assemble from soil communities [168].

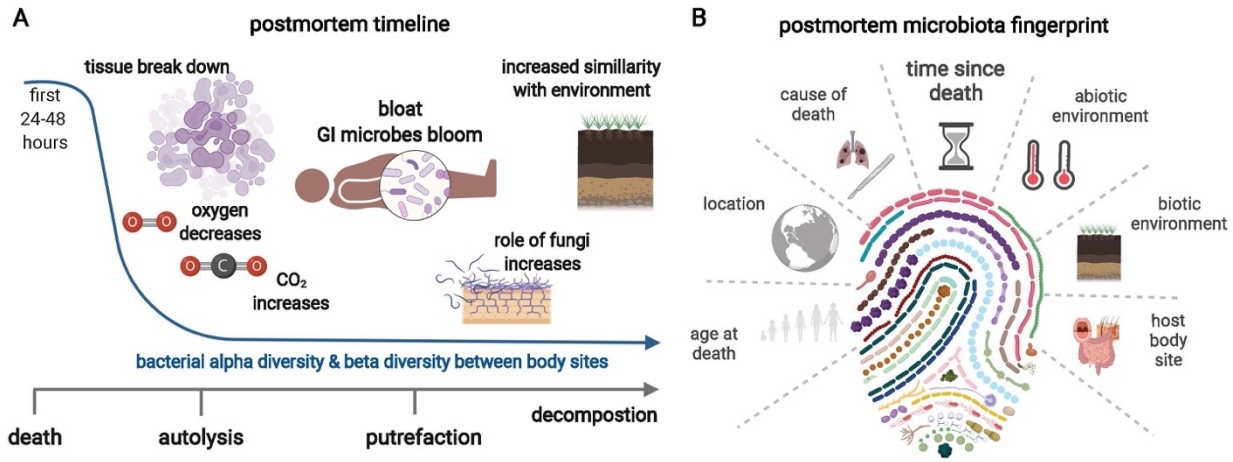

Figure 1.6. The microbiome after death. After death the microbiome is relatively stable in the first 24-48 hours, the tissue then begins to break down during autolysis leading to bloom in the gastrointestinal (GI) microbiota and a decrease in alpha diversity and a decrease in beta diversity between body sites. During putrefaction, the role of eukaryotic microorganisms increases, and the host body and the surrounding environment become more similar (A). The post-mortem microbiome is unique to each host body and is distinct between bodies based on time since death, cause of death, environment, location, age at death, and, in the beginning, between body sites (B). Created with BioRender.com.

The microbiome of death has garnered increased attention due to its implications for forensic investigations. The consistent temporal patterns of succession associated with multiple individuals and body sites are evidence that the post-mortem microbiome may serve as a bioindicator of the post-mortem interval (PMI) [168,172,176,177,179]. Post-mortem interval estimations appear more accurate during the earlier stages of decomposition (e.g. the first 2-3 weeks post-mortem) when microbial succession includes rapid turnover of community members [176,180], but are still useful in later stages of decomposition (e.g. in bone) when few lines of evidence exist for estimating PMI [181,182]. Connections with cause of death and microbial presence have also been demonstrated [170]. For example, increased detection of *Rothia* was found in the oral microbiome of individuals who died of heart disease and may be an indicator of host dysbiosis [170]. Moreover, skin microbiome shedding may contribute to trace evidence by being able to connect individuals with items they have interacted with such as cell phones [171,183]; however, the time this unique signature can be accurately matched to an individual varies on the object's material and usage [171] (**Figure 1.6B**).

1.2. Conclusions and outlook

In this review, we describe the current understanding of human resident microbial community composition across ages and different body sites. The many connections between human health and our microbial community composition are bringing an increasing interest in interventions. Interventions that focus on the whole microbial community, rather than the enrichment or elimination of a single species, require an understanding of how these communities are formed and maintained. Through studying microbial communities across the human lifespan, we may better understand these complex interactions and how to effectively push the community to a desired composition for the host. Moreover, as discussed here, these insights are being applied in several other areas such as in the field of forensics.

Although this Review has focused on the microbiome and its role in healthy aging, many conditions have been associated with accelerated aging, and are just beginning to be studied in a microbiome context, a key example being schizophrenia [184]. “Social determinants of health” have a major impact on health, aging, and longevity, both in the context of healthy and pathological aging. These factors include education, poverty, occupation, discrimination, social connections, etc. [185]. Since many of these factors have also been linked to the microbiome, understanding the role of these social determinants and how to modify their effects to promote healthy aging will be an important topic for future research linking the microbiome and aging.

Despite the enormous effort and resources being put into characterizing the microbiome, we have just scratched the surface. There are large disparities in our understanding of microbial kingdom, mainly due to technical difficulties in characterizing taxa other than bacteria [186]. The gaps in understanding virome and mycobiome community structure and cross-kingdom interactions are an area of exciting research driven by technical advances that improve accuracy and decrease the cost of DNA sequencing. However, contradictions in the field are abundant, especially in those observations driven purely by sequencing data, and more robust analyses are key to consolidating knowledge in the field (e.g., log-ratios) (**Box 1**) [187]. Measuring species beyond relative taxonomic compositions through high-throughput cultivation, metagenomics, transcriptomics, and metabolomics are rapidly expanding areas of research that are key to filling in gaps in our understanding [143].

1.3. Methods

Box 1: Sampling and quantifying microbial communities

Study design and sample collection

The human microbiome is dynamic [188]. With this in mind, it is important to design a sampling strategy that can capture the temporal and spatial variability of the microbiome, particularly when these fluctuations are relevant to the scientific questions asked. When a single sample is collected from each individual, the study is called **cross-sectional**, while sampling performed at multiple time points or at multiple body sites is referred to as **repeated measures** study. With time, the frequency of sampling should be tuned to the phenomenon researchers are attempting to observe. For example, circadian rhythm studies typically sample every 2-4 hours [189] while in IBD, it has been shown that sampling patients between 3-5 times over a period of weeks can improve disease classification [190]. In other applications, such as studying the effect of particular treatments on an individual microbiome, it may be relevant to perform an **n of one** study in which the same participant is repeatedly probed for resultant changes in their microbiome; samples collected before treatment are regarded as individual-level controls [191].

In addition to considering the frequency and location of sampling, it is important to consider how the geography and ethnicity of the sampled population impact the results of a study. For example, one of the microbes most highly associated with aging in Chinese cohorts is not detected in American cohorts [162]. Similarly, environmental factors associated with urbanized societies (i.e. the "built environment") such as decreased exposure to environmental microbes and increased use of household antimicrobials, significantly shift the human microbiome [192]. On the whole, conclusions from a given study may not generalize well to other societies and cultures. This is particularly relevant for the microbiome field given that a large majority of public human microbiome data comes from North American and European populations, with nearly half coming from the United States alone [193].

Data generation

The main categories of sequencing data that are generated from human microbiome studies are **amplicon** sequencing data and **shotgun** sequencing data. In amplicon sequencing,

the PCR products (amplicons) of established hypervariable regions are deeply sequenced, allowing identification and measurement of community members by matching to their individual "barcodes". There are two choices to be made here - the gene to amplify and which portion of that gene to amplify. Commonly amplified genes are the 16S ribosomal rRNA (rRNA) gene for bacteria, 18S rRNA for eukaryotic microbes, and internal transcribed spacer (ITS) for fungi. The choice of the hypervariable region within each specific gene to amplify depends on the particular microbes to capture, but broad, commonly used ones include the V4 from the Earth Microbiome Project [194]. In shotgun sequencing, all microbial DNA is sequenced instead of only PCR products, enabling a more specific taxonomic classification of microbes. Because it does not rely on any marker genes, shotgun sequencing is less biased than amplicon sequencing is towards certain sets of microbes.

Pairing sequencing data with other analyses

Pairing sequencing data with other analyses, including other -omics techniques can enrich the data collected. We summarize techniques commonly performed in tandem with microbiome sequencing in the table below.

Table 1.1 Methods for sampling and quantifying microbial communities.

Technique	Enhancement to amplicon or metagenomic sequencing	Citation
qPCR, FACS	Anchors relative abundance metrics to an absolute abundance	[195,196]
Host immune	ELISA, Single-cell sequencing	[197]
Culturomics	Obtain culture conditions for previously unculturable microbes	[198]
Metabolomics	Identify microbially produced metabolites; chemical effectors of microbiome function	[199]
Proteomics	Identify microbially produced proteins; another biological effector of the microbiome	
Host genomics / transcriptomics	Variant calling for how host genetics may be different; host gene expression	

Metadata

Finally, it is paramount to collect data from the subjects surveyed. Some important categories of metadata for general microbiome studies include demographics, clinical (i.e. other conditions, antibiotic use), and dietary information, however the exact metadata used will vary by study. Practices for producing standardized metadata should be adopted so that results are reusable and reproducible [200].

1.4. Acknowledgements

This work was supported by the National Institute of Justice under award number 2016-DN-BX-0194 (to R.K and J.L.M.) and the National Institutes of Health under award number U19AG063744 Project 1: Changes in Gut Microbiome (to R.K.). A.H.D. is supported by the Stein Institute for Research on Aging, Natasha Josefowitz Predoctoral Fellowship, and the Reiter Endowed Fellowship.

1.5. References

1. Chu, D. M. *et al.* Maturation of the infant microbiome community structure and function across multiple body sites and in relation to mode of delivery. *Nat. Med.* **23**, 314–326 (2017).

2. Ward, T. L. *et al.* Development of the Human Mycobiome over the First Month of Life and across Body Sites. *mSystems* **3**, (2018).

3. Oh, J. *et al.* Biogeography and individuality shape function in the human skin metagenome. *Nature* **514**, 59–64 (2014).

4. Abeles, S. R. *et al.* Human oral viruses are personal, persistent and gender-consistent. *ISME J.* **8**, 1753–1767 (2014).

5. Grice, E. A. & Segre, J. A. The Human Microbiome: Our Second Genome. *Annual Review of Genomics and Human Genetics* vol. 13 151–170 (2012).

6. Lauber, C. L., Hamady, M., Knight, R. & Fierer, N. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl. Environ. Microbiol.* **75**, 5111–5120 (2009).

7. Zengler, K. & Zaramela, L. S. The social network of microorganisms - how auxotrophies shape complex communities. *Nat. Rev. Microbiol.* **16**, 383–390 (2018).

8. Smits, S. A. *et al.* Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. *Science* **357**, 802–806 (2017).

9. Rasko, D. A. Changes in microbiome during and after travellers' diarrhea: what we know and what we do not. *J. Travel Med.* **24**, S52–S56 (2017).

10. Zheng, D., Liwinski, T. & Elinav, E. Interaction between microbiota and immunity in health and disease. *Cell Res.* **30**, 492–506 (2020).

11. Zaneveld, J. R., McMinds, R. & Vega Thurber, R. Stress and stability: applying the Anna Karenina principle to animal microbiomes. *Nat Microbiol* **2**, 17121 (2017).

12. Dini-Andreote, F., Stegen, J. C., van Elsas, J. D. & Salles, J. F. Disentangling mechanisms that mediate the balance between stochastic and deterministic processes in microbial succession. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E1326–E1332 (2015).

13. Dominguez-Bello, M. G. *et al.* Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 11971–11975 (2010).

14. Yassour, M. *et al.* Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Sci. Transl. Med.* **8**, 343ra81 (2016).

15. Bokulich, N. A. *et al.* Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Sci. Transl. Med.* **8**, 343ra82 (2016).

16. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).

17. Gregory, A. C. *et al.* The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut. *Cell Host Microbe* **28**, 724–740.e8 (2020).

18. Faith, J. J. *et al.* The long-term stability of the human gut microbiota. *Science* **341**, 1237439 (2013).

19. Thaiss, C. A. *et al.* Microbiota Diurnal Rhythmicity Programs Host Transcriptome Oscillations. *Cell* **167**, 1495–1510.e12 (2016).

20. Zaura, E. *et al.* Same Exposure but Two Radically Different Responses to Antibiotics: Resilience of the Salivary Microbiome versus Long-Term Microbial Shifts in Feces. *MBio* **6**, e01693–15 (2015).

21. Dethlefsen, L. & Relman, D. A. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proc. Natl. Acad. Sci. U. S. A.* **108 Suppl 1**, 4554–4561 (2011).

22. Hsiao, A. *et al.* Members of the human gut microbiota involved in recovery from Vibrio cholerae infection. *Nature* **515**, 423–426 (2014).

23. Chng, K. R. *et al.* Metagenome-wide association analysis identifies microbial determinants of post-antibiotic ecological recovery in the gut. *Nat Ecol Evol* **4**, 1256–1267 (2020).

24. Gibbons, S. M. Keystone taxa indispensable for microbiome recovery. *Nature microbiology* vol. 5 1067–1068 (2020).

25. Rizzatti, G., Lopetuso, L. R., Gibiino, G., Binda, C. & Gasbarrini, A. Proteobacteria: A Common Factor in Human Diseases. *Biomed Res. Int.* **2017**, 9351507 (2017).

26. Biagi, E. *et al.* Gut Microbiota and Extreme Longevity. *Curr. Biol.* **26**, 1480–1485 (2016).

27. Al Nabhani, Z. & Eberl, G. Imprinting of the immune system by the microbiota early in life. *Mucosal Immunol.* **13**, 183–189 (2020).

28. Thorburn, A. N. *et al.* Evidence that asthma is a developmental origin disease influenced by maternal diet and bacterial metabolites. *Nat. Commun.* **6**, 7320 (2015).

29. Gomez de Agüero, M. *et al.* The maternal microbiota drives early postnatal innate immune development. *Science* **351**, 1296–1302 (2016).

30. Macpherson, A. J., de Agüero, M. G. & Ganal-Vonarburg, S. C. How nutrition and the maternal microbiota shape the neonatal immune system. *Nat. Rev. Immunol.* **17**, 508–517 (2017).

31. Nakajima, A. *et al.* Maternal High Fiber Diet during Pregnancy and Lactation Influences Regulatory T Cell Differentiation in Offspring in Mice. *J. Immunol.* **199**, 3516–3524 (2017).

32. Jamalkandi, S. A. *et al.* Oral and nasal probiotic administration for the prevention and alleviation of allergic diseases, asthma and chronic obstructive pulmonary disease. *Nutr. Res. Rev.* **34**, 1–16 (2021).

33. Örtqvist, A. K., Lundholm, C., Halfvarson, J., Ludvigsson, J. F. & Almqvist, C. Fetal and early life antibiotics exposure and very early onset inflammatory bowel disease: a population-based study. *Gut* **68**, 218–225 (2019).

34. Munyaka, P. M., Eissa, N., Bernstein, C. N., Khafipour, E. & Ghia, J.-E. Antepartum Antibiotic Treatment Increases Offspring Susceptibility to Experimental Colitis: A Role of the Gut Microbiota. *PLoS One* **10**, e0142536 (2015).

35. Kiss, E. A. *et al.* Natural aryl hydrocarbon receptor ligands control organogenesis of intestinal lymphoid follicles. *Science* **334**, 1561–1565 (2011).

36. Lee, J. S. *et al.* AHR drives the development of gut ILC22 cells and postnatal lymphoid tissues via pathways dependent on and independent of Notch. *Nat. Immunol.* **13**, 144–151 (2011).

37. Qiu, J. *et al.* The aryl hydrocarbon receptor regulates gut immunity through modulation of innate lymphoid cells. *Immunity* **36**, 92–104 (2012).

38. Schulfer, A. F. *et al.* Intergenerational transfer of antibiotic-perturbed microbiota enhances colitis in susceptible mice. *Nat Microbiol* **3**, 234–242 (2018).

39. Ma, J. *et al.* High-fat maternal diet during pregnancy persistently alters the offspring microbiome in a primate model. *Nature Communications* vol. 5 (2014).

40. Torres, J. *et al.* Infants born to mothers with IBD present with altered gut microbiome that transfers abnormalities of the adaptive immune system to germ-free mice. *Gut* **69**, 42–51 (2020).

41. Milliken, S., Allen, R. M. & Lamont, R. F. The role of antimicrobial treatment during pregnancy on the neonatal gut microbiome and the development of atopy, asthma, allergy and obesity in childhood. *Expert Opin. Drug Saf.* **18**, 173–185 (2019).

42. Santacruz, A. *et al.* Gut microbiota composition is associated with body weight, weight gain and biochemical parameters in pregnant women. *Br. J. Nutr.* **104**, 83–92 (2010).

43. Trevisanuto, D. *et al.* Fetal placental inflammation is associated with poor neonatal growth of preterm infants: a case-control study. *J. Matern. Fetal. Neonatal Med.* **26**, 1484–1490 (2013).

44. Song, S. J. *et al.* Naturalization of the microbiota developmental trajectory of Cesarean-born neonates after vaginal seeding. *Med* (2021) doi:10.1016/j.medj.2021.05.003.

45. Abu-Raya, B., Michalski, C., Sadarangani, M. & Lavoie, P. M. Maternal Immunological Adaptation During Normal Pregnancy. *Front. Immunol.* **11**, 575197 (2020).

46. Hanson, L. A. *et al.* The transfer of immunity from mother to child. *Ann. N. Y. Acad. Sci.* **987**, 199–206 (2003).

47. Dominguez-Bello, M. G. *et al.* Partial restoration of the microbiota of cesarean-born infants via vaginal microbial transfer. *Nat. Med.* **22**, 250–253 (2016).

48. 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).

49. Helve, O. *et al.* 2843. Maternal fecal transplantation to infants born by cesarean section: Safety and feasibility. *Open Forum Infect. Dis.* **6**, S68–S68 (2019).

50. Subramanian, S. *et al.* Persistent gut microbiota immaturity in malnourished Bangladeshi children. *Nature* **510**, 417–421 (2014).

51. Palmer, C., Bik, E. M., DiGiulio, D. B., Relman, D. A. & Brown, P. O. Development of the human infant intestinal microbiota. *PLoS Biol.* **5**, e177 (2007).

52. Groer, M. W. *et al.* Development of the preterm infant gut microbiome: a research priority. *Microbiome* **2**, 38 (2014).

53. Henrick, B. M. *et al.* Bifidobacteria-mediated immune system imprinting early in life. *Cell* **184**, 3884–3898.e11 (2021).

54. Sela, D. A. & Mills, D. A. Nursing our microbiota: molecular linkages between bifidobacteria and milk oligosaccharides. *Trends Microbiol.* **18**, 298–307 (2010).

55. Seppo, A. E. *et al.* Infant gut microbiome is enriched with Bifidobacterium longum ssp. infantis in Old Order Mennonites with traditional farming lifestyle. *Allergy* **76**, 3489–3503 (2021).

56. Triantis, V., Bode, L. & van Neerven, R. J. J. Immunological Effects of Human Milk Oligosaccharides. *Front Pediatr* **6**, 190 (2018).

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

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

59. Odamaki, T. *et al.* Age-related changes in gut microbiota composition from newborn to centenarian: a cross-sectional study. *BMC Microbiology* vol. 16 (2016).

60. Schei, K. *et al.* Early gut mycobiota and mother-offspring transfer. *Microbiome* **5**, 107 (2017).

61. Alonso, R., Pisa, D., Fernández-Fernández, A. M. & Carrasco, L. Infection of Fungi and Bacteria in Brain Tissue From Elderly Persons and Patients With Alzheimer's Disease. *Frontiers in Aging Neuroscience* vol. 10 (2018).

62. Nagpal, R. *et al.* Gut mycobiome and its interaction with diet, gut bacteria and alzheimer's disease markers in subjects with mild cognitive impairment: A pilot study. *EBioMedicine* **59**, 102950 (2020).

63. Ahmad, H. F. *et al.* Gut Mycobiome Dysbiosis Is Linked to Hypertriglyceridemia among Home Dwelling Elderly Danes. *bioRxiv* 2020.04.16.044693 (2020) doi:10.1101/2020.04.16.044693.

64. Breitbart, M. *et al.* Metagenomic analyses of an uncultured viral community from human feces. *J. Bacteriol.* **185**, 6220–6223 (2003).

65. Liang, G. *et al.* The stepwise assembly of the neonatal virome is modulated by breastfeeding. *Nature* vol. 581 470–474 (2020).

66. Lim, E. S. *et al.* Early life dynamics of the human gut virome and bacterial microbiome in infants. *Nature Medicine* vol. 21 1228–1234 (2015).

67. Liang, G. *et al.* Dynamics of the Stool Virome in Very Early-Onset Inflammatory Bowel Disease. *J. Crohns. Colitis* **14**, 1600–1610 (2020).

68. Koren, O. & Rautava, S. *The Human Microbiome in Early Life: Implications to Health and Disease*. (Academic Press, 2020).

69. Reyes, A. *et al.* Gut DNA viromes of Malawian twins discordant for severe acute malnutrition. *Proceedings of the National Academy of Sciences* vol. 112 11941–11946 (2015).

70. Liang, G. & Bushman, F. D. The human virome: assembly, composition and host interactions. *Nat. Rev. Microbiol.* **19**, 514–527 (2021).

71. Oude Munnink, B. B. & van der Hoek, L. Viruses Causing Gastroenteritis: The Known, The New and Those Beyond. *Viruses* **8**, (2016).

72. Woolhouse, M., Scott, F., Hudson, Z., Howey, R. & Chase-Topping, M. Human viruses: discovery and emergence. *Philosophical Transactions of the Royal Society B: Biological Sciences* vol. 367 2864–2871 (2012).

73. Rascovan, N., Duraisamy, R. & Desnues, C. Metagenomics and the Human Virome in Asymptomatic Individuals. *Annual Review of Microbiology* vol. 70 125–141 (2016).

74. Mason, M. R., Chambers, S., Dabdoub, S. M., Thikkurissy, S. & Kumar, P. S. Characterizing oral microbial communities across dentition states and colonization niches. *Microbiome* **6**, 67 (2018).

75. Dzidic, M. *et al.* Oral microbiome development during childhood: an ecological succession influenced by postnatal factors and associated with tooth decay. *ISME J.* **12**, 2292–2306 (2018).

76. Merglova, V. & Polenik, P. Early colonization of the oral cavity in 6- and 12-month-old infants by cariogenic and periodontal pathogens: a case-control study. *Folia Microbiol.* **61**, 423–429 (2016).

77. Gomez, A. & Nelson, K. E. The Oral Microbiome of Children: Development, Disease, and Implications Beyond Oral Health. *Microb. Ecol.* **73**, 492–503 (2017).

78. Cephas, K. D. *et al.* Comparative analysis of salivary bacterial microbiome diversity in edentulous infants and their mothers or primary care givers using pyrosequencing. *PLoS One* **6**, e23503 (2011).

79. Crielaard, W. *et al.* Exploring the oral microbiota of children at various developmental stages of their dentition in the relation to their oral health. *BMC Med. Genomics* **4**, 22 (2011).

80. Darwazeh, A. M. & al-Bashir, A. Oral candidal flora in healthy infants. *J. Oral Pathol. Med.* **24**, 361–364 (1995).

81. Stecksén-Blicks, C., Granström, E., Silfverdal, S. A. & West, C. E. Prevalence of oral Candida in the first year of life. *Mycoses* **58**, 550–556 (2015).

82. Ghannoum, M. A. *et al.* Characterization of the oral fungal microbiome (mycobiome) in healthy individuals. *PLoS Pathog.* **6**, e1000713 (2010).

83. Abeles, S. R., Ly, M., Santiago-Rodriguez, T. M. & Pride, D. T. Effects of Long Term Antibiotic Therapy on Human Oral and Fecal Viromes. *PLOS ONE* vol. 10 e0134941 (2015).

84. Pérez-Brocal, V. & Moya, A. The analysis of the oral DNA virome reveals which viruses are widespread and rare among healthy young adults in Valencia (Spain). *PLOS ONE* vol. 13 e0191867 (2018).

85. Dye, B. A., Li, X. & Thornton-Evans, G. *Oral Health Disparities as Determined by Selected Healthy People 2020 Oral Health Objectives for the United States, 2009-2010*. (U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, 2012).

86. Baker, J. L., Bor, B., Agnello, M., Shi, W. & He, X. Ecology of the Oral Microbiome: Beyond Bacteria. *Trends in Microbiology* vol. 25 362–374 (2017).

87. Gaitanis, G. *et al.* Variation of cultured skin microbiota in mothers and their infants during the first year postpartum. *Pediatr. Dermatol.* **36**, 460–465 (2019).

88. Lee, Y. W., Yim, S. M., Lim, S. H., Choe, Y. B. & Ahn, K. J. Quantitative investigation on the distribution of Malassezia species on healthy human skin in Korea. *Mycoses* **49**, 405–410 (2006).

89. Byrd, A. L., Belkaid, Y. & Segre, J. A. The human skin microbiome. *Nat. Rev. Microbiol.* **16**, 143–155 (2018).

90. Sugita, T. *et al.* Quantitative analysis of the cutaneous Malassezia microbiota in 770 healthy Japanese by age and gender using a real-time PCR assay. *Medical Mycology* 1–5 (2009) doi:10.1080/13693780902977976.

91. Moya, A. & Brocal, V. P. *The Human Virome: Methods and Protocols*. (Springer New York, 2018).

92. Foulongne, V. *et al.* Human skin microbiota: high diversity of DNA viruses identified on the human skin by high throughput sequencing. *PLoS One* **7**, e38499 (2012).

93. Shao, Y. *et al.* Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. *Nature* **574**, 117–121 (2019).

94. Stewart, C. J. *et al.* Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature* **562**, 583–588 (2018).

95. Ainonen, S. *et al.* Antibiotics at birth and later antibiotic courses: effects on gut microbiota. *Pediatr. Res.* (2021) doi:10.1038/s41390-021-01494-7.

96. Martino, C. *et al.* Context-aware dimensionality reduction deconvolutes gut microbial community dynamics. *Nat. Biotechnol.* **39**, 165–168 (2021).

97. Furman, O. *et al.* Stochasticity constrained by deterministic effects of diet and age drive rumen microbiome assembly dynamics. *Nat. Commun.* **11**, 1904 (2020).

98. Malamitsi-Puchner, A. *et al.* The influence of the mode of delivery on circulating cytokine concentrations in the perinatal period. *Early Hum. Dev.* **81**, 387–392 (2005).

99. Stokholm, J. *et al.* Maturation of the gut microbiome and risk of asthma in childhood. *Nat. Commun.* **9**, 141 (2018).

100. Andersen, V., Möller, S., Jensen, P. B., Møller, F. T. & Green, A. Caesarean Delivery and Risk of Chronic Inflammatory Diseases (Inflammatory Bowel Disease, Rheumatoid Arthritis, Coeliac Disease, and Diabetes Mellitus): A Population Based Registry Study of 2,699,479 Births in Denmark During 1973–2016. *Clinical Epidemiology* vol. 12 287–293 (2020).

101. Blustein, J. *et al.* Association of caesarean delivery with child adiposity from age 6 weeks to 15 years. *Int. J. Obes.* **37**, 900–906 (2013).

102. Ardic, C., Usta, O., Omar, E., Yıldız, C. & Memis, E. Caesarean delivery increases the risk of overweight or obesity in 2-year-old children. *J. Obstet. Gynaecol.* **41**, 374–379 (2021).

103. Cox, L. M. *et al.* Altering the intestinal microbiota during a critical developmental window has lasting metabolic consequences. *Cell* **158**, 705–721 (2014).

104. Martinez, K. A., 2nd *et al.* Increased weight gain by C-section: Functional significance of the primordial microbiome. *Sci Adv* **3**, eaao1874 (2017).

105. Olszak, T. *et al.* Microbial exposure during early life has persistent effects on natural killer T cell function. *Science* **336**, 489–493 (2012).

106. Livanos, A. E. *et al.* Antibiotic-mediated gut microbiome perturbation accelerates development of type 1 diabetes in mice. *Nat Microbiol* **1**, 16140 (2016).

107. Moya-Pérez, A. *et al.* Intervention strategies for cesarean section–induced alterations in the microbiota-gut-brain axis. *Nutrition Reviews* vol. 75 225–240 (2017).

108. Braniste, V. *et al.* The gut microbiota influences blood-brain barrier permeability in mice. *Sci. Transl. Med.* **6**, 263ra158 (2014).

109. Forbes, J. D. *et al.* Association of Exposure to Formula in the Hospital and Subsequent Infant Feeding Practices With Gut Microbiota and Risk of Overweight in the First Year of Life. *JAMA Pediatrics* vol. 172 e181161 (2018).

110. Shenhav, L. & Azad, M. B. Using Community Ecology Theory and Computational Microbiome Methods To Study Human Milk as a Biological System. *mSystems* **7**, e0113221 (2022).

111. Bridgman, S. L. *et al.* Fecal Short-Chain Fatty Acid Variations by Breastfeeding Status in Infants at 4 Months: Differences in Relative versus Absolute Concentrations. *Front Nutr* **4**, 11 (2017).

112. Zivkovic, A. M., German, J. B., Lebrilla, C. B. & Mills, D. A. Human milk glycobiome and its impact on the infant gastrointestinal microbiota. *Proceedings of the National Academy of Sciences* vol. 108 4653–4658 (2011).

113. Ayechu-Muruzabal, V. *et al.* Diversity of Human Milk Oligosaccharides and Effects on Early Life Immune Development. *Frontiers in Pediatrics* vol. 6 (2018).

114. Murphy, K. *et al.* The Composition of Human Milk and Infant Faecal Microbiota Over the First Three Months of Life: A Pilot Study. *Scientific Reports* vol. 7 (2017).

115. Charbonneau, M. R. *Characterizing the Role of Sialylated Milk Glycans and the Infant Gut Microbiota in Growth and Metabolism*. (Washington University, 2015).

116. Bode, L. The functional biology of human milk oligosaccharides. *Early Human Development* vol. 91 619–622 (2015).

117. Bode, L. Human milk oligosaccharides: Every baby needs a sugar mama. *Glycobiology* vol. 22 1147–1162 (2012).

118. Kaetzel, C. S. Cooperativity among secretory IgA, the polymeric immunoglobulin receptor, and the gut microbiota promotes host-microbial mutualism. *Immunol. Lett.* **162**, 10–21 (2014).

119. Munblit, D., Verhasselt, V. & Warner, J. O. *Human Milk Composition and Health Outcomes in Children*. (Frontiers Media SA, 2019).

120. Mastromarino, P. *et al.* Correlation between lactoferrin and beneficial microbiota in breast milk and infant's feces. *BioMetals* vol. 27 1077–1086 (2014).

121. Agus, A., Planchais, J. & Sokol, H. Gut Microbiota Regulation of Tryptophan Metabolism in Health and Disease. *Cell Host Microbe* **23**, 716–724 (2018).

122. Coats, S. R., Pham, T.-T. T., Bainbridge, B. W., Reife, R. A. & Darveau, R. P. MD-2 Mediates the Ability of Tetra-Acylated and Penta-Acylated Lipopolysaccharides to AntagonizeEscherichia coliLipopolysaccharide at the TLR4 Signaling Complex. *The Journal of Immunology* vol. 175 4490–4498 (2005).

123. Denou, E. *et al.* Defective NOD 2 peptidoglycan sensing promotes diet-induced inflammation, dysbiosis, and insulin resistance. *EMBO Molecular Medicine* vol. 7 259–274 (2015).

124. Quinn, R. A. *et al.* Global chemical effects of the microbiome include new bile-acid conjugations. *Nature* **579**, 123–129 (2020).

125. Rooks, M. G. & Garrett, W. S. Gut microbiota, metabolites and host immunity. *Nature Reviews Immunology* vol. 16 341–352 (2016).

126. Donaldson, G. P. *et al.* Gut microbiota utilize immunoglobulin A for mucosal colonization. *Science* **360**, 795–800 (2018).

127. Fan, Y. & Pedersen, O. Gut microbiota in human metabolic health and disease. *Nature Reviews Microbiology* vol. 19 55–71 (2021).

128. Xiao, J., Fiscella, K. A. & Gill, S. R. Oral microbiome: possible harbinger for children's health. *International Journal of Oral Science* vol. 12 (2020).

129. Zhang, R., Lahens, N. F., Ballance, H. I., Hughes, M. E. & Hogenesch, J. B. A circadian gene expression atlas in mammals: implications for biology and medicine. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 16219–16224 (2014).

130. Allaband, C. *et al.* Intermittent Hypoxia and Hypercapnia Alter Diurnal Rhythms of Luminal Gut Microbiome and Metabolome. *mSystems* vol. 6 (2021).

131. Kolodziejczyk, A. A., Zheng, D. & Elinav, E. Diet–microbiota interactions and personalized nutrition. *Nat. Rev. Microbiol.* **17**, 742–753 (2019).

132. Zaramela, L. S. *et al.* Gut bacteria responding to dietary change encode sialidases that exhibit preference for red meat-associated carbohydrates. *Nat Microbiol* **4**, 2082–2089 (2019).

133. Zmora, N., Suez, J. & Elinav, E. You are what you eat: diet, health and the gut microbiota. *Nat. Rev. Gastroenterol. Hepatol.* **16**, 35–56 (2019).

134. Etemadi, A. *et al.* Mortality from different causes associated with meat, heme iron, nitrates, and nitrites in the NIH-AARP Diet and Health Study: population based cohort study. *BMJ* **357**, j1957 (2017).

135. Koeth, R. A. *et al.* Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nat. Med.* **19**, 576–585 (2013).

136. Gilbert, J. A. *et al.* Current understanding of the human microbiome. *Nat. Med.* **24**, 392–400 (2018).

137. Khanolkar, R. A. *et al.* Ecological Succession of Polymicrobial Communities in the Cystic Fibrosis Airways. *mSystems* vol. 5 (2020).

138. Durack, J. & Lynch, S. V. The gut microbiome: Relationships with disease and opportunities for therapy. *J. Exp. Med.* **216**, 20–40 (2019).

139. van Nood, E., Dijkgraaf, M. G. W. & Keller, J. J. Duodenal infusion of feces for recurrent Clostridium difficile. *The New England journal of medicine* vol. 368 2145 (2013).

140. Tariq, R., Pardi, D. S., Bartlett, M. G. & Khanna, S. Low Cure Rates in Controlled Trials of Fecal Microbiota Transplantation for Recurrent Clostridium difficile Infection: A Systematic Review and Meta-analysis. *Clin. Infect. Dis.* **68**, 1351–1358 (2019).

141. Panigrahi, P. *et al.* Corrigendum: A randomized synbiotic trial to prevent sepsis among infants in rural India. *Nature* **553**, 238 (2018).

142. Halkjær, S. I. *et al.* Faecal microbiota transplantation alters gut microbiota in patients with irritable bowel syndrome: results from a randomised, double-blind placebo-controlled study. *Gut* **67**, 2107–2115 (2018).

143. Morton, J. T. *et al.* Learning representations of microbe–metabolite interactions. *Nat. Methods* **16**, 1306–1314 (2019).

144. Kehe, J. *et al.* Positive interactions are common among culturable bacteria. *Sci Adv* **7**, eabi7159 (2021).

145. Rubin, B. E. *et al.* Species- and site-specific genome editing in complex bacterial communities. *Nature Microbiology* 1–14 (2021).

146. 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).

147. Zeevi, D. *et al.* Personalized Nutrition by Prediction of Glycemic Responses. *Cell* **163**, 1079–1094 (2015).

148. Schooley, R. T. *et al.* Development and Use of Personalized Bacteriophage-Based Therapeutic Cocktails To Treat a Patient with a Disseminated Resistant Acinetobacter baumannii Infection. *Antimicrobial Agents and Chemotherapy* vol. 61 (2017).

149. Mu, A. *et al.* Effects on the microbiome during treatment of a staphylococcal device infection. (2021).

150. Claesson, M. J. *et al.* Gut microbiota composition correlates with diet and health in the elderly. *Nature* **488**, 178–184 (2012).

151. Wu, L. *et al.* A Cross-Sectional Study of Compositional and Functional Profiles of Gut Microbiota in Sardinian Centenarians. *mSystems* vol. 4 (2019).

152. Kong, F. *et al.* Gut microbiota signatures of longevity. *Curr. Biol.* **26**, R832–R833 (2016).

153. Claesson, M. J. *et al.* Composition, variability, and temporal stability of the intestinal microbiota of the elderly. *Proc. Natl. Acad. Sci. U. S. A.* **108 Suppl 1**, 4586–4591 (2011).

154. O'Toole, P. W. & Jeffery, I. B. Gut microbiota and aging. *Science* **350**, 1214–1215 (2015).

155. Shibagaki, N. *et al.* Aging-related changes in the diversity of women's skin microbiomes associated with oral bacteria. *Scientific Reports* vol. 7 (2017).

156. Liu, S., Wang, Y., Zhao, L., Sun, X. & Feng, Q. Microbiome succession with increasing age in three oral sites. *Aging* **12**, 7874–7907 (2020).

157. Schwartz, J. L. *et al.* Old age and other factors associated with salivary microbiome variation. *BMC Oral Health* **21**, 490 (2021).

158. Strati, F. *et al.* Age and Gender Affect the Composition of Fungal Population of the Human Gastrointestinal Tract. *Frontiers in Microbiology* vol. 7 (2016).

159. Wu, L. *et al.* Age-Related Variation of Bacterial and Fungal Communities in Different Body Habitats across the Young, Elderly, and Centenarians in Sardinia. *mSphere* **5**, (2020).

160. Nagpal, R. *et al.* Gut microbiome and aging: Physiological and mechanistic insights. *Nutr Healthy Aging* **4**, 267–285 (2018).

161. Wilmanski, T. *et al.* Gut microbiome pattern reflects healthy ageing and predicts survival in humans. *Nat Metab* **3**, 274–286 (2021).

162. Huang, S. *et al.* Human Skin, Oral, and Gut Microbiomes Predict Chronological Age. *mSystems* **5**, (2020).

163. Gill-King, H. Chemical and ultrastructural aspects of decomposition. in *Forensic taphonomy: the postmortem fate of human remains* 93–108 (CRC Press, 1997).

164. Janaway, R. C., Percival, S. L. & Wilson, A. S. Decomposition of Human Remains. *Microbiology and Aging* 313–334 (2009) doi:10.1007/978-1-59745-327-1_14.

165. Forbes, S. L., Perrault, K. A. & Comstock, J. L. Microscopic post-mortem changes: The chemistry of decomposition. in *Taphonomy of Human Remains: Forensic Analysis of the Dead and the Depositional Environment* 26–38 (John Wiley & Sons, Ltd, 2017).

166. Heimesaat, M. M. *et al.* Comprehensive Postmortem Analyses of Intestinal Microbiota Changes and Bacterial Translocation in Human Flora Associated Mice. *PLoS ONE* vol. 7 e40758 (2012).

167. Parkinson, R. A. *et al.* Microbial Community Analysis of Human Decomposition on Soil. in *Criminal and Environmental Soil Forensics* (eds. Ritz, K., Dawson, L. & Miller, D.) 379–394 (Springer Netherlands, 2009).

168. Metcalf, J. L. *et al.* Microbial community assembly and metabolic function during mammalian corpse decomposition. *Science* **351**, 158–162 (2016).

169. DeBruyn, J. M. & Hauther, K. A. Postmortem succession of gut microbial communities in deceased human subjects. *PeerJ* vol. 5 e3437 (2017).

170. Pechal, J. L., Schmidt, C. J., Jordan, H. R. & Benbow, M. E. A large-scale survey of the postmortem human microbiome, and its potential to provide insight into the living health condition. *Sci. Rep.* **8**, 5724 (2018).

171. Kodama, W. A. *et al.* Trace Evidence Potential in Postmortem Skin Microbiomes: From Death Scene to Morgue. *Journal of Forensic Sciences* vol. 64 791–798 (2019).

172. Hauther, K. A., Cobaugh, K. L., Jantz, L. M., Sparer, T. E. & DeBruyn, J. M. Estimating Time Since Death from Postmortem Human Gut Microbial Communities. *J. Forensic Sci.* **60**, 1234–1240 (2015).

173. Burcham, Z. M. *et al.* Fluorescently labeled bacteria provide insight on post-mortem microbial transmigration. *Forensic Science International* vol. 264 63–69 (2016).

174. Burcham, Z. M. *et al.* Bacterial Community Succession, Transmigration, and Differential Gene Transcription in a Controlled Vertebrate Decomposition Model. *Frontiers in Microbiology* vol. 10 (2019).

175. Balzan, S., de Almeida Quadros, C., de Cleva, R., Zilberstein, B. & Cecconello, I. Bacterial translocation: Overview of mechanisms and clinical impact. *Journal of Gastroenterology and Hepatology* vol. 22 464–471 (2007).

176. Metcalf, J. L. *et al.* A microbial clock provides an accurate estimate of the postmortem interval in a mouse model system. *eLife* vol. 2 (2013).

177. Hyde, E. R., Haarmann, D. P., Petrosino, J. F., Lynne, A. M. & Bucheli, S. R. Initial insights into bacterial succession during human decomposition. *International Journal of Legal Medicine* vol. 129 661–671 (2015).

178. Javan, G. T., Finley, S. J., Smith, T., Miller, J. & Wilkinson, J. E. Cadaver Thanatomicrobiome Signatures: The Ubiquitous Nature of Clostridium Species in Human Decomposition. *Frontiers in Microbiology* vol. 8 (2017).

179. Johnson, H. R. *et al.* A Machine Learning Approach for Using the Postmortem Skin Microbiome to Estimate the Postmortem Interval. *PLoS One* **11**, e0167370 (2016).

180. Belk, A. *et al.* Microbiome Data Accurately Predicts the Postmortem Interval Using Random Forest Regression Models. *Genes* vol. 9 104 (2018).

181. Metcalf, J. L. Estimating the postmortem interval using microbes: Knowledge gaps and a path to technology adoption. *Forensic Science International: Genetics* vol. 38 211–218 (2019).

182. Deel, H. *et al.* A Pilot Study of Microbial Succession in Human Rib Skeletal Remains during Terrestrial Decomposition. *mSphere* vol. 6 (2021).

183. Metcalf, J. L. *et al.* Microbiome Tools for Forensic Science. *Trends Biotechnol.* **35**, 814–823 (2017).

184. Nguyen, T. T., Hathaway, H., Kosciolek, T., Knight, R. & Jeste, D. V. Gut microbiome in serious mental illnesses: A systematic review and critical evaluation. *Schizophr. Res.* **234**, 24–40 (2021).

185. Jeste, D. V., Koh, S. & Pender, V. B. Perspective: Social Determinants of Mental Health for the New Decade of Healthy Aging. *The American Journal of Geriatric Psychiatry* (2022) doi:10.1016/j.jagp.2022.01.006.

186. Matijašić, M. *et al.* Gut Microbiota beyond Bacteria-Mycobiome, Virome, Archaeome, and Eukaryotic Parasites in IBD. *Int. J. Mol. Sci.* **21**, (2020).

187. Morton, J. T. *et al.* Establishing microbial composition measurement standards with reference frames. *Nat. Commun.* **10**, 2719 (2019).

188. Gerber, G. K. The dynamic microbiome. *FEBS Lett.* **588**, 4131–4139 (2014).

189. Zarrinpar, A., Chaix, A., Yooseph, S. & Panda, S. Diet and feeding pattern affect the diurnal dynamics of the gut microbiome. *Cell Metab.* **20**, 1006–1017 (2014).

190. Vázquez-Baeza, Y. *et al.* Guiding longitudinal sampling in IBD cohorts. *Gut* vol. 67 1743–1745 (2018).

191. Kane, P. B., Bittlinger, M. & Kimmelman, J. Individualized therapy trials: navigating patient care, research goals and ethics. *Nat. Med.* **27**, 1679–1686 (2021).

192. McCall, L.-I. *et al.* Home chemical and microbial transitions across urbanization. *Nat Microbiol* **5**, 108–115 (2020).

193. Abdill, R. J., Adamowicz, E. M. & Blekhman, R. Public human microbiome data dominated by highly developed countries. *bioRxiv* 2021.09.02.458641 (2021) doi:10.1101/2021.09.02.458641.

194. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* vol. 551 457–463 (2017).

195. Marotz, C. A. *et al.* Improving saliva shotgun metagenomics by chemical host DNA depletion. *Microbiome* **6**, 42 (2018).

196. Barlow, J. T., Bogatyrev, S. R. & Ismagilov, R. F. A quantitative sequencing framework for absolute abundance measurements of mucosal and lumenal microbial communities. *Nat. Commun.* **11**, 2590 (2020).

197. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).

198. Lagier, J.-C. *et al.* Culturing the human microbiota and culturomics. *Nat. Rev. Microbiol.* **16**, 540–550 (2018).

199. Bauermeister, A., Mannochio-Russo, H., Costa-Lotufo, L. V., Jarmusch, A. K. & Dorrestein, P. C. Mass spectrometry-based metabolomics in microbiome investigations. *Nat. Rev. Microbiol.* (2021) doi:10.1038/s41579-021-00621-9.

200. Vangay, P. *et al.* Microbiome Metadata Standards: Report of the National Microbiome Data Collaborative's Workshop and Follow-On Activities. *mSystems* vol. 6 (2021).

201. Janssen, S. *et al.* Phylogenetic Placement of Exact Amplicon Sequences Improves Associations with Clinical Information. *mSystems* **3**, (2018).

202. McDonald, D. *et al.* An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* **6**, 610–618 (2012).

Chapter 2. Robust Aitchison PCA reveals microbiome perturbations

Abstract

The central aims of many host or environmental microbiome studies is to elucidate factors associated with microbial community compositions, and to relate microbial features to outcomes. However, these aims are often complicated by difficulties stemming from high-dimensionality, non-normality, sparsity, and the compositional nature of microbiome datasets. A key tool in microbiome analysis is beta diversity, underpinned by definitions of the distance between two samples and resulting in a sample-by-sample distance matrix. Many different distance metrics have been proposed, all with varying discriminatory power on data with differing characteristics. Here, we propose a compositional beta diversity metric rooted in a center log-ratio transformation and matrix completion called Robust Aitchison PCA. We demonstrate the benefits of compositional transformations upstream of beta diversity calculations through simulations. We then demonstrate consistently improved effect size and classification accuracy over the current state of the art on several decreased samples subsets of real microbiome datasets. Finally, we highlight the ability of this new beta diversity metric to retain the feature loadings linked to sample ordinations revealing salient inter-community niche feature importance.

Importance

By accounting for the sparse compositional nature of microbiome datasets, Robust Aitchison PCA can yield high discriminatory power and salient feature ranking between microbial niches. The software to perform this analysis is available under an open-source license and can be obtained at https://github.com/cameronmartino/DEICODE, additionally a QIIME 2 plugin is provided to perform this analysis.

2.1. Introduction

Beta diversity is an ecological concept that describes differentiation in taxonomic or phylogenetic composition between communities. Beta diversity methods are a major component of many microbiome statistical analysis pipelines. These analyses enable an overview of complex microbial communities, identifying environmental factors differentiating microbial communities. However, there are dozens of distance metrics available to microbial ecologists to analyze their data, with each distance metric tailored to capture specific data characteristics. Beta diversity plots can therefore look dramatically different depending on the distance metric chosen, contributing to differences in interpretation of raw data (1).

One major confounding factor in beta diversity analysis is that microbiome datasets are sparse (i.e. most microorganisms are not found in most datasets), which has been shown to give rise to spike and horseshoe patterns in ordination plots (2, 3), complicating analysis. Furthermore, principal component analysis (PCA) has common assumptions of normally distributed and linearly related variables, often violated by biological data (4–7). As a result, classical distance metrics that only take into account the presence/absence of taxa, such as the Jaccard index, or metrics that explicitly account for relative abundances, such as Bray-Curtis symmetrized distance, are commonly used. Microbial beta diversity estimation was greatly improved with the incorporation of phylogenetic information, as was shown with UniFrac (8), which can be used as either a presence/absence (unweighted) or relative abundance (weighted) metric. However, presence/absence methods often yield substantial differences between communities that are obscured by abundance-based methods. This might seem paradoxical, because abundance-based methods are integrating more information about the community - it is counter-intuitive that such methods would reduce the signal compared to their presence/absence-based counterparts. However, if the key players are rare rather than abundant species, or if abundant species display

large fluctuations unrelated to function, abundance information may obscure rather than clarify the result, even with phylogenetic metrics (9).

This phenomenon can arise from mathematical problems rather than from real biology. Failure to reveal associations between phenotype and the microbiome overall may also be symptoms of methods that do not properly account for the relative changes of microbial taxa abundances. To demonstrate this principle, consider the scenario in Figure 2.1A, where three taxa are simulated over time. In this scenario, Taxon 1 has a much lower abundance than the other two taxa, but it is growing exponentially over time. Taxon 2 has a high abundance and is stable over time. Taxon 3 also has a high abundance but fluctuates randomly. The Euclidean distance between the first community and the other two time points is extremely variable, and does not capture the change induced by the exponential growth of Taxon 2. This variability in the Euclidean distance is largely driven by the random fluctuations in the high-abundance taxa.

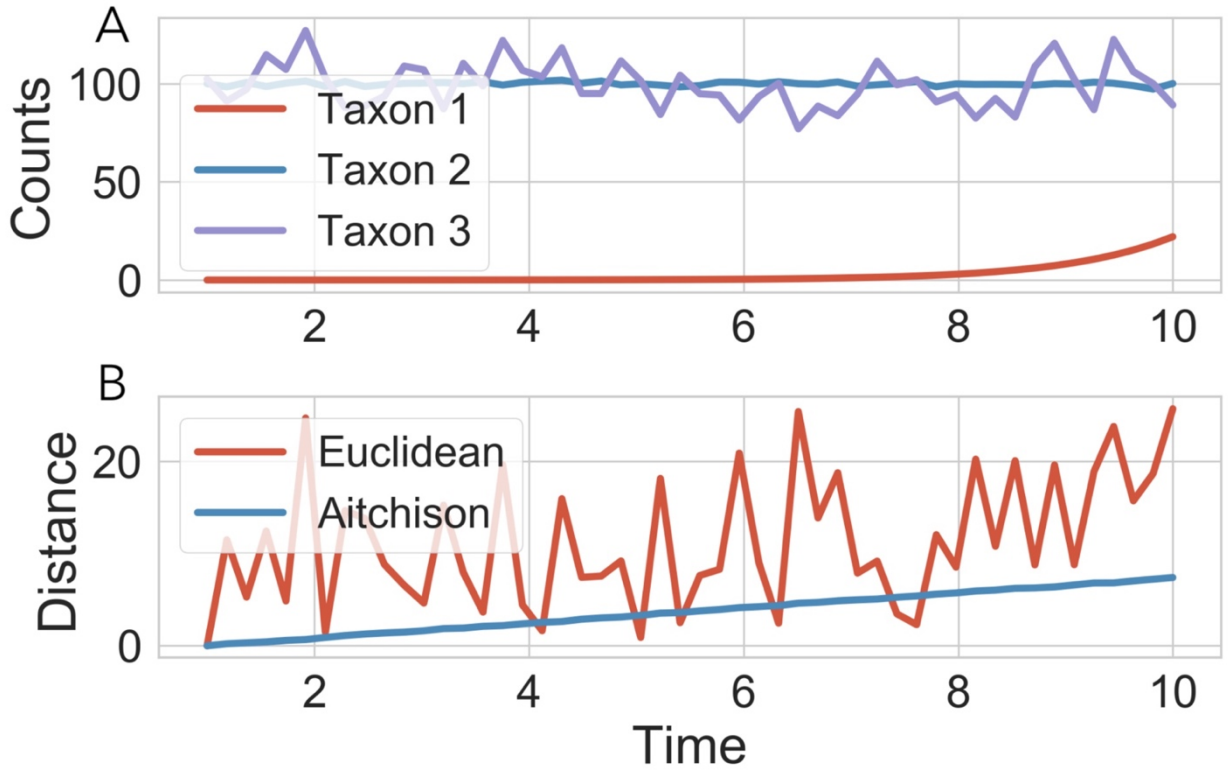

Figure 2.1. Benchmarking the rclr preprocessing step. Toy example with simple 3-taxa community sampled over time (A). Distance calculated between the t=1 community and subsequent communities demonstrates the robustness of Aitchison distance compared to Euclidean distance (B).

In contrast to Euclidean distance, compositional distance metrics, such as the Aitchison distance (Equation 2), can properly account for such relative changes (10). Here, the Aitchison distance only factors in the log fold change, reflecting the fact that deviations in the high abundance taxa are large on an absolute scale but small on a relative scale. The difference between 100 counts and 120 counts is 20 counts, which is large compared to the abundance of the first taxon, but is only a 20% increase. In contrast, the first taxon increased around 2,000%, and as a result, the Aitchison distance is driven by the large changes in the low-abundance species.

Aitchison distance is sensitive to relative changes between samples. As a result, microbes that display large fold change across samples will be weighted more heavily in the calculation of the Aitchison distance. However, this distance metric cannot handle zeros, and is thus challenging to apply to the sparse datasets that characterize microbiome studies. Here we propose a novel, compositional distance metric that can also explicitly handle sparse data through the use of matrix completion.

Matrix completion was originally developed in the context of recommender systems to predict user-item ratings (11) as a natural solution for handling sparse data. For example, the Netflix database contains a matrix detailing all customers by all movies where the entries are the movie ratings. However, each user only rates a small portion of the possible movies available on Netflix, so that only about 1% of the database contains non-zero values. As a result, when trying to recommend specific movies to specific customers, models need to be trained on the available ratings that customers have provided. Matrix completion tasks have become one of the state-of-the-art methods for performing these sorts of tasks.

Here, using simulation benchmarks and two case studies, we demonstrate the utility of preprocessing sparse microbiome datasets with matrix completion to allow compositional ordination and to preserve information about the features driving differences among samples.

2.2. Results

Description of Robust Aitchison PCA

Matrix completion can be interpreted as a robust dimensionality reduction technique, where PCA is performed accounting only for the observed entries (i.e. ignoring the zeros). Matrix completion relies on two major assumptions. First, it assumes that data are missing at random, meaning that the missing entries in the matrix are uniformly distributed. Second, because matrix completion is a robust form of PCA, it assumes that the data are normally distributed and centered around zero (12). To meet this assumption, a commonly applied approach is to subtract the row and column means (13, 14). However, because microbiome sequencing data are represented as counts (15), the data are strictly positive and skewed towards zero, which confounds PCA. A workaround is to first log transform the nonzero values before centering the data - we will refer to this preprocessing procedure as the robust center log ratio (rclr) due to its links to the clr transform commonly used in compositional data analysis (10) (Fig. 2.2A-B). A similar procedure using interquartiles was suggested previously (16).

This procedure produces a transformed table with missing values that can be used as input for matrix completion, or robust principal components analysis (RPCA), which provides the sample and feature loadings. These sample and feature loadings contain the ordination information directly used in beta diversity plotting and feature biclustering (Fig. 2.2C-E). Because PCA preserves feature information, we can use the feature loadings to determine which taxa drive the differences among sample types (Fig. 2.2F).

Simulations

To benchmark the effectiveness of the rclr preprocessing step we generated simulations from a study comparing microbial communities on keyboards and human fingertips (keyboard dataset) (17) (see Methods for detail). Simulated data was chosen as an initial proof-of-concept

benchmark due to the ease of changing dataset characteristics across which to interrogate, here the primarily focus was on sequencing depth.

The simulated data was generated with two clusters over varying sequencing depths from 1,000 to 10,000 reads per sample. At each sequencing depth, the output of the RPCA with and without the rclr transformation was compared by Kullback-Leibler divergence (KL) (18) to the simulation ground truth between rclr preprocessed and raw count data. Additionally, ordination output was compared by Permutational multivariate analysis of variance (PERMANOVA) F-statistic and supervised k-nearest neighbor (KNN) classification cross-validation (40:60) split.

When rclr preprocessing was applied, we saw a decrease in mean KL, demonstrating a more closely matched probability distribution when using the rclr (Fig 2.3A). Furthermore, when the rclr was applied, the F-statistic demonstrated a 4-fold increase (Fig 2.3B) and KNN classification accuracy (Fig 2.3C) increased by between 30-40%. All of the metrics, when applied to rclr RPCA, improved as the sequencing depth improved, following the logic that a good fit should increase performance as sequencing depth increases. A negative-control simulation with no group discrimination revealed no biclustering, RPCA clustering (Fig. 2.3E), low KNN classification accuracy, and PERMANOVA significance compared to a positive control (Fig. 2.3D) with two distinct groups (see Table AA.1.S1 in the supplemental material). This demonstrates a proof of concept that rclr is less affected by outliers, and is reliably reproducible at low and high sequencing depths.

Case Studies

Next, we demonstrated the utility of RPCA compared to the current state of the art. To do this we used two 16S rRNA gene amplicon sequencing datasets. The first dataset is a subset of the Sponge Microbiome Project (sponges) (19), and we compared sponge microbial communities classified by health status (i.e. stressed or healthy). The second dataset derives from a sleep apnea study; it consists of mouse fecal samples and focuses on comparing the gut microbiome

of animals exposed to intermittent hypoxia and hypercapnia (IHH; as a model of obstructive sleep apnea) to controls exposed to room air (air) (20).

Many different metrics exist for beta diversity distance comparison. We compared RPCA to two of the most commonly employed abundance based methods, Bray-Curtis and Weighted UniFrac, over 10-fold random subsamples of the data. The distances between the highlighted metadata categories for the two datasets were compared over subsamples with PERMANOVA (Fig. 2.4A,C). The Principal coordinates analysis (PCoA) was compared by supervised KNN classification cross-validation (40:60 split) accuracy for both datasets over subsamples (Fig. 2.4B,D). In all subsample comparisons the Robust Aitchison (distance metric derived from RPCA) outperformed Bray-Curtis and Weighted UniFrac. The results are qualitatively demonstrated in the PCoA clustering between metadata categories for low and high subsample depths (Fig. 2.4E,F).

A key benefit of RPCA over metrics, such as Weighted UniFrac and Bray-Curtis, is direct access to the feature loadings. With Euclidean distance it is also possible to obtain feature loadings, but due to multiple undesirable properties of Euclidean distance, such as artifacts in clustering patterns (2) and weak discrimination in high dimensional sparse data (2, 6, 7, 21, 22), these values are not reliable. Feature loadings allow us to rank the taxa in the data in relation to the samples and the metadata. When sorted, often referred to as biclustering, this method results in a table that reveals which taxa are driving the clustering seen in the ordinations.

In this case, we have a two-block table represented by clr-transformed heatmaps for the sponges (Fig. 2.5A) and sleep apnea (Fig. 2.5B) datasets. It is evident from the heatmap and ordination plots that there are some taxonomic abundance changes between the categories that are dividing the clusters. In order to compare two taxa directly we applied log ratios on highly weighted features. The highest loaded features (positive and negative maximums) correspond to the most influential taxa driving the clustering. To visualize these changes, the log ratios of highly

and lowly ranked microbes were compared between the sample loading (PC1) clusters of the sponge (Fig. 2.5C) and sleep apnea (Fig. 2.5D) datasets. The highly and lowly ranked log ratios revealed a strong (R^2 = 0.97 and 0.93) and weak (R^2 = 0.26 and 0.36) Pearson correlation to the sample PC1, respectively.

The highly weighted log ratios in the sponge case study indicate that two sOTUs can explain a great deal of variation between healthy and thermally stressed sponges. The sOTUs most strongly associated with healthy and stressed sponges, respectively, were classified at the lowest assignment level to *Candidatus* Synechococcus spongiarum (species, numerator) and *Nitrosopumilus* (genus, denominator). Both of these groups are known sponge symbionts (23, 24). *Nitrosopumilus* are ammonia-oxidizing archaea, which nitrify ammonia to nitrate and nitrification by sponge-associated microbiota is thought to remove ammonia waste produced by the host sponge (23, 25). It has been proposed that ammonium, urea, and creatine leaking from host sponge tissue could promote growth of *Nitrosopumilus (26)*, and this leakage may be more active in stressed hosts. *Candidatus* Synechococcus spongiarum have been found in numerous sponge species around the globe (24) and their photosynthetic products may contribute to host nutrition (27). From this analysis, this sOTU and several other sOTUs of *Candidatus* Synechococcus spongiarum (28) appear to be strongly associated with healthy sponges relative to stressed sponges.

In the sleep apnea dataset, the highly weighted log ratios revealed a strong clustering of air vs. IHH. These sOTUs were classified as Coriobacteriaceae (family) and *Clostridium* (genus). This trend was also observed by Tripathi et al. (20) where it was corroborated by the perturbations in the small molecular products attributed to members of these taxonomic classes. For example, changes in *Clostridium* were reflected in downstream changes in intestinal bile acids as members of this genera are known to transform bile acids (29). Previous studies (30, 31) have also reported changes in these taxonomic classes in cardio-metabolic comorbidities of sleep apnea, which suggests that our method potentially guides biologically relevant observations.

2.3. Discussion

Here we demonstrated the ability of rclr preprocessing and RPCA to reveal salient, beta diversity ordination and factor loading. We demonstrated through simulations that rclr preprocessing dramatically improved RPCA. In two case studies (sponge and sleep apnea), RPCA presented higher PERMANOVA F-statistics and KNN classifier accuracy in small subsamples of the data. In addition, RPCA revealed qualitatively increased the discriminative ability of clusters obtained from the ordination than beta diversity techniques widely used in the field, both at low and high levels of subsampling.

We have shown that Aitchison distance has numerous other desirable properties, such as scale invariance, negating the need to perform rarefaction. This feature is critical when one lacks access to absolute microbial abundance, because scale invariant distances ensure equivalence between distances computed from absolute and relative abundance measurements (See Methods for equation). Aitchison distance is also known to be subcompositionally coherent (32). This guarantees that distances will never decrease if additional taxa are observed (e.g. by using PCR primers with broader specificity), which has important implications for reproducibility across distance-based analyses, especially across studies that use different molecular methods.

The increased cluster separation at smaller subsamples of the dataset highlight the compositionaly coherent properties of the method. Significant partitioning of sample categories on smaller sample cohorts is particularly important in a clinical setting, due to the difficulty of large volume sample collection. In addition, rapid resolutions of taxa driving ordination is of principal importance in translational results.

Importantly, because RPCA provides linked sample and feature information, one can directly identify which taxa are likely driving sample clustering (which are typically separate workflows in canonical amplicon analysis). However, RPCA does not currently take into account

phylogenetic relationships among features. Adding this component to the current workflow could potentially improve the resulting ordinations.

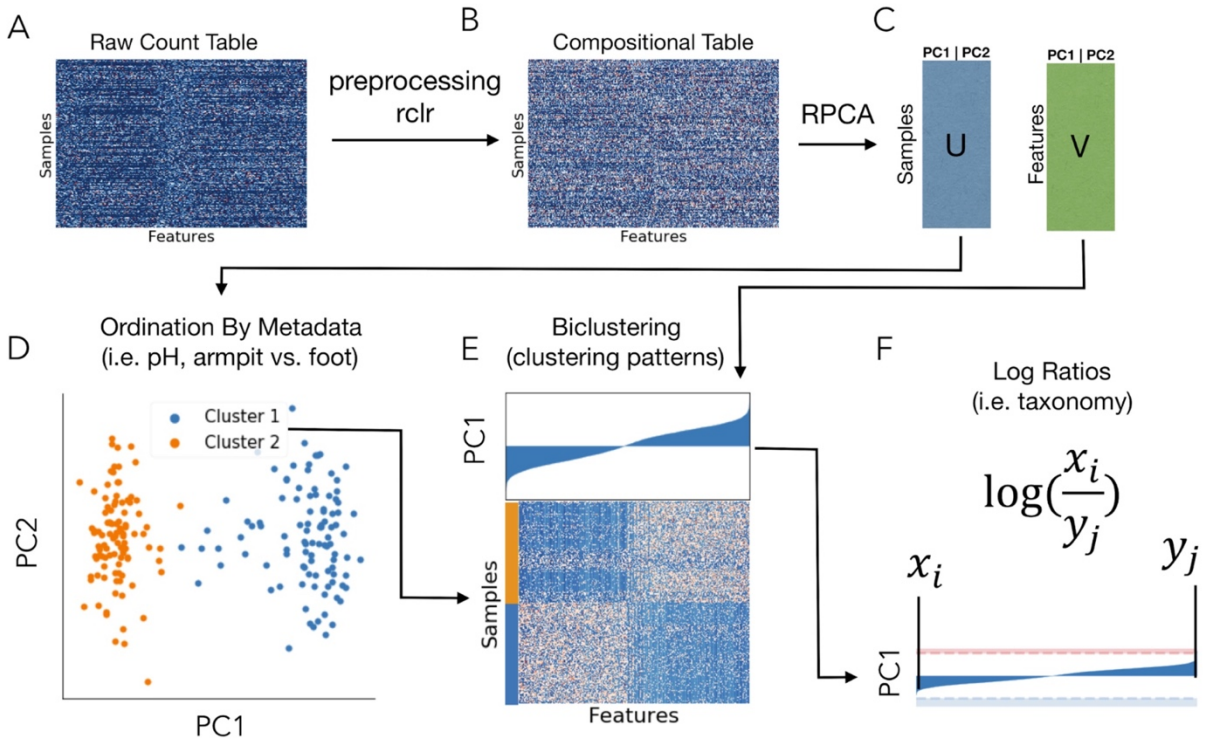

Figure 2.2. A general overview of the workflow. (A) A sparse, raw sequencing count table with samples on the y-axis and features (i.e. OTUs, Genes) on the x-axis. (B) The data is preprocessed by a robust centered log ratio transform (rclr) on only the known (non-zero) values. (C) Matrix completion with a robust principal component analysis (RPCA) that operates on only the observed values in the table resolves a loading by samples and by features. These loadings can be directly used for ordination (D), biclustering (E), and the identification of important taxa driving clustering in both the previous plots (F).

Additionally, much work remains to better understand matrix completion methods, particularly in the context of compositional data analysis. Previous methods have been developed to handle zeros in compositional datasets. In particular, zCompositions (33) contains several methods that could potentially be adapted to microbiome datasets. However, these algorithms are not currently appropriate for microbiome datasets, because the number of microbial features typically vastly outnumber the number of samples within an experiment. With matrix completion techniques, it may be possible to extend existing compositional methods to handle missing data.

Furthermore, overfitting is still a topic that must be addressed with these methods. Given the high dimensional nature of microbial datasets, the number of parameters required to fit robust principal components can grow very quickly. As a result, it is still possible to overfit these methods, making them potentially sensitive to outliers and reducing their predictive power (34), although we did not notice these effects in our simulations. We therefore recommend starting fitting RPCA models with a low rank of either two or three.

A low rank constraint can possibly cause misleading results in the case of high rank datasets. High rank datasets may occur in microbiome datasets as a gradient between samples and features. To give intuition of what types of data may contain high rank structure we provide two published examples. The first example is a study of soil microbiomes representing different pH environments (35)(see Methods for detail). The second example is a case study of the gut colonization of an infant over time (36) (see Methods for detail). In both cases, a gradient forms because very few samples contain similar microbes (see Fig. 2.S1 in the supplemental material). For example, in the infant development study very few microbes are shared between subsequent samples over time. Although the rclr transform eases the problem, it can still lead to misinterpretation in ordination (see Fig. 2.S2 in the supplemental material). There are many possible future directions for incorporating regularization or Bayesian priors to better fit these models.

In light of the current limitations, we have shown that matrix completion resolves numerous outstanding problems in beta diversity analysis including sparsity, compositional effects, and uneven sequencing depths, all while giving information about the taxa driving microbial perturbations. This method can also be applied to or combined with other omics paradigms (e.g. metabolomics, metatranscriptomics, and metagenomics), and provides the opportunity to initiate standardization of beta diversity analyses in the microbiome field.

2.4. Methods

Preprocessing with rclr

Prior to running matrix completion, the data needs to be centered around zero and approximately normally distributed. Centered log ratio (clr) transformation is commonly applied in compositional data analysis before applying PCA. This log transforms each value then centers them around zero. This is particularly useful when one assumes that the data is lognormally distributed as proposed in (37), since log transformed lognormal distributed data is normally distributed. The clr transform is given below.

$$clr(\vec{x}) = [log \frac{x_1}{g(\vec{x})}, \ldots, log \frac{x_D}{g(\vec{x})}] = log\, \vec{x} - \bar{log\, \vec{x}} \quad (1)$$

Where $g(\vec{x})$ is the geometric mean of all of the taxa. The Aitchison distance can be directly calculated from the Euclidean distance of the clr transformed data. This is given as follows

$$d_A(x, y) = \sqrt{\sum_{i=1}^{D} (clr(x)_i - clr(y)_i)^2} = \sqrt{\sum_{i=1}^{D} (log \frac{x_i}{x_j} - log \frac{y_i}{y_j})^2} \quad (2)$$

One can show that this transformation is scale variant as follows

$$d_A(x,y) = \sqrt{\sum_{i=1}^{D} \left(log\frac{x_i}{x_j} - log\frac{y_i}{y_j}\right)^2} = \sqrt{\sum_{i=1}^{D} \left(log\frac{np_{x_i}}{np_{x_j}} - log\frac{np_{y_i}}{np_{y_j}}\right)^2} = \sqrt{\sum_{i=1}^{D} \left(log\frac{p_{x_i}}{p_{x_j}} - log\frac{p_{y_i}}{p_{y_j}}\right)^2} = d_A(p_x, p_y) \quad (3)$$

The Aitchison distance between the absolute abundances is equivalent to the Aitchison distance on the proportions.

In order to center the samples around zero, the average clr transformed sample needs to be calculated, then subtracted from the remaining samples. Thus, the clr transformed results will be as follows:

$$y_{ij} = log\, x_{ij} - log\bar{}\,\overrightarrow{x_{i.}} - log\bar{}\,\overrightarrow{x_{.J}} \quad (4)$$

This centering procedure is commonly used prior to performing PCA and eliminates the need to explicitly compute bias constants (38).

The issue with applying the clr transform directly to sparse count data is that the log of zero is undefined. This motivated the construction of an approximate clr transform only defined on non-zero counts. The robust clr transform is given as follows:

$$rclr(\vec{x}) = \left[log\frac{x_1}{g_r(\vec{x})}, \ldots, log\frac{x_D}{g_r(\vec{x})}\right] \quad (5)$$

$$g_r(\vec{x}) = \left(\prod_{i \in \Omega_x} x_i\right)^{1/|\Omega_x|} \quad (6)$$

Where x_i is the abundance of taxa i, Ω_x is the set of observed taxa in sample x and $g_r(x)$ is the geometric mean only defined on observed taxa. The rationale behind this procedure is that due to the high dimensionality of these datasets, the robust geometric mean (the geometric mean of the log-transformed non-zero data) can serve as an approximation to the true geometric mean. We know from the Central Limit Theorem that as we collect more independent measurements we approach the true geometric mean:

$$\frac{1}{|\Omega_x|}\sum_{i\in\Omega_x} x_i \rightarrow E[log\ \vec{x}] \text{ as } |\Omega_x| \rightarrow |\vec{x}| \quad (7)$$

From this we can redefine the transformed result as follows:

$$y_{ij} = log\ x_{ij} - \frac{1}{|\Omega_{x_{i.}}|}\sum_{k\in\Omega_{x_{i.}}} x_k - \frac{1}{|\Omega_{x_{.j}}|}\sum_{i\in\Omega_{x_{.j}}} x_k \quad (8)$$

Where y_{ij} is only defined when $x_{ij} > 0$. The matrix completion methods can then be directly applied to this transformed result.

Matrix completion

OptSpace is a matrix completion algorithm based on a singular value decomposition (SVD) optimized on a local manifold. It has been shown to be quite robust to noise in low-rank datasets (39). The objective function that it optimizes over is given by:

$$min_{u,v}|\Lambda(Y - USV^T)|_2{}^2 \quad (9)$$

where U and V are the matrices that are trying to be estimated and S is analogous to a matrix of eigenvalues. Y is the observed values and Λ is a function such that the errors between Y and USV are only computed on the nonzero entries.

Simulations

Simulations were designed to replicate real datasets with low-rank clusters as a proof of concept testing of OptSpace with and without the rclr preprocessing step. The keyboard dataset was chosen as a representative dataset to fit the simulation parameters due to the three distinct microbial community clusters observed in the study (M2, M3, and M9). Simulations were built by drawing blocks of Nsequences with the microbial proportions given as follows (40).

$$x_{ij} = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(\frac{(\mu_i - g_j)^2}{2\sigma^2}\right) \quad (10)$$

$$p_{ij} = \frac{x_{ij}}{\Sigma_k x_{kj}} \quad (11)$$

The resulting simulation was induced by multiple noise sources. The was normally distributed error that was applied to the entire matrix. There was also normally distributed errors that were randomly applied to a subset of the entries in the matrix. In addition, there were subsampling errors that were simulated from the Poisson-log normal (PLN) distribution with an overdispersion parameter ϕ (41) where the final subsampled simulation y_{ij} is represented by:

$$\lambda_{ij} = n\, p_{ij} \quad (12)$$

$$y_{ij} \sim PLN(\lambda_{ij}, \phi) \quad (13)$$

The resulting optimized parameters are optimized rank (number of clusters), the intensity of noise, sequencing depth, the distribution parameters(μ andσ), and overlap of features between clusters (i.e. effect size). To resolve the most realistic simulation possible these parameters were optimized to minimize the KL-divergence between the real data and the simulation with a Broyden–Fletcher–Goldfarb–Shanno (BFGS) optimization. The resolved parameters were used to run the simulation at a rank of 2 over sequencing depths ranging from 100 to 10,000 reads/sample. At each depth, before the introduction of noise and subsampling, the sampled data was stored as a base truth to be compared to the reconstruction. Furthermore, the same noisy and subsampled simulation was run with OptSpace with or without rclr preprocessing. The resulting matrix$USV^T$was compared by KL-Divergence to the base truth. The rclr preprocessed data was inverse transformed by taking the exponential of USV^Tbefore comparison to the base truth. In addition, the simulation, base truth, sample orientation U and feature loadings V were saved at each iteration and compared visually.

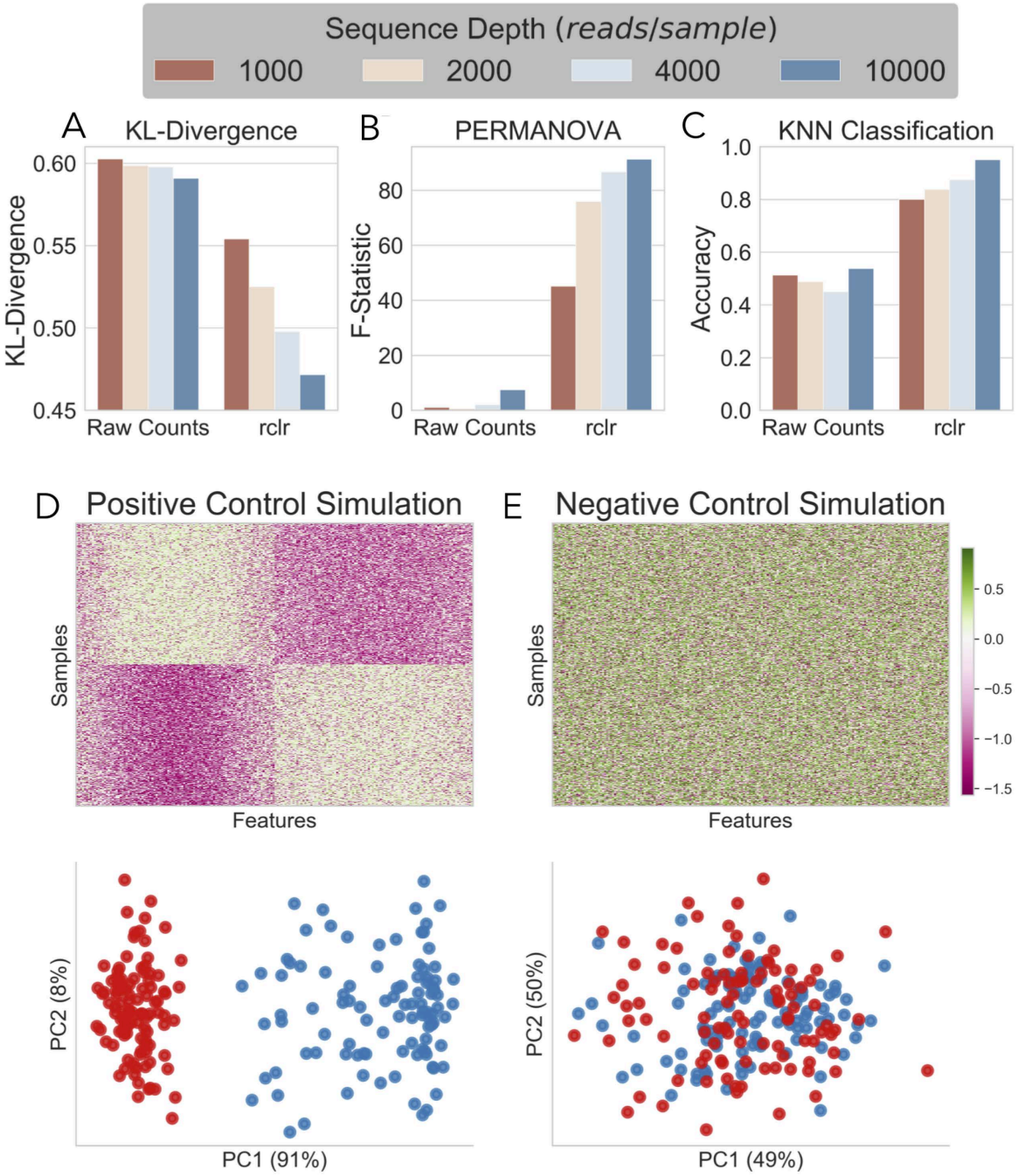

Figure 2.3. The robust centered log-ratio improves imputation and dimensionality reduction. (A) Comparison of KL-Divergence (y-axis) between simulated base truth data between RPCA output from raw count data and rclr preprocessed data. Comparison between RPCA ordination by PERMANOVA F-statistic (B) and KNN classifier accuracy (C). All at varying sequencing depths from 1000 to 10,000 reads per sample. (D and E) Comparison of positive- (D) and negative-control (E) simulation by biclustering (top) and RPCA ordination (bottom).

The simulation results of improved clustering at uneven sequencing depths was also compared in the real keyboard dataset (see case studies for data processing). The data was compared between two subjects at 500 and 100 reads/sample. Ordination and PERMANOVA results were compared for jaccard, bray-curtis and RPCA with rclr preprocessing. RPCA with rclr preprocessing alleviated the clustering by sequencing depth in the real dataset. This was seen both qualitatively (see Fig. 2.S2 in the supplemental material) and through the PERMANOVA F-statistic by subject id (see Table AA.1.S1 in the supplemental material).

Case studies

Case studies on real-world datasets were used to compare robust Aitchison PCA to the current state of the art in beta diversity comparison. The sponge,sleep apnea, infant, keyboard, and 88 soils datasets were acquired on 9/20/2018 from Qiita (42) with IDs of 10793,10422, 101, 232, and 103 respectively. Each dataset was run through Qiita with default trimming and Deblur (v. 1.1.0) sOTU (43) picking approach, using QIIME 2 (v. 2018.6.0) (44). The resulting biom (45) tables were then filtered for samples greater than 1000 reads per sample. Phylogeny was built using the most up to date GreenGenes using SEPP (46) and taxonomy was assigned through scikit-learn with default QIIME 2 parameters.

The sponge dataset was filtered using the metadata so that it only contained samples with either the label healthy or stressed. This resulted in a comparison with 248 remaining samples. Similarly, the sleep apnea study was filtered for IHH and air control samples, with a treatment duration of 6 weeks resulting in 184 remaining samples. The Infant gut colonization case study was filtered for samples over 500 reads/samples and for a single sample form the mother with the title 101.Mother. The 88 soils dataset was filtered for samples over 500 reads/samples. The keyboard dataset was filtered for samples over 500 reads/samples and 15 reads/sOTU. Additionally, only subject ids corresponding to M3, M2 and M9 were retained, giving 67 samples.

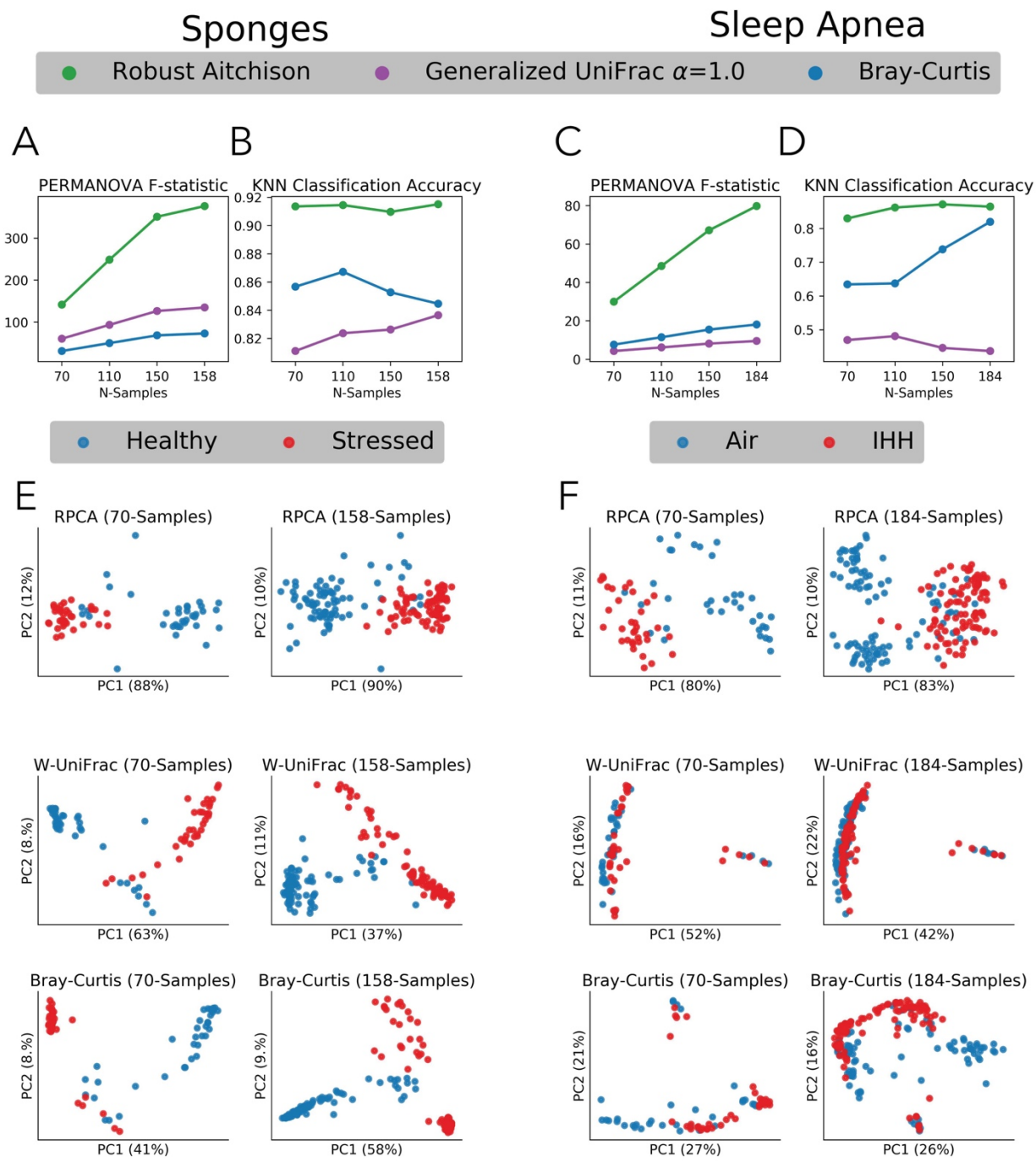

Figure 2.4. A case study of RPCA on real datasets; sponge (left panel, A,B,E) and sleep apnea (right panel C,D,F). PERMANOVA F test statistic (y-axis) (A,C) or KNN classifier accuracy (B,D) by subsamples of the datasets. Ordination plots between 70 samples total (left) and maximum number of samples (right) compared between RPCA (top) Generalized Weighted UniFrac (alpha=1) (middle) and Bray-Curtis (bottom) (E,F). Sponge dataset plotted between healthy (blue) and stressed (red) (E) along with sleep apnea dataset plotted between air (blue) and IHH (red).

Both datasets were then preprocessed with the robust centered log ratio (rclr) transform and RPCA was run with a rank of 2 because there were two metadata categories of interest in each comparison. Weighted UniFrac distances was calculated using Generalized UniFrac with an alpha of one (47). Bray-Curtis distances were calculated through QIIME 2 (44). Both Weighted UniFrac and Bray-Curtis distances were calculated on tables rarefied to 1000 reads per sample. PCoA and PERMANOVA analysis for the Bray-Curtis, RPCA distance matrix, and Weighted UniFrac were calculated through scikit-bio. The resulting PCoA and PCA axis were plotted through matplotlib (48) with PC1 and PC2 in the x and y-axis respectively.

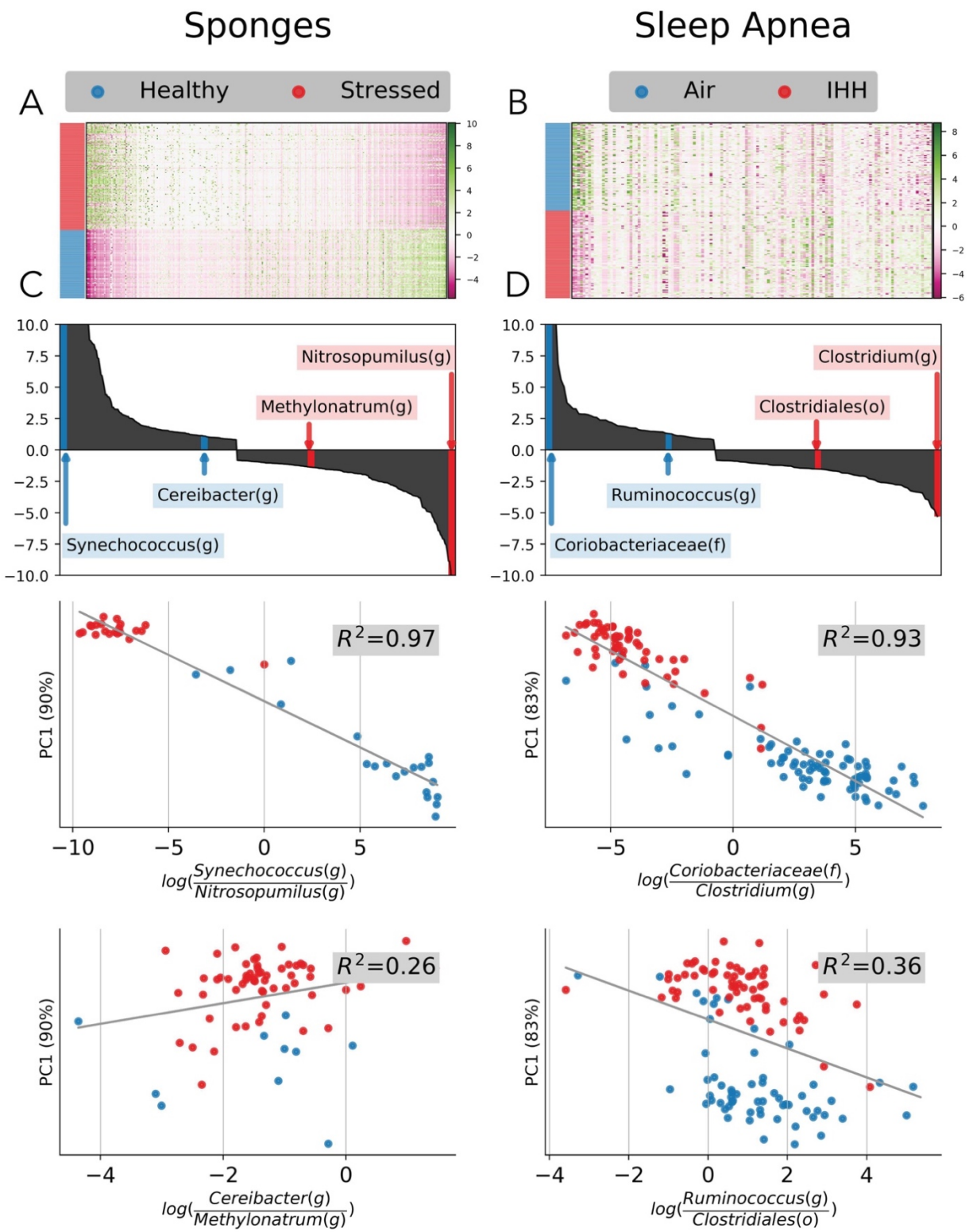

Figure 2.5. A case study of RPCA feature loadings on real datasets; sponge (left panel, A,C) and sleep apnea (right panel B,D). Heatmaps of clr transformed sOTU tables with samples sorted by metadata and features sorted by RPCA feature loadings (A,B). Absolute highest (middle) and lowest (bottom) feature loading sOTUs (top) plotted as log ratios (x-axis) by sample loading PC1 (y-axis).

The original unprocessed (raw count) tables were sorted by features loadings from RPCA. Features with a count sum of less than 10 across all samples were filtered out. The resulting table was then clr transformed with a pseudo-count of one and plotted as a heat map. Each sOTU was given the lowest classification for the sleep apnea and sponge datasets, respectively.

The features in the PC1 axis of the feature loadings from RPCA were selected to represent a manageable number of taxa to compare between sub-groups. Those selected features (sOTUs) from the feature loadings were used for log ratios. Log ratios were calculated from the table used to calculate them. The samples that contained zeros in either the numerator or denominator were removed before calculating the ratios. The correlations between the log ratio and PC1 axis were performed by pearson correlation via Scipy (49).

2.5. Acknowledgements

We would like to thank Todd P. Coleman (Department of Bioengineering, University of California at San Diego) for his insights into matrix completion. We thank Torsten Thomas for discussions about sponge-associated microbes. This material is based upon work supported by the National Science Foundation under Grant No. 1332344, the National Institutes of Health under Grant No. AR071731, and the U.S. Department of Energy, Office of Science, Office of Biological & Environmental Research Awards DE-SC0012658 and DE-SC0012586. Cameron Martino is supported in part by the UC San Diego Frontiers of Innovation Scholars Program (FISP). James T. Morton is supported in part by the National Science Foundation under Grant No. 1144086.

2.6. References

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

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

3. Morton JT, Toran L, Edlund A, Metcalf JL, Lauber C, Knight R. 2017. Uncovering the Horseshoe Effect in Microbial Analyses. mSystems 2.

4. Greig-Smith P. 1980. The development of numerical classification and ordination. Vegetatio 42:1–9.

5. Potvin C, Roff DA. 1993. Distribution-free and robust statistical methods: viable alternatives to parametric statistics. Ecology 74:1617–1628.

6. Tabachnick BG, Fidell LS. 2013. Using Multivariate Statistics: Pearson New International Edition. Pearson Education Limited.

7. Ginter JL, Thorndike RM. 1979. Correlational Procedures for Research. J Mark Res 16:600.

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

9. Lozupone CA, Hamady M, Kelley ST, Knight R. 2007. Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. Appl Environ Microbiol 73:1576–1585.

10. Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R. 2015. Modeling and Analysis of Compositional Data. John Wiley & Sons.

11. Candès EJ, Li X, Ma Y, Wright J. 2011. Robust principal component analysis? J ACM 58:1–37.

12. Tipping ME, Bishop CM. 1999. Probabilistic Principal Component Analysis. J R Stat Soc Series B Stat Methodol 61:611–622.

13. Jiang B, Ma S, Causey J, Qiao L, Hardin MP, Bitts I, Johnson D, Zhang S, Huang X. 2016. Corrigendum: SparRec: An effective matrix completion framework of missing data imputation for GWAS. Sci Rep 6:37365.

14. Cai T, Tony Cai T, Zhang A. 2016. Structured Matrix Completion with Applications to Genomic Data Integration. J Am Stat Assoc 111:621–633.

15. McMurdie PJ, Holmes S. 2014. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. PLoS Comput Biol 10:e1003531.

16. Finding the centre: corrections for asymmetry in high-throughput sequencing datasets.

17. Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R. 2010. Forensic identification using skin bacterial communities. Proc Natl Acad Sci U S A 107:6477–6481.

18. Kullback S, Leibler RA. 1951. On Information and Sufficiency. Ann Math Stat 22:79–86.

19. Moitinho-Silva L, Nielsen S, Amir A, Gonzalez A, Ackermann GL, Cerrano C, Astudillo-

Garcia C, Easson C, Sipkema D, Liu F, Steinert G, Kotoulas G, McCormack GP, Feng G, Bell JJ, Vicente J, Björk JR, Montoya JM, Olson JB, Reveillaud J, Steindler L, Pineda M-C, Marra MV, Ilan M, Taylor MW, Polymenakou P, Erwin PM, Schupp PJ, Simister RL, Knight R, Thacker RW, Costa R, Hill RT, Lopez-Legentil S, Dailianis T, Ravasi T, Hentschel U, Li Z, Webster NS, Thomas T. 2017. The sponge microbiome project. Gigascience 6:1–7.

20. Tripathi A, Melnik AV, Xue J, Poulsen O, Meehan MJ, Humphrey G, Jiang L, Ackermann G, McDonald D, Zhou D, Knight R, Dorrestein PC, Haddad GG. 2018. Intermittent Hypoxia and Hypercapnia, a Hallmark of Obstructive Sleep Apnea, Alters the Gut Microbiome and Metabolome. mSystems 3.

21. Dollhopf SL, Hashsham SA, Tiedje JM. 2001. Interpreting 16S rDNA T-RFLP Data: Application of Self-Organizing Maps and Principal Component Analysis to Describe Community Dynamics and Convergence. Microb Ecol 42:495–505.

22. Aggarwal CC, Hinneburg A, Keim DA. 2001. On the Surprising Behavior of Distance Metrics in High Dimensional Space, p. 420–434. *In* Database Theory — ICDT 2001. Springer Berlin Heidelberg.

23. Feng G, Sun W, Zhang F, Karthik L, Li Z. 2016. Inhabitancy of active Nitrosopumilus-like ammonia-oxidizing archaea and Nitrospira nitrite-oxidizing bacteria in the sponge Theonella swinhoei. Sci Rep 6:24966.

24. Usher KM. 2008. The ecology and phylogeny of cyanobacterial symbionts in sponges. Mar Ecol 29:178–192.

25. Diaz MC, Ward BB. 1997. Sponge-mediated nitrification in tropical benthic communities. Mar Ecol Prog Ser 156:97–107.

26. Moitinho-Silva L, Díez-Vives C, Batani G, Esteves AIS, Jahn MT, Thomas T. 2017. Integrated metabolism in sponge–microbe symbiosis revealed by genome-centered metatranscriptomics. ISME J 11:1651.

27. WATERBURY, B J. 1986. Biological and ecological characterization of the marine unicellular cyanobacterium Synechococcus. Photosynthetic Picoplankton 71–120.

28. Erwin PM, Thacker RW. 2008. Cryptic diversity of the symbiotic cyanobacterium Synechococcus spongiarum among sponge hosts. Mol Ecol 17:2937–2947.

29. Studer N, Desharnais L, Beutler M, Brugiroux S, Terrazos MA, Menin L, Schürch CM, McCoy KD, Kuehne SA, Minton NP, Stecher B, Bernier-Latmani R, Hapfelmeier S. 2016. Functional Intestinal Bile Acid 7α-Dehydroxylation by Clostridium scindens Associated with Protection from Clostridium difficile Infection in a Gnotobiotic Mouse Model. Front Cell Infect Microbiol 6:191.

30. Jie Z, Xia H, Zhong S-L, Feng Q, Li S, Liang S, Zhong H, Liu Z, Gao Y, Zhao H, Zhang D, Su Z, Fang Z, Lan Z, Li J, Xiao L, Li J, Li R, Li X, Li F, Ren H, Huang Y, Peng Y, Li G, Wen B, Dong B, Chen J-Y, Geng Q-S, Zhang Z-W, Yang H, Wang J, Wang J, Zhang X,

Madsen L, Brix S, Ning G, Xu X, Liu X, Hou Y, Jia H, He K, Kristiansen K. 2017. The gut microbiome in atherosclerotic cardiovascular disease. Nat Commun 8:845.

31. Kameyama K, Itoh K. 2014. Intestinal colonization by a Lachnospiraceae bacterium contributes to the development of diabetes in obese mice. Microbes Environ 29:427–430.

32. Greenacre M, Lewi P. 2005. Distributional Equivalence and Subcompositional Coherence in the Analysis of Contingency Tables, Ratio-Scale Measurements and Compositional Data.

33. Palarea-Albaladejo J, Martín-Fernández JA. 2015. zCompositions—R package for multivariate imputation of left-censored data under a compositional approach. Chemometrics Intellig Lab Syst 143:85–96.

34. Keshavan RH, Montanari A. 2010. Regularization for matrix completion, p. 1503–1507. *In* 2010 IEEE International Symposium on Information Theory.

35. Lauber CL, Hamady M, Knight R, Fierer N. 2009. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. Appl Environ Microbiol 75:5111–5120.

36. Koenig JE, Spor A, Scalfone N, Fricker AD, Stombaugh J, Knight R, Angenent LT, Ley RE. 2011. Succession of microbial consortia in the developing infant gut microbiome. Proc Natl Acad Sci U S A 108 Suppl 1:4578–4585.

37. Paulson JN, Stine OC, Bravo HC, Pop M. 2013. Differential abundance analysis for microbial marker-gene surveys. Nat Methods 10:1200–1202.

38. Abdi H, Williams LJ, Valentin D. 2013. Multiple factor analysis: principal component analysis for multitable and multiblock data sets. WIREs Comp Stat 5:149–179.

39. Keshavan RH, Oh S, Montanari A. 2009. Matrix completion from a few entries2009 IEEE International Symposium on Information Theory.

40. Atchison J, Shen SM. 1980. Logistic-normal distributions:Some properties and uses. Biometrika 67:261–272.

41. Aitchison J, Ho CH. 1989. The multivariate Poisson-log normal distribution. Biometrika 76:643–653.

42. Gonzalez A, Navas-Molina JA, Kosciolek T, McDonald D, Vázquez-Baeza Y, Ackermann G, DeReus J, Janssen S, Swafford AD, Orchanian SB, Sanders JG, Shorenstein J, Holste H, Petrus S, Robbins-Pianka A, Brislawn CJ, Wang M, Rideout JR, Bolyen E, Dillon M, Caporaso JG, Dorrestein PC, Knight R. 2018. Qiita: rapid, web-enabled microbiome meta-analysis. Nat Methods 551:457.

43. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, Kightley EP, Thompson LR, Hyde ER, Gonzalez A, Knight R. 2017. Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. mSystems 2.

44. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet C, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Titus Brown C, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope E, Da Silva R, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley G, Janssen S, Jarmusch AK, Jiang L, Kaehler B, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu Y-X, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton J, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Michael S Robeson II, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Knight R, Gregory Caporaso J. 2018. QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. e27295v1. PeerJ Preprints.

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

46. Janssen S, McDonald D, Gonzalez A, Navas-Molina JA, Jiang L, Xu ZZ, Winker K, Kado DM, Orwoll E, Manary M, Mirarab S, Knight R. 2018. Phylogenetic Placement of Exact Amplicon Sequences Improves Associations with Clinical Information. mSystems 3.

47. Chen J, Bittinger K, Charlson ES, Hoffmann C, Lewis J, Wu GD, Collman RG, Bushman FD, Li H. 2012. Associating microbiome composition with environmental covariates using generalized UniFrac distances. Bioinformatics 28:2106–2113.

48. Hunter JD. 2007. Matplotlib: A 2D Graphics Environment. Computing in Science Engineering 9:90–95.

49. Jones E, Oliphant T, Peterson P. 2001--. {SciPy}: Open source scientific tools for {Python}.

Chapter 3. Context-aware dimensionality reduction deconvolutes gut microbial community dynamics

3.1. Introduction

The translational power of human microbiome studies is limited by high inter-individual variation. We describe a dimensionality reduction tool, compositional tensor factorization (CTF), that incorporates information from the same host, across multiple samples, to reveal patterns driving differences in microbial composition across phenotypes. CTF identifies robust patterns in sparse, compositional datasets, allowing for the detection of microbial changes associated with specific phenotypes that are reproducible across datasets.

3.2. Discussion

Host-associated microbiomes are often host-specific, with the subject driving the majority of the variation. This host-specific variation can obscure microbial changes that are broadly associated with a given phenotype. Collecting multiple samples from the same participant, either longitudinally or from different body sites (i.e., “repeated measures”), is a valid experimental approach to control for inter-individual variation. However, there are multiple challenges to leveraging this type of experimental design due to the nature of microbiome sequencing datasets.

One common way to explore microbiome sequencing data is by performing dimensionality reduction on a distance matrix (e.g. principal coordinates analysis (PCoA)), which describes the relationship among samples, allowing global differences across a dataset to be observed. Nonetheless, when applied to repeated measures, this approach does not account for the inherent temporal or spatial correlation structure. An alternative to analyze repeated measures microbiome data is by using supervised methods, which are focused on generative models

inferring the dynamics of these communities (e.g., generalized Lotka Volterra)[1–4]. Although these methods account for the correlation structure induced by repeated measures, as well as for sparsity and compositionality, their output does not directly allow clustering of phenotypes by microbial community dynamics.

To address these challenges simultaneously, we developed compositional tensor factorization (CTF), which allows an unsupervised dimensionality reduction for repeated measures data, producing both a traditional beta-diversity analysis as well as a differential feature abundance assessment. In the first step, a two-dimensional matrix is transformed using the robust, centered-log-ratio technique[5] to account for the inherent sparse and compositional nature of next-generation sequencing datasets[6] (Fig. 3.1a). Next, this transformed matrix is restructured into a three-dimensional tensor, which relates microbial sequences, sampled host (or subject), and time or space (Fig. 3.1b). Decomposition (i.e., factorization) of this tensor provides distinct vectors for subjects (“U”), microbial features (“V”), and timepoints (“W”) (Fig. 3.1c). Analogous to the concept of reference frames[7], these vectors are unit-scaled and therefore can be ordered, where their ranking indicates their association to the underlying phenotypic groups. From here on we will refer to the ordering of these vectors as ‘rankings’ (i.e., “feature rankings”). Notably, CTF assumes the data harbors an underlying low-rank structure, where only a few phenotypic factors explain the majority of the variance[5] (Fig. 3.1d-g).

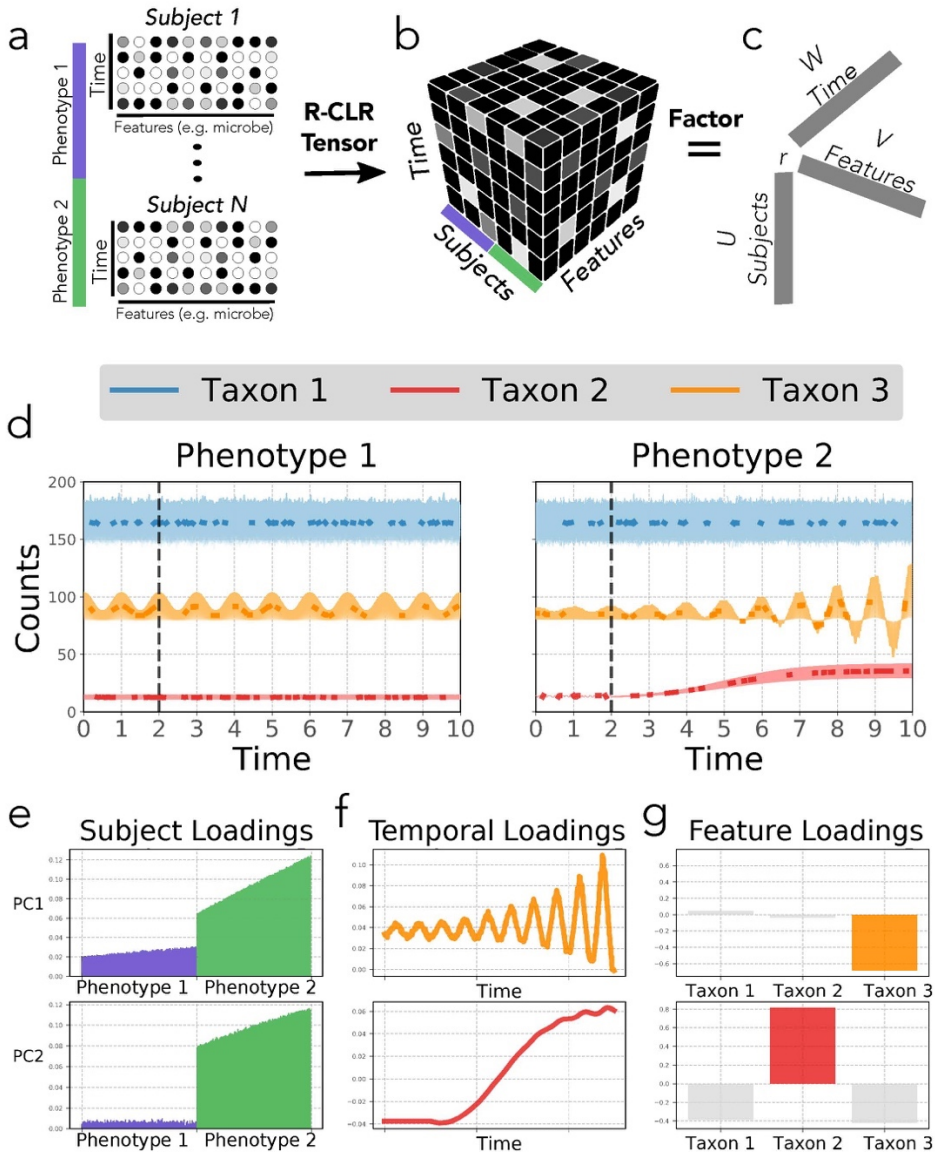

Figure 3.1. Overview of the CTF algorithm. (a) CTF utilizes feature abundance matrices for subjects over time. For each subject with a phenotype of interest, the data is represented as relative abundances of features (abundance gradient represented in grayscale) over time. (b) The matrices are concatenated, robust-centered log-ratio transformed (R-CLR) and structured into a tensor format with modes corresponding to subjects, features and time. (c) The resulting tensor is then factored based only on observed data into loading vectors for each dimension (i.e. subject, timepoint, and feature). (d) Simulated count data is plotted on the y-axis for three taxa with the mean counts in bold and missing values absent from the bold line. Standard deviation of distributions are shaded behind. Two phenotypes are compared; a control unchanging in time (left) and a dynamic phenotype with a perturbation at time point 2 (right). Taxon 1 (blue) is highly abundant and noisy, taxon 2 (red) is lowly abundant but growing exponentially in phenotype 2, and taxon 3 (orange) is oscillatory with increasing amplitude in phenotype 2. The first two principal component axes (i.e. loadings) from CTF (PC1 (top) and PC2 (bottom)) are plotted on the y-axis with the corresponding sample (e), time (f), and feature loadings (g). In PC1, phenotype 2 is linked to the unstable oscillatory waveform of highly loaded taxon 3 (orange, top). Similarly, in PC2, phenotype 2 is linked to the sigmoidal waveform of highly loaded taxon 2 (red, bottom).

To demonstrate the utility of CTF, we applied it to a simulated longitudinal dataset with two phenotypic groups. Simulations were generated based on distributions in real longitudinal 16S data from Halfvarson et al.[8] while varying the sequencing depth and temporal sampling densities as described by Äijö et al.[3] This dataset was chosen because there were strong differences in microbial composition and beta diversity between subjects with and without Crohn's disease[8].We compared CTF to state-of-the-art beta-diversity metrics through PCoA including Jaccard[9], Bray Curtis[10], Aitchison[11], unweighted UniFrac[12], and weighted UniFrac[13]. K-nearest neighbor (KNN) classification by disease state in each of our simulations revealed that CTF exhibited higher accuracy than existing methods regardless of sequencing depth or the number of longitudinally collected samples (Fig. 2.2, Table AA.1.S1). CTF also exhibited higher discriminatory power by PERMANOVA F-statistic across all levels of sequencing depth and at higher sampling densities (≥ 3 time points; Fig. 2.2).

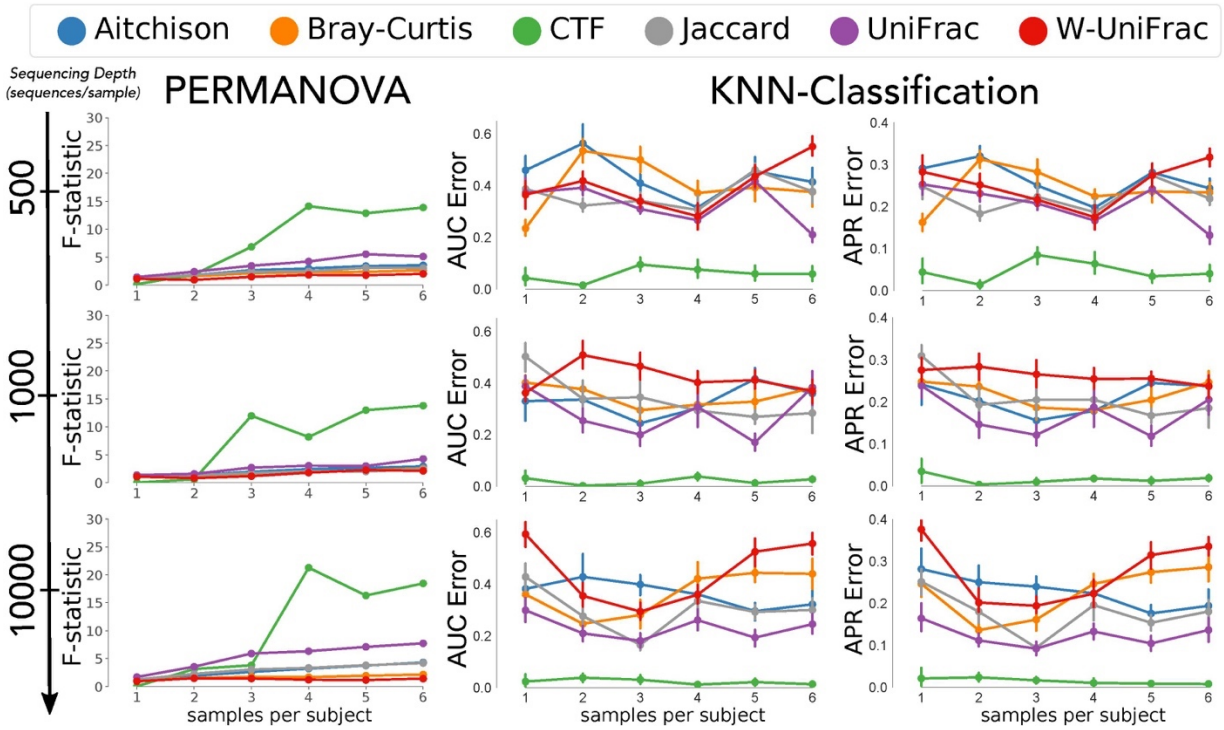

Figure 3.2. CTF outperforms popular distance metrics in longitudinal in silico data-driven simulations. Increasing sequencing depth (500 - 10,000; rows) over differing temporal sampling densities (x-axis) evaluated for PERMANOVA F-statistic as a measure of discriminatory power (left column), in addition to KNN-classification cross-validation by AUC (n=100; middle column), and APR (n=100; right column). Compared among CTF (green) and popular distance metrics Aitchison (blue), Bray-Curtis (orange), Jaccard (grey), unweighted (purple), and weighted (red) UniFrac. Error bars represent standard error of the mean.

We next applied CTF to two published datasets that tracked infant gut development over time. The datasets abbreviated as ECAM (n-subjects=43)[14] and DIABIMMUNE (n-subjects=39)[15] followed infants for the first 2 and 3 years of life, respectively. Both datasets observed that birth mode (i.e., vaginal delivery or caesarean section) differentiated microbial community composition. Similar to our results from the simulated data, CTF is 10-fold better at discriminating vaginally from caesarean born infants compared to state-of-the-art beta-diversity metrics (Fig. 3.S2a&b, Fig. 3.S3a&b, Table AB.2.S2).

We sought to examine CTF's ability to reproducibly identify differentially abundant microbes in an unsupervised manner. To this end, we compared the feature rankings between the ECAM and DIABIMMUNE datasets along the first axis of variation and found they were significantly correlated (Pearson correlation; R^2=0.974, $P<10^{-10}$) (Fig. 3.S2). While these 2 datasets had <50% overlap at the sOTU level (Fig. 3.S2d), highly ranked sOTUs grouped at the genus level were similar across both datasets (Fig. 3.S2e). We note that although these datasets were collected and processed using distinct protocols and by different labs, CTF identified the same taxa driving gut microbiome differentiation by birth mode, suggesting a robust microbial structure across infants.

We constructed a birth-mode log-ratio of vaginally to cesarean features using the sOTUs most associated with vaginal and cesarean birth in each dataset (Fig. 3.S4; Methods). Samples were significantly separated by birth-mode in both datasets along time (Fig. 3.S5, Table AB.2.S3). We note that these birth-mode microbial signatures are not confounded by established differentiators such as antibiotics usage or feeding mode (Fig. 3.S5). Nonetheless, we cannot rule out the possibility of unmeasured confounders. We next combined those sOTUs common to both ECAM and DIABIMMUNE birth-mode ratios to create a ‘microbial birth-mode signature’.

To examine the robustness of this microbial birth-mode signature, we tested its discriminatory ability in data from the American Gut Project (AGP, n=8,099), a large cross-sectional dataset[16]. We found that this signature significantly differentiated participants under the

age of four by birth mode (t-test; p-value=0.042; Fig. 3.S6), consistent with our previous findings. The robustness of this microbial signature, across multiple datasets, highlights the ability of CTF to identify differentially abundant features reproducibly associated with a phenotype.

In both the ECAM and DIABIMMUNE datasets we observed that throughout infant development samples from vaginally versus cesarean born infants became less distinct (Fig. 3.S2a&b). Similarly, the microbial birth-mode signature no longer differentiated participants by birth mode in samples from participants above the age of four in the AGP dataset (Fig. 3.S6).

CTF is the only unsupervised method that allows full utilization of repeated measures while accounting for the inherent properties of microbiome sequencing datasets, namely high-dimensionality, sparsity, and compositionality. In both simulated and real datasets, CTF outperformed the current state-of-the-art beta-diversity metrics. Although CTF can reveal robust microbial signatures, several considerations are necessary when applying this tool. First, CTF relies on an assumption that the underlying data is of low rank. This assumption can be violated, making CTF inappropriate to use, such as when the data are driven by a gradient rather than discrete groupings (for example the 88 Soils dataset[17]). Our implementation of CTF estimates the underlying rank and informs the user if the data does not meet this requirement[18]. Second, CTF, like other beta-diversity metrics, does not directly account for the presence of confounders that may affect downstream clustering, requiring additional validations similar to the one presented in Fig. 3.S5. Finally, although CTF leverages repeated measures to account for inter-individual variation and is optimal in the case of a synchronization event (e.g., treatment, diet), it is permutation invariant and does not take into account the ordering of longitudinal data.

In addition to longitudinal datasets as benchmarked here, CTF could also be used for spatially repeated measurements. This includes studies where samples are collected contemporaneously, for example where multiple body sites are measured (e.g., skin and saliva) or sites with different phenotypes (e.g., lesioned versus adjacent non-lesioned skin). Furthermore, CTF could be used to analyze other types of datasets that contain a high amount of inter-individual

variation, such as metabolomics or proteomics. In summary, CTF leverages the power of repeated measures study design to elucidate biological changes while accounting for inter-individual variability. We propose the use of this tool both for the re-analysis of existing datasets and for future microbial community research.

3.3 Methods

Preprocessing with robust-clr

Prior to running tensor factorization, we use the robust centered log-ratio transformation (robust-clr) to center the data around zero and approximate a normal distribution[5]

$$rclr(x) = \left[\log\frac{x_1}{g_r(x)}, \ldots, \log\frac{x_D}{g_r(x)}\right] \quad (1)$$

$$g_r(x) = \left(\prod_{i \in \Omega_x} x_i\right)^{1/|\Omega_x|} \quad (2)$$

where x_i is the abundance of microbe i, Ω_x is the set of observed microbes in sample x and $g_r(x)$ is the geometric mean only defined on microbes with abundance > 0. Unlike the traditional clr transformation, the robust-clr handles the high level of sparsity found in microbial datasets without requiring imputation. Furthermore, this transformation has shift invariant properties that allow the restructuring of the matrix into tensor form.

Tensor factorization via alternating least squares minimization

Here we follow the tensor notations of Lim[24] and Anandkumar et al.[25] , for a full notation see the supplemental methods. To perform tensor factorization on sparse data we followed a procedure introduced by Jain and Oh[26]. Due to the high level of sparsity in microbiome datasets

we would like to find the minimum rank representation of *T* that best explains only *observed* values defined as Ω. We use the projection $P_\Omega(T)_{ijt}$

$$P_\Omega(T)_{ijt} = f(x) = \begin{cases} T_{ij}, & if(i,j,t) \in \Omega \\ 0, & Otherwise \end{cases} \quad (3)$$

The objective function being optimized through alternating least squares minimization (ALS) is given by

$$\min_{\{\sigma_i, a_i, b_i, c_i\}_{i \in [r]}} \| P_\Omega(T) - P_\Omega \left(\sum_{i=1}^{r} \sigma_i \, (a_i \otimes b_i \otimes c_i) \right) \|_F^2 \quad (4)$$

where a, b, and c are unstructured, orthogonal, and have a Euclidean norm of 1. The low rank representations a, b, and c correspond to loadings for the first, second and third tensor modes respectively. It is important to note that this factorization is permutation invariant, meaning the order of time or space is not a factor in the subsequent loadings of c.

Factorization trajectories

Here, we focus on the interpretation of tensor factorization for biological data. We are primarily concerned with 3rd-order tensors from studies following multiple subjects over several timepoints. In this tensor the first mode is the subjects or environments sampled. The second mode is biological features such as microbes, metabolites, or genes. The third mode is timepoints where subjects/environments were sampled repeatedly. Of utmost interest is the relation between subject or features and the third mode of time. To obtain easily interpretable loadings we introduce trajectories given by

$$Subject\ Trajectory = a \odot c = [a_1 \otimes c_1, \ldots, a_r \otimes c_r] \in \mathbb{R}^{d^2 \times r}$$

$$Feature\ Trajectory = b \odot c = [b_1 \otimes c_1, \ldots, b_r \otimes c_r] \in \mathbb{R}^{d^2 \times r}$$

where ⊙ represents the Khatri-Rao product. These trajectories are of the shape (subjects × time, rank) or (features × time, rank) where each rank-1 column has an accompanying singular value σ_r.

Log-ratio feature selection

In order to explore how feature rankings in b or b ⊙ c partitioned subjects we used log-ratios between highly (positive) and lowly (negative) ranked features along the first axis of variation. To avoid the use of pseudo-counts we explore the sum of the minimum number of highly and lowly ranked features summed across all samples, such that no log-ratio contains a zero value. For ECAM 1400 and DIABIMMUNE 750 total features were used and split between numerator and denominator evenly such that no samples were dropped due to zero values (Fig 3.S5). We then used a Linear Mixed Effects (LME) model via statsmodels (v. 0.11.0) to test the if the log-ratio changed over time and in response to birth mode for ECAM and DIABIMMUNE separately. The LME model produced residual R^2 values of 0.976 and 0.986 for DIABIMMUNE and ECAM respectively. The resulting p-values from the LME were significant ($P < .05$) by birth mode, time in days, and the interaction of the two (Table S4). To produce the microbial birth-mode signature, we used only sequences shared among ECAM, DIABIMMUNE, and the American Gut Project (1,064 features total). We used the ranking structure inferred from ECAM and DIABIMMUNE to evenly divide these shared features into vaginal or cesarean-associated taxa (532 each in the numerator and denominator, respectively). A t-test via SciPy (v. 1.4.1) was used on the microbial birth-mode signature (i.e., log-ratio) to test for significance between birth modes stratified by age or time point for both data sets, respectively.

Data driven simulation benchmarks

Data driven simulations were designed to benchmark different characteristics of data without making assumptions about microbial dynamics. The IBD dataset was chosen due to its high temporal resolution and two-group (low-rank) comparison. Simulations were generated using a procedure from Äijö et al.[3] modified to use a Poisson-lognormal distribution (PLN)[27] as opposed to a Poisson-Multinomial distribution. This simulation was repeated for different levels of dispersion, subsampling (i.e. sparsity), sampling density (i.e. number of timepoints) and percentage of randomly missing samples.

Case Study Sequence Processing

Raw sequences were quality controlled, trimmed at 100 nucleotides, and clustered as amplicon sequence variants (sOTUs) using QIIME 2 release 2019.7 and Deblur (v. 1.1.0) [28,29]. The phylogenetic tree was created using SEPP sequence insertion with the Greengenes tree 13.8 release as the reference tree[30,31]. Taxonomy assignments were made using a Naive Bayes classifier as implemented in QIIME2 (v. 2019.7). All data preprocessing was conducted on Qiita[32] where all the data used here is freely available. All other visualizations were plotted through Matplotlib.

Quantitative comparison of metrics

All comparisons were made between Jaccard, Bray-Curtis, Weighted UniFrac, Unweighted UniFrac, Aitchison, and CTF distances. All distance metrics were calculated through QIIME2 (v. 2019.7). PERMANOVA on distances between subject groupings (i.e. vaginal vs. caesarean birth mode) was performed through scikit-bio (v. 0.5.5). Dimensionality reduction on distances was performed through PCoA via scikit-bio (v. 0.5.5). The first three components of each dimensionality reduction were evaluated through k-nearest neighbors (KNN) classification via scikit-learn (v. 0.21.2). To assess the classification accuracy, KNN classification was

performed with 100-fold 40:60 cross-validation evaluating AUC and APR prediction accuracy at each fold-iteration via scikit-learn (v. 0.21.2).

Basis for simulations

Halfvarson et al. The IBD cohort used as the introduction example is a previously published dataset by Halfvarson et al. (Qiita ID 1629)[8]. The dataset consists, after filtering as described below, of 23 subjects (14 Crohn's disease (CD), 9 Control) each with one to eight samples for a total of 134 samples. Samples were filtered from the original data for only CD and Control. For the data-driven simulations, only the first 6 time points were retained to reduce the missing time points across subjects. The resulting data was then run through the data-driven simulation protocol described above for a sequencing depth of 500, 1000, and 10000 mean reads per sample. CTF was performed on each simulated data set through gemelli (v. 0.0.5) with a set rank of 2.

Case study: ECAM

The ECAM dataset published by Bokulich et al. followed 43 infants (19 c-section, 24 vaginally delivered) from birth over the first year of life with monthly fecal sampling (Qiita ID 10249)[14]. Three months (month 6, 15, and 19) were removed for a lack of subjects represented and CTF analysis was run with a set rank of 2. Features with < 5 total counts across samples were filtered. Samples with < 2000 reads per sample were removed.

Case study: DIABIMMUNE

The DIABIMMUNE dataset, published by Yassour et al., followed 39 infants (4 c-section, 35 vaginally delivered) from the 2nd month after birth over the first three years of life with monthly fecal sampling (Qiita ID 11884)[15]. Two months (month 28 and 30) were removed for a lack of

subjects represented and CTF analysis was run with a set rank of 4. Features with < 5 total counts across samples were filtered. Samples with < 2000 reads per sample were removed.

Case study: American Gut

The American Gut Project data and metadata tables were acquired from ftp://ftp.microbio.me/AmericanGut/manuscript-package/ which was provided in McDonald et al.[16]. From this data the combined ECAM and DIABIMMUNE log-ratio feature set was used on the subset of the data with age and birth-mode labels provided (8,436 total samples).

Data availability

The sequences and biom tables for the IBD, ECAM, DIABIMMUNE, and AGP datasets can be found on Qiita (http://qiita.microbio.me) under study IDs 1629, 10249, 11884, and 10317 and at EBI or BioProject under ERP020401, ERP016173, PRJNA290381, and ERP012803.

Code availability

The CTF codebase named Gemelli is a fully unit tested open-source python package, and is installable through pip or conda. Additionally, CTF is wrapped in a QIIME2 plugin: https://github.com/biocore/gemelli; All the code and analyses are available in the 'Code Ocean' capsule: https://dx.doi.org/10.24433/CO.5938114.v1.

3.4. Acknowledgments

C.M., L.S., and R.K. conceived, initiated, and coordinated the project. C.M., L.S., D.M. and Y.V-B coordinated, compiled and performed analysis. C.M., C.M., and G.A. wrote the code for CTF. C.M., L.S., and C.M. wrote the manuscript. J.T.M, A.D.S., and M.G.D-B. provided

essential discussion and advice. E.H. and R.K. supervised the project. All authors discussed the experiments and results, read, and approved the manuscript.

This work was partially supported by the C&D Research Fund (M.G.D.B.), the EMCH fund for human microbiome studies, the Norwegian Institute of Public Health (2019-0350), the Emerald Foundation, the NIH Pioneer award (1DP1AT010885), the National Institute of Justice (2016-DN-BX-4194), the San Diego Digestive Diseases Research Center (NIDDK 1P30DK120515) and Janssen Pharmaceuticals (20175015). CM was funded by the NIDCR (1F31DE028478-01). E.H. and L.S were partially supported by the National Science Foundation (Grant No. 1705197) and by NIH 1R56MD013312. E.H. was also partially supported by NIH/NHGRI HG010505-02, NIH 1R01MH115979, NIH 5R25GM112625, and NIH 5UL1TR001881.

3.5. References

1. Gibson, T. E. & Gerber, G. K. Robust and Scalable Models of Microbiome Dynamics. *arXiv [stat.ML]* (2018).

2. Shenhav, L. *et al.* Modeling the temporal dynamics of the gut microbial community in adults and infants. *PLoS Comput. Biol.* **15**, e1006960 (2019).

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

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

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

6. 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).

7. Morton, J. T. *et al.* Establishing microbial composition measurement standards with reference frames. *Nat. Commun.* **10**, 2719 (2019).

8. Halfvarson, J. *et al.* Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat Microbiol* **2**, 17004 (2017).

9. Jaccard, P. The distribution of the flora in the alpine zone. 1. *New Phytol.* **11**, 37–50 (1912).

10. Bray, J. R. & Curtis, J. T. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecol. Monogr.* **27**, 325–349 (1957).

11. Aitchison, J. Principal component analysis of compositional data. *Biometrika* **70**, 57–65 (1983).

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

13. McDonald, D. *et al.* Striped UniFrac: enabling microbiome analysis at unprecedented scale. *Nat. Methods* **15**, 847–848 (2018).

14. Bokulich, N. A. *et al.* Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Science Translational Medicine* vol. 8 343ra82–343ra82 (2016).

15. Yassour, M. *et al.* Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Sci. Transl. Med.* **8**, 343ra81 (2016).

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

17. Lauber, C. L., Hamady, M., Knight, R. & Fierer, N. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl. Environ. Microbiol.* **75**, 5111–5120 (2009).

18. Keshavan, R. H., Montanari, A. & Oh, S. Low-rank matrix completion with noisy observations: A quantitative comparison. in *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)* 1216–1222 (2009).

19. Alhagamhmad, M. H., Day, A. S., Lemberg, D. A. & Leach, S. T. An overview of the bacterial contribution to Crohn disease pathogenesis. *J. Med. Microbiol.* **65**, 1049–1059 (2016).

20. Vázquez-Baeza, Y. *et al.* Guiding longitudinal sampling in IBD cohorts. *Gut* vol. 67 1743–1745 (2018).

21. Cekin, A. H. A microbial signature for Crohn's disease. *The Turkish Journal of Gastroenterology* vol. 28 237–238 (2017).

24. Lek-Heng Lim. Singular values and eigenvalues of tensors: a variational approach. in *1st IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, 2005.* 129–132 (2005).

25. Anandkumar, A., Ge, R. & Janzamin, M. Guaranteed Non-Orthogonal Tensor Decomposition via Alternating Rank-1 Updates. *arXiv [cs.LG]* (2014).

26. Jain, P. & Oh, S. Provable Tensor Factorization with Missing Data. in *Advances in Neural Information Processing Systems 27* (eds. Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D. & Weinberger, K. Q.) 1431–1439 (Curran Associates, Inc., 2014).

27. Aitchison, J. & Ho, C. H. The multivariate Poisson-log normal distribution. *Biometrika* **76**, 643–653 (1989).

28. Amir, A. *et al.* Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems* **2**, (2017).

29. Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857 (2019).

30. Janssen, S. *et al.* Phylogenetic Placement of Exact Amplicon Sequences Improves Associations with Clinical Information. *mSystems* **3**, (2018).

31. McDonald, D. *et al.* An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* **6**, 610–618 (2012).

32. Gonzalez, A. *et al.* Qiita: rapid, web-enabled microbiome meta-analysis. *Nat. Methods* **551**, 457 (2018).

Chapter 4. Naturalization of the microbiota developmental trajectory of Cesarean-born neonates after vaginal seeding

Summary

Background

Early microbiota perturbations are associated with disorders that involve immunological underpinnings. Cesarean section (CS)-born babies show altered microbiota development in relation to babies born vaginally. Here we present the first statistically powered longitudinal study to determine the effect of restoring exposure to maternal vaginal fluids after CS birth.

Methods

Using *16S rRNA* gene sequencing, we followed the microbial trajectories of multiple body sites in 177 babies over the first year of life; 98 were born vaginally and 79 were born by CS, of which 30 were swabbed with a maternal vaginal gauze right after birth.

Findings

Compositional tensor factorization analysis confirmed that microbiota trajectories of exposed CS-born babies aligned closer to that of vaginally born babies. Interestingly, the majority of amplicon sequence variants from maternal vaginal microbiomes on the day of birth were shared with other maternal sites, in contrast to non-pregnant women from the HMP study.

Conclusions

The results of this observational study prompt the urgent need of randomized clinical trials to test whether microbial restoration reduces the increased disease risk associated with CS birth, and the underlying mechanisms. Also, it provides evidence for the pluripotential nature of maternal vaginal fluids to provide pioneer bacterial colonizers for the body newborn body sites.

This is the first study showing long term naturalization of the microbiota of CS-born infants by restoring microbial exposure at birth.

Funding

C&D, Emch Fund, CIFAR, Chilean CONICYT and SOCHIPE, Norwegian Institute of Public Health Emerald Foundation, NIH, National Institute of Justice, Janssen.

4.1. Introduction

Over the past few decades, we have learned a great deal about the multitude of ways that microbiota affect the development of their hosts. Studies on model organisms show that fetal development can be modulated by microbial products from the pregnant mother's microbiota, and early colonization is critical for immune system development (Gensollen et al., 2016); (Al Nabhani and Eberl, 2020).

Natural transmission and colonization of maternal microbes is impaired by delivery via cesarean section (CS) (Bokulich et al., 2016; Dominguez-Bello et al., 2010; Shao et al., 2019; Stewart et al., 2018; Yassour et al., 2016). Furthermore, CS birth is associated with reduced levels of various cytokines and their receptors (Malamitsi-Puchner et al., 2005), increased risk of opportunistic neonatal infections (Shao et al., 2019), immune diseases (Stokholm et al., 2018) (Andersen et al., 2020) and obesity (Ardic et al., 2020; Blustein et al., 2013). These associations have been shown to be causal in mouse models for conditions such as obesity (Cox et al., 2014; Martinez et al., 2017), and immune disorders (Livanos et al., 2016; Olszak et al., 2012). Neuroendocrine abnormalities including cognitive and behavioral disorders have also been associated with early microbiome perturbations (Braniste et al., 2014; Moya-Perez et al., 2017). Understanding the contribution of microbionts to healthy development remains a crucial challenge to address the current epidemic of immune and metabolic diseases in urban societies.

Although used without medical indication in many countries, CS delivery is often medically necessary and a life-saving procedure, and thus, restoration may be one solution to help reduce the risk of associated disorders related to the microbiome. Two proof of concept studies have demonstrated the principle of engraftment of maternal bacteria on CS born babies after deliberate microbial exposure: the first one by Dominguez-Bello et al. using maternal vaginal gauze as a source (Dominguez-Bello et al., 2016), and the second recent pilot study by Korpela et al. using maternal feces (Korpela et al., 2020). Here we present the first large observational study of the long-term effect of maternal vaginal seeding after CS delivery to restore microbial development during the first year of life.

4.2. Results

Vaginal seeding of CS born infants

A total of 177 infants born to 174 mothers were studied (Figure 4.S1a), of which 101 were born in USA, 50 in Chile, 6 in Bolivia, and 20 in Spain (Table 4.1). 98 infants were born vaginally and 79 were delivered by CS, of which 30, who complied with inclusion criteria (see Star Methods), were swabbed with a maternal vaginal gauze at birth (vaginal seeding (Dominguez-Bello et al., 2016)). The microbiota development was followed during the first year of life. A total of 8,104 samples from stool, mouth, and skin of infants and their mothers were obtained, with additional nasal and vaginal samples from mothers (Figure 4.S1a-c). None of the seeded infants had any complications, and all children developed normally during the 12 months of the study.

Vaginal seeding partly normalizes microbiome trajectories in C-section-delivered infants

Across the different body sites, the samples yielded good overall sequencing depth (mean depth of 63,035 paired-end reads per sample), with a low probability of sample contamination as indicated by a survey of negative controls (Figure 4.S1d). Analysis of the vaginal gauzes stored in the vagina for an hour before the CS procedure with which the neonates were swabbed showed

that ~76% of bacterial amplicon sequence variants (ASVs, see Star Methods) contained in maternal vaginal swabs were also present in the gauzes (Figure 4.S2a).

Some studies have reported decreased alpha diversity in CS born versus vaginally-born infants (Jakobsson et al., 2014). Yet, others have reported no differences by birth mode (Bokulich et al., 2016; Yassour et al., 2016). Using a linear mixed-effects model, we found inconsistent results depending on the body site and alpha-diversity metric (Supplementary Methods 4.S1). One possibility for this inconsistency is that the dynamic nature of the developing microbiome can be highly non-linear, and often times, data collected longitudinally vary in frequency and timing across individuals. To account for these potential irregularities we applied a novel method called Bayesian Sparse Functional Principal Components Analysis (SFPCA; Jiang et al. (2020) to estimate individual trajectories (see STAR Methods). Using SFPCA, we found that alpha diversity trajectories did not differ among birth modes when measured as Shannon diversity (Figure 4.S3), or when accounting for phylogenetic relatedness (SFPCA on Faith's PD, data not shown).

However, significant birth group differences were found in beta diversity when using an unsupervised dimensionality reduction method called Compositional Tensor Factorization (CTF) (Martino et al., 2020). CTF accounts for the repeated measurements allowing comparisons of beta-diversity over time ('trajectory') while accounting for the sparse, compositional nature of next-generation microbiome sequencing data (Gloor et al., 2017; Morton et al., 2019). The trajectory of gut microbiota development in CS born infants diverged from that of vaginally born infants through the entire first year of life (Fig. 4.1). These results are consistent with findings from previous studies that used more traditional analysis approaches (Bokulich et al., 2016; Yassour et al., 2016). CTF also detected measurable differences in the microbial development of the mouth (Fig. 4.2) and skin (Fig. 4.3), underscoring the importance of birth mode in affecting multiple microbial niches during human development.

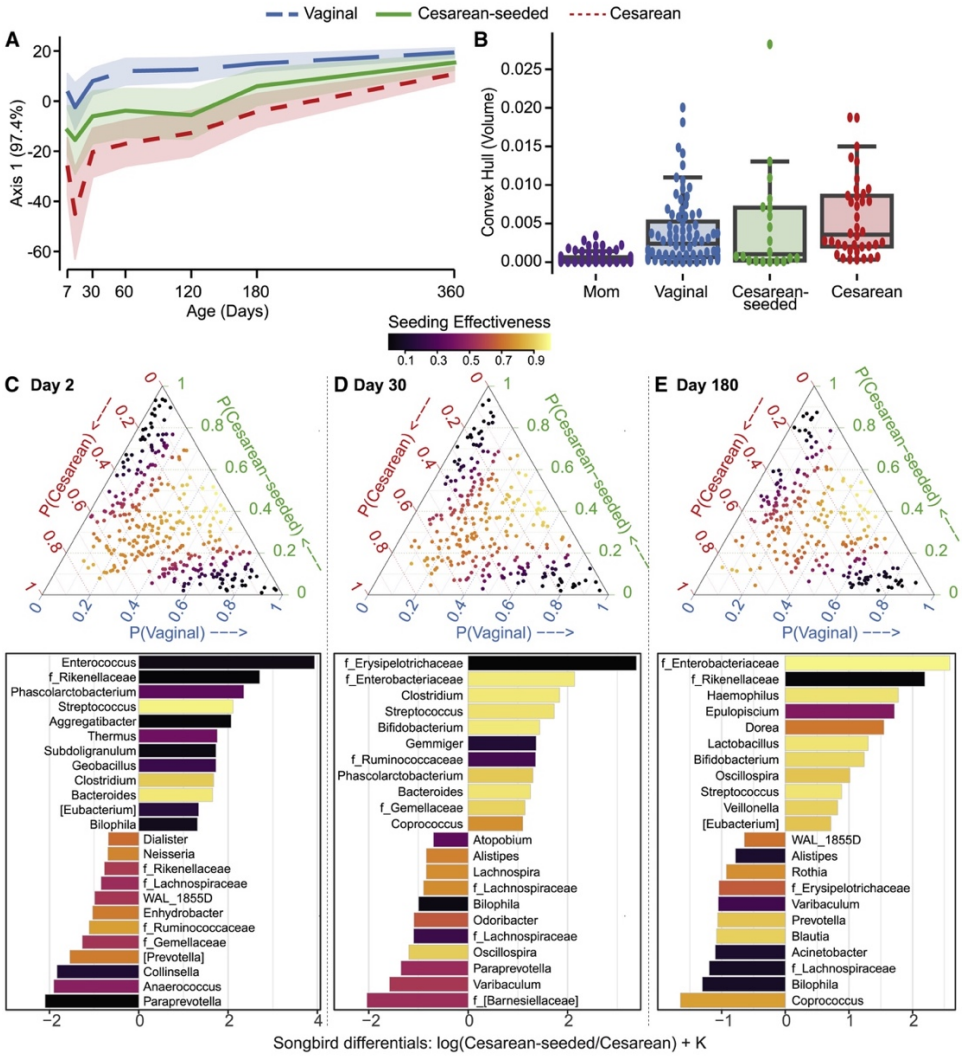

Figure 4.1. Fecal microbiota development during the first year of life in babies discordant to birth mode/exposure. (a) Compositional Tensor Factorization (CTF) first principal component (Y-axis) of infant samples over age in days (X-axis). (b) Convex hull volume (Y-axis) on the first three Principal Coordinates (unweighted UniFrac distances) in mothers (purple) and infants by birth mode or exposure (X-axis). Infants show highest volumes in Cesarean born and lowest in Vaginally born, with Cesarean-seeded babies showing intermediate volumes; all pairwise comparisons are significant using Mann-Whitney test with Bonferroni corrections at 0.05 level (Table AC.2.S3). (c-e) Songbird differentials shown for day 2, 30, and 180 after birth; ternary plots of the inverse additive log-ratio transform (inverse-ALR) of Songbird differentials give the estimated probability of a microbe being observed in Cesarean (left-axes; red), Vaginal (bottom-axes; blue), or Cesarean-seeded (right-axes; green). The color of the dots depicts the seeding effectiveness, with yellow color indicating effectively seeded/suppressed and black indicating not effectively seeded. Below each triangle, bar plots of top and bottom 20% Songbird differentials summarized at genus-level taxa between Cesarean-seeded and Cesarean born babies; a positive value indicates higher association with the Cesarean-seeded group, a negative value indicates higher with Cesarean. Bars are colored by the ASVs' seeding effectiveness. The majority of taxa discordant overrepresented in the Cesarean-seeded group over the Cesarean group are yellow-orange, indicating ASVs effectively seeded in the Cesarean seeded group, and these are observed at all ages. *See also Figure 4.S4, Supplementary Methods 4.S2-S7.*

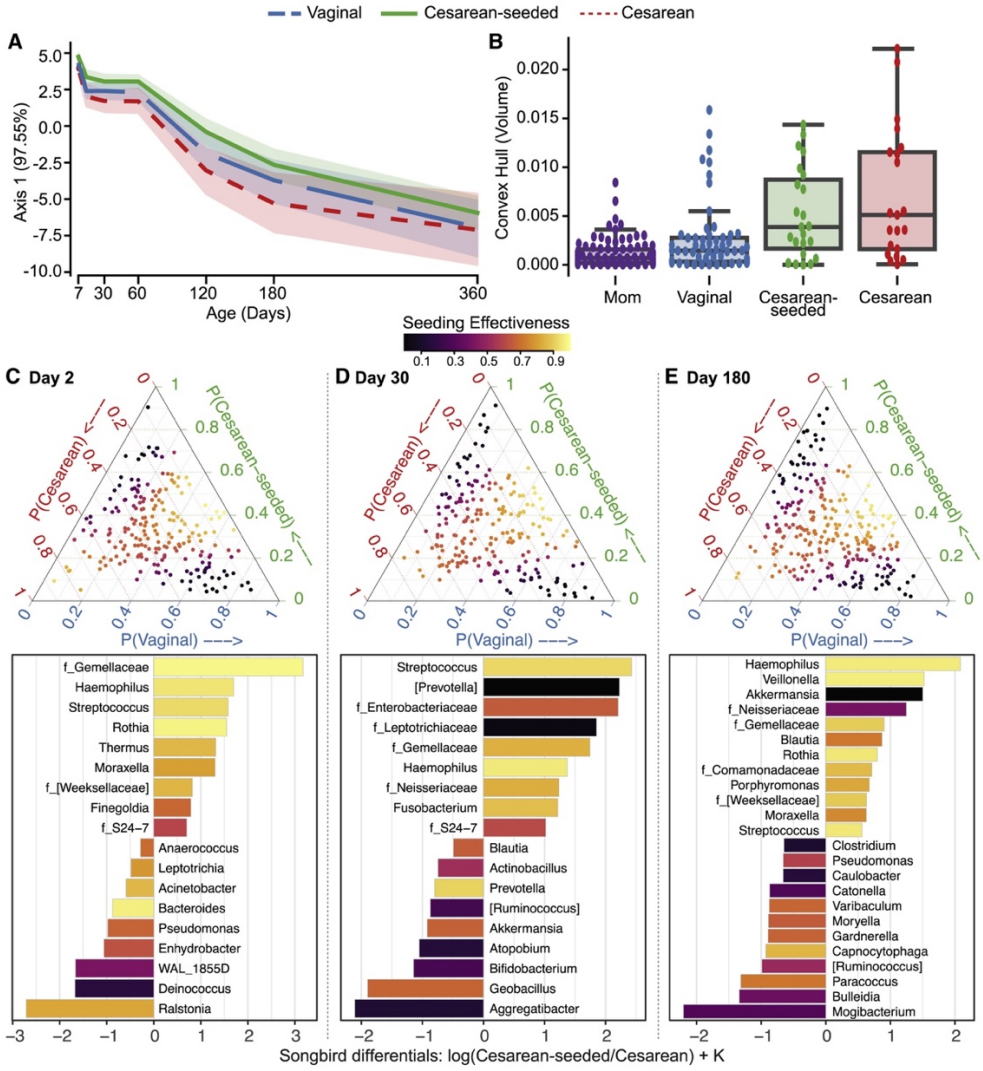

Figure 4.2. Oral microbiota development during the first year of life in babies discordant to birth mode/exposure. (a) Compositional Tensor Factorization (CTF) first principal component (Y-axis) of infant samples over age in days (X-axis). (b) Convex hull volume (Y-axis) on the first three Principal Coordinates (unweighted UniFrac distances) in mothers (purple) and infants by birth mode or exposure (X-axis). Infants show highest volumes in Cesarean born and lowest in Vaginally born, with Cesarean-seeded babies showing intermediate volumes; all pairwise comparisons are significant using Mann-Whitney test with Bonferroni corrections at 0.05 level (Table AC.2.S3). (c-e) Songbird differentials shown for day 2, 30, and 180 after birth; ternary plots of the inverse additive log-ratio transform (inverse-ALR) of Songbird differentials give the estimated probability of a microbe being observed in Cesarean (left-axes; red), Vaginal (bottom-axes; blue), or Cesarean-seeded (right-axes; green). The color of the dots depicts the seeding effectiveness, with yellow color indicating effectively seeded/suppressed and black indicating not effectively seeded. Below each triangle, bar plots of top and bottom 20% Songbird differentials summarized at genus-level taxa between Cesarean-seeded and Cesarean born babies; a positive value indicates higher association with the Cesarean-seeded group, a negative value indicates higher with Cesarean. Bars are colored by the ASVs' seeding effectiveness. The majority of taxa discordant overrepresented in the Cesarean-seeded group over the Cesarean group are yellow–orange, indicating ASVs effectively seeded in the Cesarean seeded group, and these are observed at all ages. *See also Figure 4.S4, Supplementary Methods 4.S2-S7.*

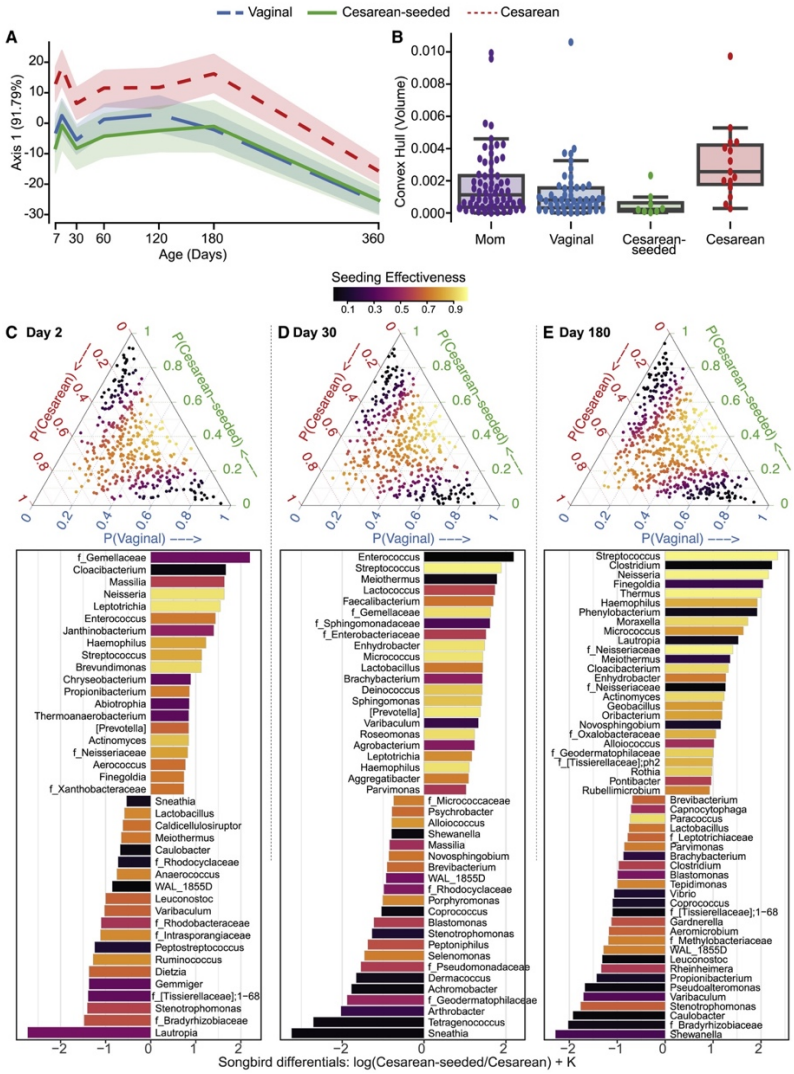

Figure 4.3. Skin microbiota development during the first year of life in babies discordant to birth mode/exposure. (a) Compositional Tensor Factorization (CTF) first principal component (Y-axis) of infant samples over age in days (X-axis). (b) Convex hull volume (Y-axis) on the first three Principal Coordinates (unweighted UniFrac distances) in mothers (purple) and infants by birth mode or exposure (X-axis). Infants show highest volumes in Cesarean born and lowest in Vaginally born, with Cesarean-seeded babies showing intermediate volumes; all but one pairwise comparison are significant using Mann-Whitney test with Bonferroni corrections at 0.05 level (Table .4.S3). (c-e) Songbird differentials shown for day 2, 30, and 180 after birth; ternary plots of the inverse additive log-ratio transform (inverse-ALR) of Songbird differentials give the estimated probability of a microbe being observed in Cesarean (left-axes; red), Vaginal (bottom-axes; blue), or Cesarean-seeded (right-axes; green). The color of the dots depicts the seeding effectiveness, with yellow color indicating effectively seeded/suppressed and black indicating not effectively seeded. Below each triangle, bar plots of top and bottom 20% Songbird differentials summarized at genus-level taxa between Cesarean-seeded and Cesarean born babies; a positive value indicates higher association with the Cesarean-seeded group, a negative value indicates higher with Cesarean. Bars are colored by the ASVs' seeding effectiveness. The majority of taxa discordant overrepresented in the Cesarean-seeded group over the Cesarean group are yellow–orange, indicating ASVs effectively seeded in the Cesarean seeded group, and these are observed at all ages. *See also Figure 4.S4, Supplementary Methods S2-S7.*

Seeding CS-born infants led to a developmental trajectory that more closely resembled that of vaginally-born infants most prominently in feces (Fig. 4.1, 4.S4a,b) and skin (Fig. 4.3, 4.S4a,b); this trend held when considering only the 101 babies born in the US (Fig. 4.S4c), while other countries lacked sufficient sample size for individual analysis. Furthermore, a stepwise redundancy analysis based on the first three principal components of CTF ordination (Falony et al., 2016) confirmed that birth mode significantly contributed to differences in microbial community structures in the gut and on skin, but not in the mouth, with effect sizes of 0.17 (R^2) in fecal samples and 0.09 in skin samples (Supplementary Methods 4.S2). Analyzing these data using more conventional tools for comparing beta diversity that do not account for interindividual variation in repeated measure studies (Supplementary Methods 4.S3-S4), evaluated through PERMANOVA (on unweighted Unifrac distance) or RDA (on PCoA PCs), expectedly reveals individuals as the primary driver of variation (PERMANOVA F-statistic = 5.45, P-value <= 0.001; RDA adjusted R^2 = 0.113, Supplementary Methods 4.S5). High interindividual variation obscured the ability to detect differences due to more muted factors such as birth mode using these methods. Together, these findings reveal that birth mode affects the development of microbial communities, and that this effect may be undetected upon analyses with traditional bioinformatic tools.

Differences in microbial composition stability have been used to differentiate phenotypes in longitudinal studies (Halfvarson et al., 2017; Zaneveld et al., 2017). Accordingly, we next compared variability across samples over time within a given individual. To leverage the dense sampling design, we calculated the volume of the shape determined by an individual's samples in the first 3 principal coordinates of unweighted UniFrac space using a convex hull analysis (see STAR Methods). As expected, the average variability of the microbiome over an infant's first year of life was much greater than the variability in the mother's microbiome (Fig. 4.1b, 4.2b, 4.3b). CS born infants had significantly greater microbial variability than vaginally born infants, and the variability of seeded infants was intermediate (Fig. 4.1b, 4.2b, 4.3b, Supplementary Methods

4.S6). This finding held true for fecal, oral and skin samples, suggesting that vaginal seeding may also help stabilize microbiome development. This trend can also be observed using data within the first 6 months (Supplementary Methods 4.S7). Possible confounders such as antibiotic consumption (which was similar between baby groups; Table 4.1) were discarded; in the CS born and restored babies, stepwise RDA did not recognize antibiotic consumption as a factor altering seeding efficiency. In summary, these results indicate that vaginal seeding resulted in partial recovery of the microbiome in CS-delivered infants.

Bacterial taxa associated with effective seeding

To determine whether specific microbial taxonomies were being seeded well or the overall seeding across all microbes was partial, we first identified which taxa were most associated with a vaginal birth compared to a CS birth using Songbird (Morton et al., 2019), and then calculated a seeding-effectiveness score for those taxa (see STAR Methods; zero indicates poor seeding and one indicates effective seeding or effectively suppressed). Effectively seeded microbes are those shared by vaginal and CS-seeded infants. Effectively suppressed microbes are those highly associated only with unseeded CS infants, indicating that seeding excludes that microbe. Many taxa highly associated with CS-seeded infants had a seeding effectiveness score greater than 0.8, indicating that the vaginal seeding method was able to establish microbes missing in CS born babies (Fig. 4.1, 4.2, 4.3, c-e). Notably, in the infant gut, ASVs from common gut-associated genera such as *Bacteroides, Streptococcus,* and *Clostridium* were identified to be enriched in CS-seeded infants and have high seeding effectiveness scores in early time points (Fig. 4.1 c-e, Supplementary Methods 4.S8-S9). Especially of note, *Bacteroides* was consistently identified as being associated with vaginal seeding (Supplementary Methods 4.S10) using other algorithms such as ANCOM (Supplementary Methods 4.S11), MaAsLin2 (Supplementary Methods 4.S12) and LEfSe (Supplementary Methods 4.S13). In the mouth, bacteria with high seeding effectiveness scores included ASVs from *Gemellaceae, Haemophilus, and Streptococcus* (Fig.

4.2 c-e). In the skin, taxa included ASVs from *Streptococcus, Neisseria, Thermus,* and *Neisseriaceae* (Fig. 4.3 c-e). However, across all three body sites, most of the taxa associated with CS had a moderate to low seeding effectiveness score, indicating that this method was not effective at attenuating the presence of microbes typically depleted in vaginally born babies.

Maternal sites contribute to the infant microbiota

In order to determine which body sites from the mother were most likely to have the highest contributions towards shaping the infant microbiome, we also used the source-tracking tool FEAST (Shenhav et al., 2019). The first 2 days of life showed a prominent maternal vaginal source in the oral and skin sites of infants exposed to vaginal fluids; however, within the first few days, a large proportion of the microbiota colonizing the infants' sites was shared with the corresponding maternal site, regardless of birth mode or seeding status (Fig. 4.4). Selection by the specific body site was evidenced by the lack of overrepresentation of *Lactobacillus*, a dominant member of the mother's vagina, among infants born vaginally or exposed to the vaginal gauze when compared to CS-born babies. Unsurprisingly, we found that the infant oral microbiota most resembled that of the mother's mouth and areola (Fig. 4.4 h, k), and that the infant skin microbiota resembled that of the mother's skin (Fig. 4.4 o), consistent with exposure patterns and differential selection exerted by different body sites in the baby.

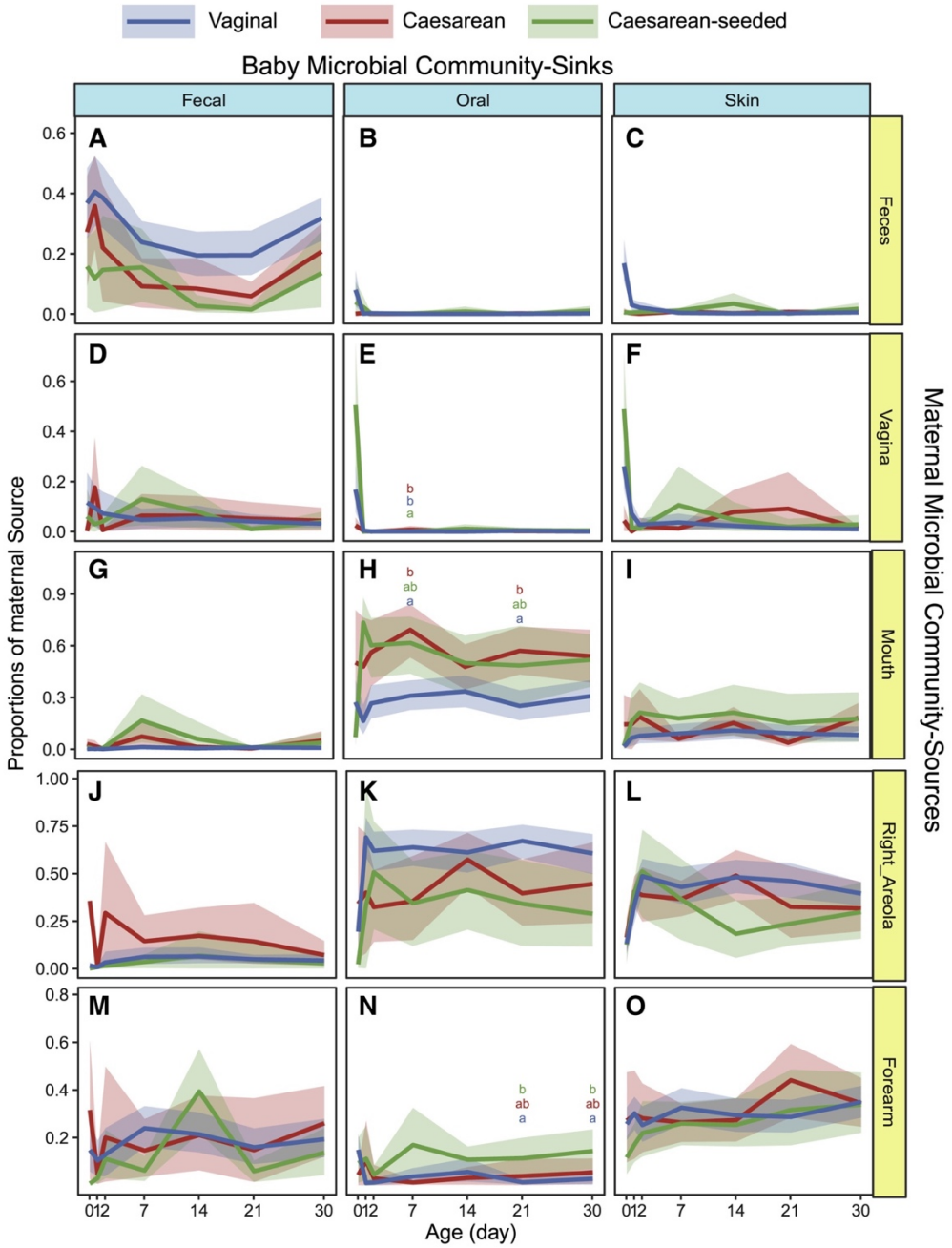

Figure 4.4. Microbial source tracking of the neonate microbiome (first month) through fast expectation-maximization microbial source tracking (FEAST). Contributions (y-axes) of various maternal sources (rows) to the infant microbial community (columns) are estimated across age in days (x-axes) for the first month of life, in 15 mother-baby pairs. Error bars show 95% confidence interval of the mean calculated by bootstrapping; Dunn test based on Kruskal-Wallis were performed on each time points by each maternal source for each baby sink, significant differences are marked by different letters in each panel. The vaginal source -prominent in day “0” for oral and skin in babies exposed to vaginal fluids (vaginal and CS-seeded; panel e, f)- as not prominent later in any baby site. Baby site specific communities resemble the corresponding maternal site (panels a, h, o), consistent with specific site selection of bacteria. The maternal right areola appears as a source for baby oral bacteria (panel k), which likely means that baby oral bacteria is transmitted to the mother’s areola during lactation.

Interestingly, we observed a notable taxonomic overlap between the maternal vagina and other maternal body sites, especially feces, on the day of giving birth: nearly 30% of the bacterial ASVs in vaginal samples were shared with feces (5.5% with feces alone, and 24.5% with feces and some other body sites), and 22.3% with more distant body sites such as arm skin, mouth, and nose (Fig. 4.5 a, Supplementary Methods 4.S14). These trends showing the pluripotent nature of the perinatal vaginal microbiome held true when examining the mothers in different countries from this study, despite variations in specific proportions (Fig. 4.S2b,c, Supplementary Methods 4.S15). In contrast, women who are not pregnant -from the HMP study- shared less than 20% of vaginal ASVs with other body sites, predominantly with skin, and none with fecal samples (Fig. 4.5 b). Together, these results point to the importance of maternal sources of microbes on the developing infant consortium.

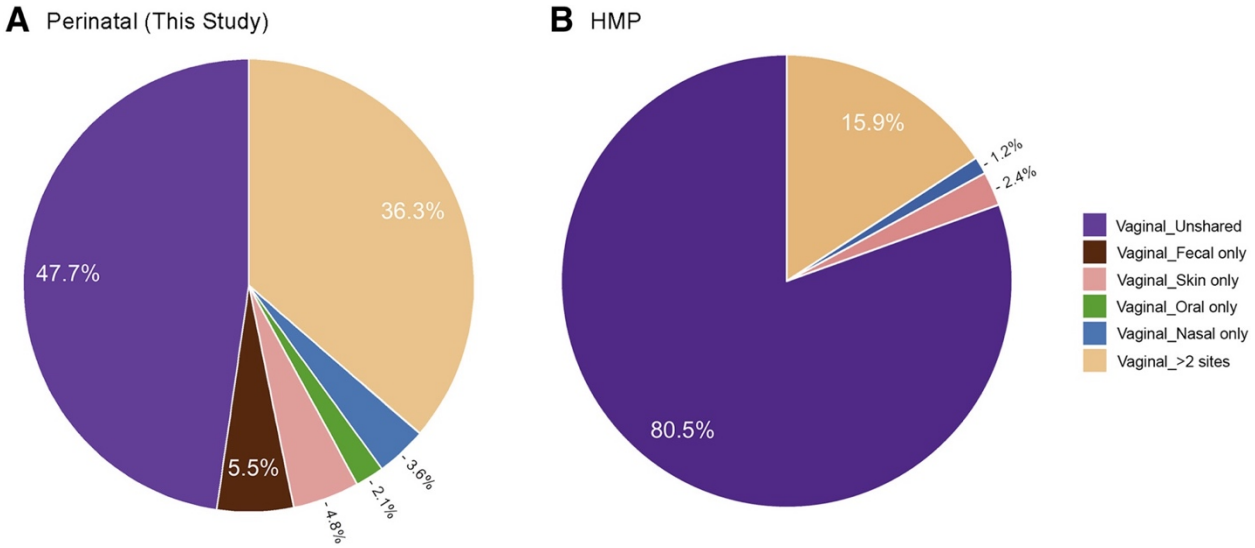

Figure 4.5. Proportions of bacterial vaginal ASVs shared with other body sites in the mothers of the current study, at the day of delivery (a) and in non-pregnant women (b). (a) V4 sequences from vaginal swabs and gauzes obtained from 97 parturient mothers form this study at the day of birth. Current study data were sequenced by Illumina HiSeq and processed by QIIME2 using the same pipeline as for the HMP data. (b) HMP V4 data from vaginal swabs obtained from 105 non-pregnant women; ASVs included in the analyses were present in at least 10% of the samples in the respective body site. Roche 454 V3V5 sequences were trimmed to obtain the V4 region.
See also Figure S2 b-c, Supplementary Methods S13-S14.

4.3. Discussion

This intervention study expands the findings of previous smaller studies, further demonstrating that microbial differences associated with delivery mode can be reduced by exposure to a vaginal microbial source at birth. The study only included scheduled C-sections on healthy mothers (mostly due to multiple previous C-sections and to malposition presentations), since infants born by emergency C-section after rupture of the chorioamniotic membrane are likely exposed the maternal microbes, given enough time before the C-section procedure (Azad et al., 2013).

Using advanced and longitudinally-aware methods, we found that birth mode significantly differentiated infant gut and skin microbiome development, and that seeding worked to adjust the trajectory of CS-delivered infants through partial restoration of microbiome features associated with a vaginal delivery. For example, differential abundance analyses confirmed previous findings that in the gut, *Bacteroides* and *Parabacteroides*-- both common gut-associated taxa--are highly associated with vaginally born infants. Our study further shows that seeding works to effectively restore these and other taxa associated with a vaginal birth. However, there are several other taxa that do not appear to establish well in the seeded infants (e.g. *Bilophila*). Also of interest, while we observed a significant association of *Enterococcus* with CS-born infants (which in previous studies has been noted as a potential opportunistic pathogen), we did not see a weakened association of this taxon, or most other CS-associated taxa, with seeded babies. Further research is needed to determine why certain gut taxa may show a higher effectiveness for seeding while other taxa may exhibit more resilience after a seeding procedure, and what roles these microbes may play in the developing infant microbiome.

An interesting facet of our study is the finding that vaginal seeding led to converging microbial compositions in the infant gut, despite the exposure coming from a vaginal source. The same pattern was observed in the skin environment. Our results clearly indicate that from very

early timepoints, the microbiota of an infant largely resembles the same maternal site, supporting the idea of strong site selection occurring from very early ages (i.e. that different body sites will select for specific microbes out of a diverse population). This is further supported by the finding that *Lactobacillaceae*, the most dominant member of the mother's vagina, was not identified as one of the most differentially abundant taxa among infants at any of the three body sites observed. Site selection is also consistent with the recent evidence of successful engraftment after fecal microbiota transplant from the mother to CS neonates (Korpela et al., 2018), and with previous evidence of fecal bacteria in the infant gut (Ferretti et al., 2018; Helve et al., 2019). Indeed, bacterial transfer from homologous sites from the mother and other family members surely occur after birth. However, this may only be a part of the story. Our results show that unlike in non-pregnant women, more ASVs from the vaginal microbiome from parturient women overlap with those in other body sites, mostly the proximal rectum (which in mammals is next to the reproductive canal), but also more distant sites. This strongly suggests a pluripotent capacity of vaginal fluids to seed different sites of the baby's body. Transmission and colonization by these pioneer species may then modulate the succession that proceeds, influencing engraftment of later colonizers to each body site (Martinez et al., 2018). Major changes in the vaginal microbiota during pregnancy have been described (Stout et al., 2017), although the changes in the last semester have not been deeply characterized. This begs the question of whether the vaginal microbiome becomes specifically primed during pregnancy to deliver key pioneer colonizers tailored towards multiple body sites of the infant. This hypothesis is supported by previous work demonstrating the by-phasic dynamics in gestational changes in which after decreasing diversity in the first two thirds of gestation, in the last gestational trimester diversity increases at the expense of *Lactobacillus* from week 24 of pregnancy until birth (Rasmussen et al., 2020); increase in vaginal diversity continues in the postpartum vaginal tract for up to 1 year following birth (DiGiulio et al., 2015).

This study provides solid evidence that deliberate, early microbial seeding can help naturalize the microbiome developmental trajectory of CS born infants. While overall trajectories

do appear to head towards convergence over time, studies show that early perturbations during the crucial developmental window of very early life seem to have irreversible consequences (Huh et al., 2012; Pistiner et al., 2008; Sevelsted et al., 2015; Thavagnanam et al., 2008). Restoring natural exposures at birth may thus be one way to reduce the risk of CS-associated diseases such as obesity, asthma, allergies, and immune disfunctions. However, randomized clinical trials on large cohorts are needed to gain conclusive evidence for microbial restoration at birth improving health outcomes (Mueller et al., 2019). Moreover, in light of recent research showing that oral administration of maternal fecal microbes is also effective in restoring the microbiome in CS-delivered infants (Korpela et al., 2020), future research investigating the effects of exposure to both sources explicitly compared to either single source will help determine the best routes to restoring the neonate microbiome. In this study we exposed infants to freshly collected maternal vaginal/perineal microbes, but it is unknown how storage would alter the microbiota composition. More research is needed to determine whether it is optimum that they receive their own mother's microbiome or achieve defined universal cocktails that can be used to restore neonates.

Limitations of study

This study is limited by the cohort size, particularly in countries outside of the United States, the follow up time of the first year of life, since any longer-term consequences of seeding were not assessed, and by the *16S rDNA* amplicon sequencing, which excludes functional characterizations as well as fungi and viruses. Future studies capturing longer timeframes, larger and broader cultural and geographic representations, and additional data types are needed to gain a more understanding of how seeding affects the microbiome and ultimately the health of CS-delivered infants.

4.4. Acknowledgements

This work was partially supported by the C&D Research Fund, Emch Fund for human microbiome studies, and CIFAR FS20-078 #125869 (M.G.D.B), the Chilean CONICYT PIA/ANILLO Grant ACT172097, the Chilean SOCHIPE Project 022019 (P.H.), the Norwegian Institute of Public Health (2019-0350), the Emerald Foundation, the NIH Pioneer award (1DP1AT010885), the National Institute of Justice (2016-DN-BX-4194), the San Diego Digestive Diseases Research Center (NIDDK 1P30DK120515), Janssen Pharmaceuticals (20175015), and by in-kind donations from Illumina, MoBio/QIAGEN, and the Center for Microbiome Innovation at UC San Diego (R.K.). We acknowledge the contribution of students, technicians and MDs who participated in obtaining the samples and preparing the metadata: Sukhleen Bedi, Allison Horan, and Yi Cai from NYUSM; Noraliz Garcia, Hebe Rosado, Selena Marie Rodriguez, and Keimari Mendez from UPR; Maricruz Mojica, Magaly Magariños and Myriam Corrales from Universidad Real y Pontificia San Francisco Javier de Chuquisaca, and the University Hospital in Sucre, Bolivia; Mauricio Sandoval, Marlene Ortiz, Carolina Serrano from the PUC. We thank Gail Ackermann for her help curating the metadata and uploading sequences to Qiita, and James T. Morton for providing advice in the use of Songbird differentials.

Author Contributions

M.G.D.B. designed the study.

M.G.D.B., S.J.S, J.W., P.R.H., C.D.H., N.H., E.A., D.N., C.G., V.M., W.S., M.J., J.C., H.S., M.C.C., I.G.M, F.G.R., J.I.R.V., M.C.R., J.F.R.C., R.K. collected and processed specimens.

M.G.D.B., R.K. sequenced and generated data.

M.G.D.B., S.J.S., J.W., C.M., L.J., W.K.T., L.S., D.McD., R.K. analyzed data including performing, overseeing the statistical analysis.

M.G.D.B. and R.K. has unrestricted access to all data.

M.G.D.B. S.J.S. and R.K. drafted the manuscript.

All authors reviewed, agreed to submit the final manuscript, read and approved the final draft and take full responsibility of its contents, including the accuracy of the data and the fidelity of the trial to the registered protocol and its statistical analysis.

Declaration of Interest

New York University has filed an U.S. patent application (number 62161549) on behalf of M.G.D.B., related to methods for restoring the microbiota of newborns.

4.5 References

1. Aitchison, J. (1982). The Statistical-Analysis of Compositional Data. J Roy Stat Soc B Met *44*, 139-177.

2. Aitchison, J. (1983). Principal Component Analysis of Compositional Data. Biometrika *70*, 57-65.

3. Al Nabhani, Z., and Eberl, G. (2020). Imprinting of the immune system by the microbiota early in life. Mucosal Immunol *13*, 183-189.

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

5. Andersen, V., Moller, S., Jensen, P.B., Moller, F.T., and Green, A. (2020). Caesarean Delivery and Risk of Chronic Inflammatory Diseases (Inflammatory Bowel Disease, Rheumatoid Arthritis, Coeliac Disease, and Diabetes Mellitus): A Population Based Registry Study of 2,699,479 Births in Denmark During 1973-2016. Clin Epidemiol *12*, 287-293.

6. Anderson, M.J. (2017). Permutational Multivariate Analysis of Variance (PERMANOVA). In Wiley StatsRef: Statistics Reference Online, pp. 1-15.

7. Ardic, C., Usta, O., Omar, E., Yildiz, C., and Memis, E. (2020). Caesarean delivery increases the risk of overweight or obesity in 2-year-old children. J Obstet Gynaecol, 1-6.

8. Azad, M.B., Konya, T., Maughan, H., Guttman, D.S., Field, C.J., Chari, R.S., Sears, M.R., Becker, A.B., Scott, J.A., Kozyrskyj, A.L.*, et al.* (2013). Gut microbiota of healthy Canadian infants: profiles by mode of delivery and infant diet at 4 months. CMAJ *185*, 385-394.

9. Blustein, J., Attina, T., Liu, M., Ryan, A.M., Cox, L.M., Blaser, M.J., and Trasande, L. (2013). Association of caesarean delivery with child adiposity from age 6 weeks to 15

years. Int J Obes (Lond) *37*, 900-906.

10. Bokulich, N.A., Chung, J., Battaglia, T., Henderson, N., Jay, M., Li, H., A, D.L., Wu, F., Perez-Perez, G.I., Chen, Y.*, et al.* (2016). Antibiotics, birth mode, and diet shape microbiome maturation during early life. Sci Transl Med *8*, 343ra382.

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

12. Braniste, V., Al-Asmakh, M., Kowal, C., Anuar, F., Abbaspour, A., Toth, M., Korecka, A., Bakocevic, N., Ng, L.G., Kundu, P.*, et al.* (2014). The gut microbiota influences blood-brain barrier permeability in mice. Sci Transl Med *6*, 263ra158.

13. Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S.M., Betley, J., Fraser, L., Bauer, M.*, et al.* (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J *6*, 1621-1624.

14. Cox, L.M., Yamanishi, S., Sohn, J., Alekseyenko, A.V., Leung, J.M., Cho, I., Kim, S.G., Li, H., Gao, Z., Mahana, D.*, et al.* (2014). Altering the intestinal microbiota during a critical developmental window has lasting metabolic consequences. Cell *158*, 705-721.

15. DiGiulio, D.B., Callahan, B.J., McMurdie, P.J., Costello, E.K., Lyell, D.J., Robaczewska, A., Sun, C.L., Goltsman, D.S.A., Wong, R.J., Shaw, G.*, et al.* (2015). Temporal and spatial variation of the human microbiota during pregnancy. Proceedings of the National Academy of Sciences *112*, 11060.

16. Dominguez-Bello, M.G., Costello, E.K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., and Knight, R. (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. Proc Natl Acad Sci U S A *107*, 11971-11975.

17. Dominguez-Bello, M.G., De Jesus-Laboy, K.M., Shen, N., Cox, L.M., Amir, A., Gonzalez, A., Bokulich, N.A., Song, S.J., Hoashi, M., Rivera-Vinas, J.I.*, et al.* (2016). Partial restoration of the microbiota of cesarean-born infants via vaginal microbial transfer. Nat Med *22*, 250-253.

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

19. Ferretti, P., Pasolli, E., Tett, A., Asnicar, F., Gorfer, V., Fedi, S., Armanini, F., Truong, D.T., Manara, S., Zolfo, M.*, et al.* (2018). Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. Cell Host Microbe *24*, 133-145 e135.

20. Gensollen, T., Iyer, S.S., Kasper, D.L., and Blumberg, R.S. (2016). How colonization by

microbiota in early life shapes the immune system. Science (New York, NY) *352*, 539-544.

21. Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V., and Egozcue, J.J. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. Front Microbiol *8*, 2224.

22. Halfvarson, J., Brislawn, C.J., Lamendella, R., Vazquez-Baeza, Y., Walters, W.A., Bramer, L.M., D'Amato, M., Bonfiglio, F., McDonald, D., Gonzalez, A.*, et al.* (2017). Dynamics of the human gut microbiome in inflammatory bowel disease. Nat Microbiol *2*, 17004.

23. Helve, O., Korpela, K., Kolho, K.-L., Saisto, T., Skogberg, K., Dikareva, E., Stefanovic, V., Salonen, A., de Vos, W.M., and Andersson, S. (2019). 2843. Maternal Fecal Transplantation to Infants Born by Cesarean Section: Safety and Feasibility. Open Forum Infect Dis *6*, S68-S68.

24. Huh, S.Y., Rifas-Shiman, S.L., Zera, C.A., Edwards, J.W., Oken, E., Weiss, S.T., and Gillman, M.W. (2012). Delivery by caesarean section and risk of obesity in preschool age children: a prospective cohort study. Arch Dis Child *97*, 610-616.

25. Jakobsson, H.E., Abrahamsson, T.R., Jenmalm, M.C., Harris, K., Quince, C., Jernberg, C., Bjorksten, B., Engstrand, L., and Andersson, A.F. (2014). Decreased gut microbiota diversity, delayed Bacteroidetes colonisation and reduced Th1 responses in infants delivered by caesarean section. Gut *63*, 559-566.

26. James, G.M., Hastie, T.J., and Sugar, C.A. (2000). Principal component models for sparse functional data. Biometrika *87*, 587-602.

27. Janssen, S., McDonald, D., Gonzalez, A., Navas-Molina, J.A., Jiang, L., Xu, Z.Z., Winker, K., Kado, D.M., Orwoll, E., Manary, M.*, et al.* (2018). Phylogenetic Placement of Exact Amplicon Sequences Improves Associations with Clinical Information. mSystems *3*.

28. Jari, O., Blanchet, F.G., Michael, F., Roeland, K., Pierre, L., Dan, M., Peter, R.M., Hara, R.B.O., Gavin, L.S., Peter, S.*, et al.* (2019). vegan: Community Ecology Package.

29. Jiang, L., Zhong, Y., Elrod, C., Natarajan, L., Knight, R., and Thompson, W.K. (2020). BayesTime: Bayesian Functional Principal Components for Sparse Longitudinal Data. Arxiv.

30. Korpela, K., Costea, P., Coelho, L.P., Kandels-Lewis, S., Willemsen, G., Boomsma, D.I., Segata, N., and Bork, P. (2018). Selective maternal seeding and environment shape the human gut microbiome. Genome Res *28*, 561-568.

31. Korpela, K., Helve, O., Kolho, K.L., Saisto, T., Skogberg, K., Dikareva, E., Stefanovic, V., Salonen, A., Andersson, S., and de Vos, W.M. (2020). Maternal Fecal Microbiota Transplantation in Cesarean-Born Infants Rapidly Restores Normal Gut Microbial Development: A Proof-of-Concept Study. Cell *183*, 324-334 e325.

32. Livanos, A.E., Greiner, T.U., Vangay, P., Pathmasiri, W., Stewart, D., McRitchie, S., Li, H., Chung, J., Sohn, J., Kim, S.*, et al.* (2016). Antibiotic-mediated gut microbiome perturbation accelerates development of type 1 diabetes in mice. Nat Microbiol *1*, 16140.

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

34. Malamitsi-Puchner, A., Protonotariou, E., Boutsikou, T., Makrakis, E., Sarandakou, A., and Creatsas, G. (2005). The influence of the mode of delivery on circulating cytokine concentrations in the perinatal period. Early Hum Dev *81*, 387-392.

35. Martinez, I., Maldonado-Gomez, M.X., Gomes-Neto, J.C., Kittana, H., Ding, H., Schmaltz, R., Joglekar, P., Cardona, R.J., Marsteller, N.L., Kembel, S.W*., et al.* (2018). Experimental evaluation of the importance of colonization history in early-life gut microbiota assembly. Elife *7*.

36. Martinez, K.A., 2nd, Devlin, J.C., Lacher, C.R., Yin, Y., Cai, Y., Wang, J., and Dominguez-Bello, M.G. (2017). Increased weight gain by C-section: Functional significance of the primordial microbiome. Sci Adv *3*, eaao1874.

37. Martino, C., Shenhav, L., Marotz, C.A., Armstrong, G., McDonald, D., Vázquez-Baeza, Y., Morton, J.T., Jiang, L., Dominguez-Bello, M.G., Swafford, A.D*., et al.* (2020). Context-aware dimensionality reduction deconvolutes gut microbial community dynamics. Nature Biotechnology, 1-4.

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

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

40. Moya-Perez, A., Luczynski, P., Renes, I.B., Wang, S., Borre, Y., Anthony Ryan, C., Knol, J., Stanton, C., Dinan, T.G., and Cryan, J.F. (2017). Intervention strategies for cesarean section-induced alterations in the microbiota-gut-brain axis. Nutr Rev *75*, 225-240.

41. Mueller, N.T., Dominguez-Bello, M.G., Appel, L.J., and Hourigan, S.K. (2019). 'Vaginal seeding' after a caesarean section provides benefits to newborn children: FOR: Does exposing caesarean-delivered newborns to the vaginal microbiome affect their chronic disease risk? The critical need for trials of 'vaginal seeding' during caesarean section. BJOG.

42. Olszak, T., An, D., Zeissig, S., Vera, M.P., Richter, J., Franke, A., Glickman, J.N., Siebert, R., Baron, R.M., Kasper, D.L*., et al.* (2012). Microbial exposure during early life has persistent effects on natural killer T cell function. Science *336*, 489-493.

43. Pistiner, M., Gold, D.R., Abdulkerim, H., Hoffman, E., and Celedon, J.C. (2008). Birth by cesarean section, allergic rhinitis, and allergic sensitization among children with a parental history of atopy. J Allergy Clin Immunol *122*, 274-279.

44. Rasmussen, M.A., Thorsen, J., Dominguez-Bello, M.G., Blaser, M.J., Mortensen, M.S.,

Brejnrod, A.D., Shah, S.A., Hjelmsø, M.H., Lehtimäki, J., Trivedi, U.*, et al.* (2020). Ecological succession in the vaginal microbiota during pregnancy and birth. The ISME Journal *14*, 2325-2335.

45. Schiffer, L., Azhar, R., Shepherd, L., Ramos, M., Geistlinger, L., Huttenhower, C., Dowd, J.B., Segata, N., and Waldron, L. (2019). HMP16SData: Efficient Access to the Human Microbiome Project Through Bioconductor. Am J Epidemiol *188*, 1023-1026.

46. Sevelsted, A., Stokholm, J., Bonnelykke, K., and Bisgaard, H. (2015). Cesarean section and chronic immune disorders. Pediatrics *135*, e92-98.

47. Shao, Y., Forster, S.C., Tsaliki, E., Vervier, K., Strang, A., Simpson, N., Kumar, N., Stares, M.D., Rodger, A., Brocklehurst, P.*, et al.* (2019). Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. Nature *574*, 117-121.

48. Shenhav, L., Thompson, M., Joseph, T.A., Briscoe, L., Furman, O., Bogumil, D., Mizrahi, I., Pe'er, I., and Halperin, E. (2019). FEAST: fast expectation-maximization for microbial source tracking. Nat Methods *16*, 627-632.

49. Stewart, C.J., Ajami, N.J., O'Brien, J.L., Hutchinson, D.S., Smith, D.P., Wong, M.C., Ross, M.C., Lloyd, R.E., Doddapaneni, H., Metcalf, G.A.*, et al.* (2018). Temporal development of the gut microbiome in early childhood from the TEDDY study. Nature *562*, 583-588.

50. Stokholm, J., Blaser, M.J., Thorsen, J., Rasmussen, M.A., Waage, J., Vinding, R.K., Schoos, A.M., Kunoe, A., Fink, N.R., Chawes, B.L.*, et al.* (2018). Maturation of the gut microbiome and risk of asthma in childhood. Nat Commun *9*, 141.

51. Stout, M.J., Zhou, Y., Wylie, K.M., Tarr, P.I., Macones, G.A., and Tuuli, M.G. (2017). Early pregnancy vaginal microbiome trends and preterm birth. Am J Obstet Gynecol *217*, 356 e351-356 e318.

52. Thavagnanam, S., Fleming, J., Bromley, A., Shields, M.D., and Cardwell, C.R. (2008). A meta-analysis of the association between Caesarean section and childhood asthma. Clin Exp Allergy *38*, 629-633.

53. Thompson, L.R., Sanders, J.G., McDonald, D., Amir, A., Ladau, J., Locey, K.J., Prill, R.J., Tripathi, A., Gibbons, S.M., Ackermann, G.*, et al.* (2017). A communal catalogue reveals Earth's multiscale microbial diversity. Nature *551*, 457-463.

54. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J.*, et al.* (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods *17*, 261-272.

55. Yassour, M., Vatanen, T., Siljander, H., Hamalainen, A.M., Harkonen, T., Ryhanen, S.J., Franzosa, E.A., Vlamakis, H., Huttenhower, C., Gevers, D.*, et al.* (2016). Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. Sci Transl Med *8*, 343ra381.

56. Zaneveld, J.R., McMinds, R., and Vega Thurber, R. (2017). Stress and stability: applying

the Anna Karenina principle to animal microbiomes. Nat Microbiol *2*, 17121.

Appendix A. Supplemental Information for Robust Aitchison PCA reveals microbiome perturbations

AA.1. Supplemental Figures

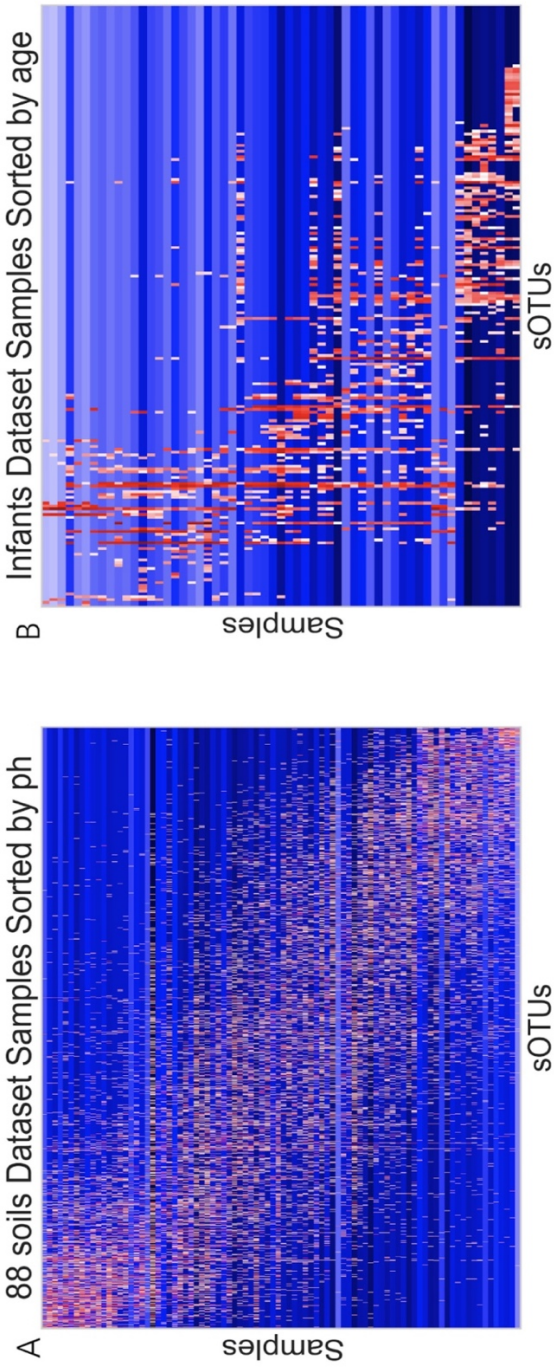

Figure AA.1.S1. Sorted heatmap plots of example gradient structured high rank datasets for 88 soils (A) and infant development dataset (B).

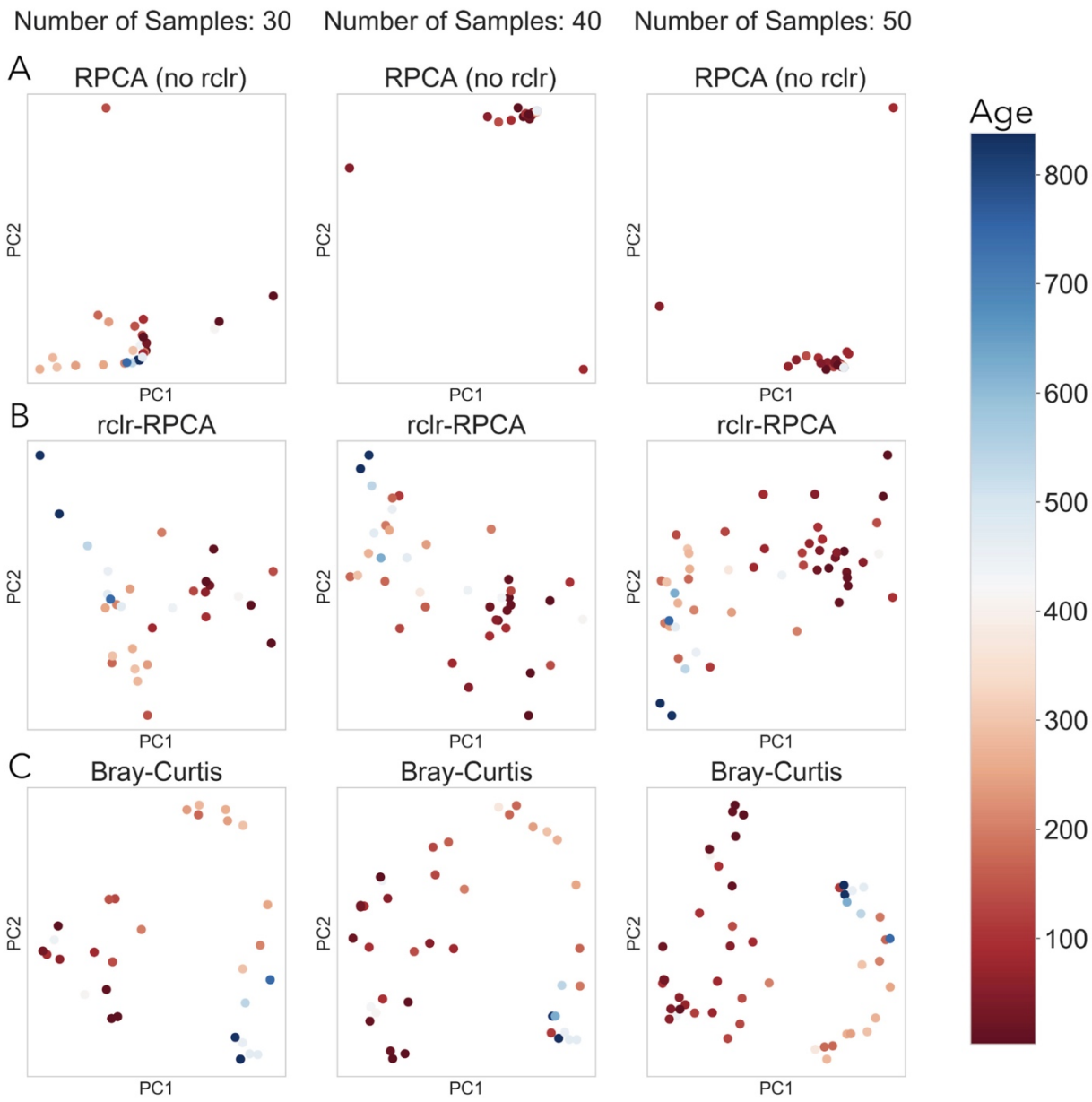

Figure AA.1.S2. Comparison of methods RPCA without rclr (A), RPCA with rclr (B), and Bray-Curtis (C) in high-rank infant development dataset at varying number of samples of 30, 40, and 50 from left to right.

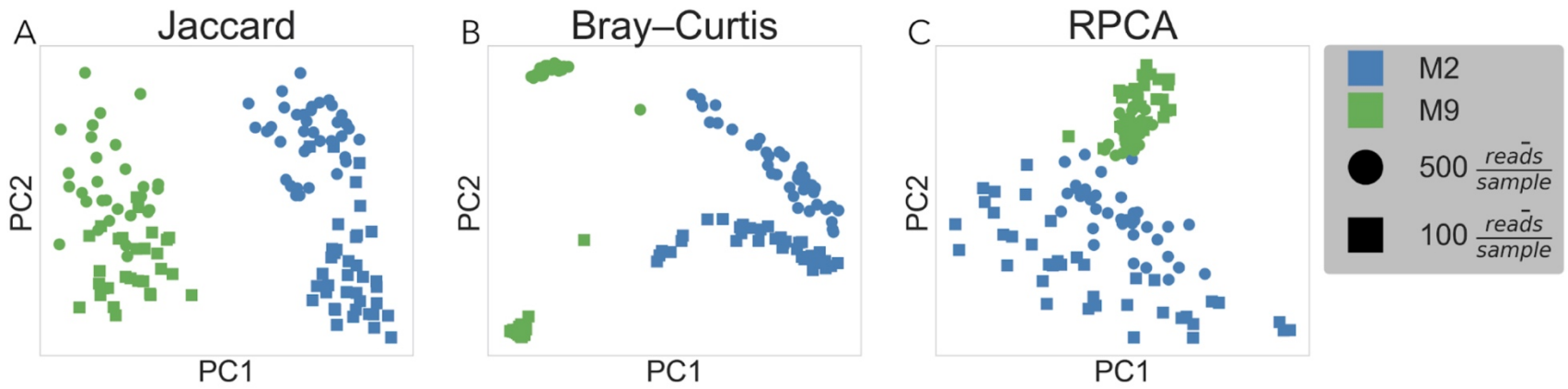

Figure AA.1.S3. Comparison between two subjects (geen and blue) from the keyboard dataset compared between sequencing depth 500 (circles) and 100 (squares) over different beta-diversity methods Jaccard (A), Bray-Curtis (B), RPCA (C).

AA.2. Supplemental Tables

Table AA.2.S1. Comparison of PERMANOVA and KNN classifier accuracy between positive- and negative-control simulations.

Negative Control			Positive Control		
F-statistic	**p-value**	**Accuracy**	**F-statistic**	**p-value**	**Accuracy**
0.650725	0.512	0.65	158.543171	<0.001	1
PERMANOVA	**PERMANOVA**	**KNN-Classifier**	**PERMANOVA**	**PERMANOVA**	**KNN-Classifier**

Table AA.2.S2. Comparison of PERMANOVA F-statistic and p-value between between subject id clusters in the keyboard dataset with uneven sequencing of 500 and 100 reads/sample.

	Subject ID	
	F-statistic	**p-value**
RPCA (with rclr)	**75.5**	<.001
Bray-Curtis	54.6	<.001
Jaccard	6.26	<.001

Appendix B. Supplemental Information for Context-Aware Dimensionality Reduction Deconvolutes Dynamics of Gut Microbial Community Development

AB.1 Supplemental Figures

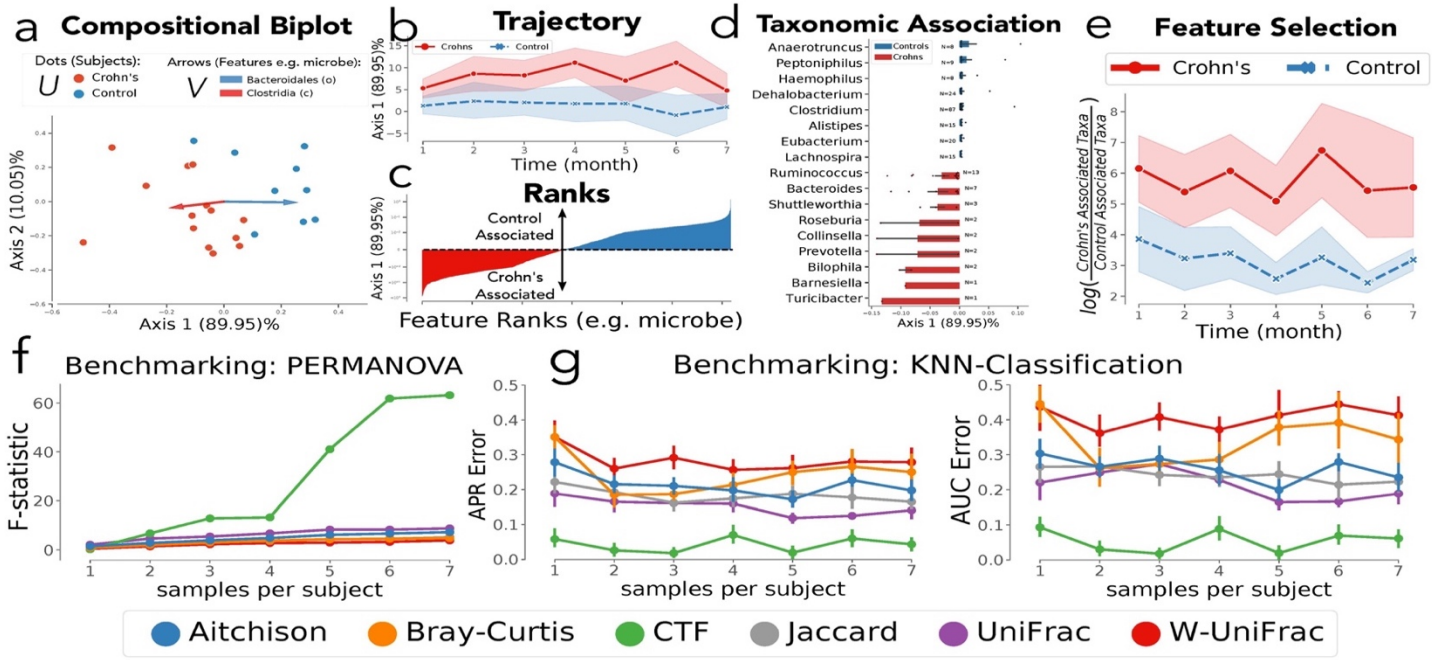

Figure AB.1.S1. IBD dataset benchmarking. CTF was applied to longitudinal 16S data from Halfvarson et al. Multiple distinct downstream analyses were generated from the CTF output: (a) The subject (U) and feature (V) loadings can be visualized as compositional biplot ordinations, using the top two ordination axes where each point represents a subject's time series and the arrows represent the top-ranked microbial features differentiating the subjects. (b) The compositional biplot axis driving phenotypically relevant separation was used to track sample distance over time (referred to here as a 'trajectory'). This trajectory demonstrates that Crohn's disease samples are compositionally distinct from healthy control samples across the entire sampling timeline. (c) Differentially abundant microbes most associated with the phenotypes were identified by plotting the feature rankings by the major axis of separation. The highest ranked sOTUs revealed by CTF are associated with Crohn's-disease while the lowest are associated with the control group. (d) sOTU feature rankings averaged by genus colored by control (n=9 subjects; 950 sOTUs; blue) and Crohn's (n=14 subjects; 950 sOTUs; red) (N sOTUs in each genus annotated on plot). (e) These phenotype-associated taxa were chosen based on the feature ranks to identify reference frames (log-ratios of sequencing counts) that can differentiate phenotypes, distinguishing healthy (n=9 subjects) from Crohn's (n=14 subjects) disease samples. These trajectories and differentially abundant taxa are supported by previous findings in the IBD literature8,19–21 (f) Disease status PERMANOVA f-statistic among different distance metrics. (g) K-Nearest Neighbor classification compared by AUPR (left) and AUC (right) among different distance metrics. Error bars represent standard error of the mean.

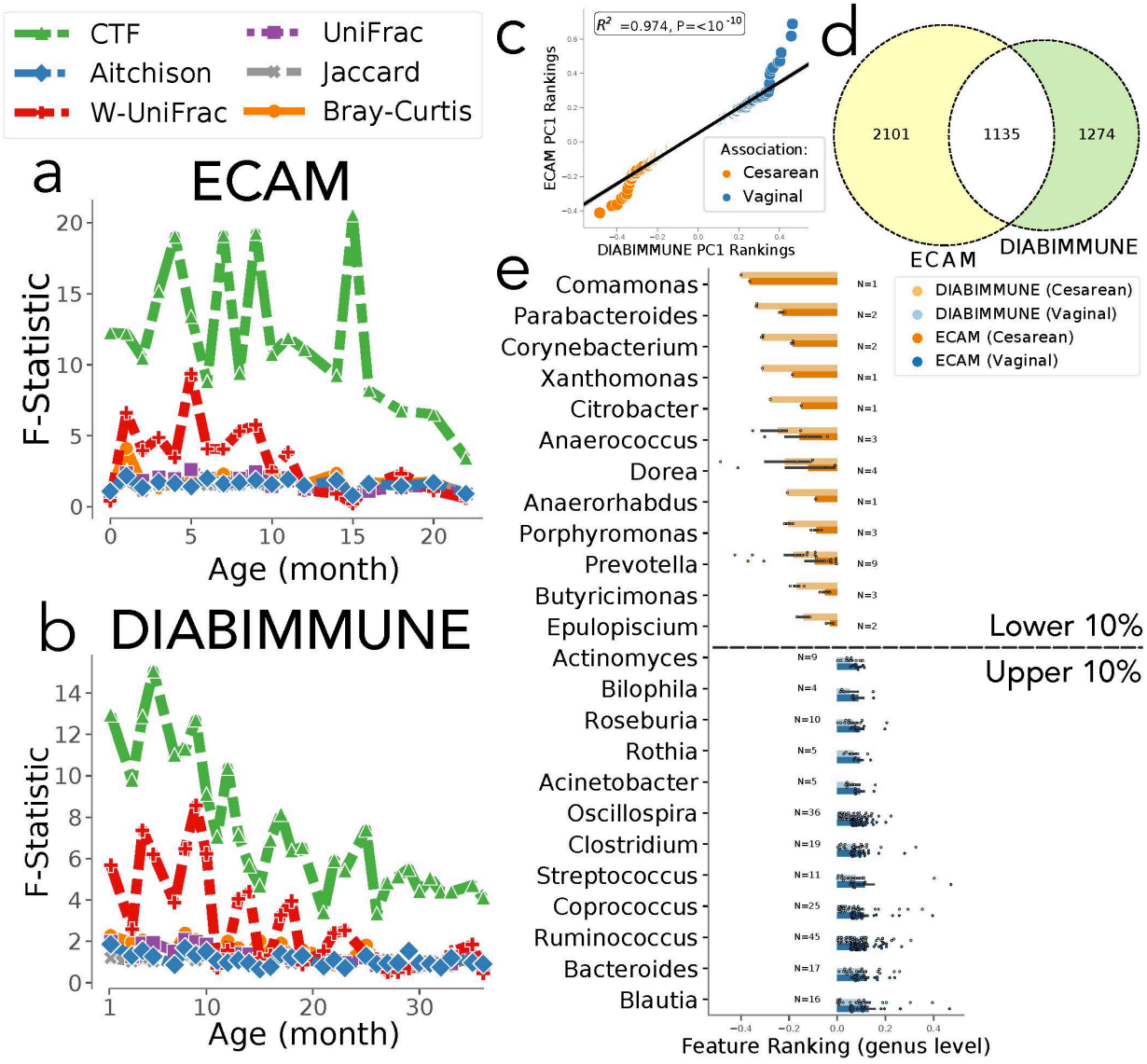

Figure AB.1.S2. Feature rankings distinguishing birth-modes across the ECAM and DIABIMMUNE datasets are tightly correlated.
(a & b) PERMANOVA F-statistic (y-axis) separating vaginal vs cesarean birth-mode colored by distance metric for ECAM (top) and DIABIMMUNE (bottom). (c) Regression plot between sOTUs ranked in ECAM and DIABIMMUNE datasets; Pearson correlation shown. (d) Venn diagram of the number of disjoint and shared sOTUs between datasets. (e) The top and bottom 10% ranked sOTUs averaged by genus in ECAM and DIABIMMUNE colored by vaginal (blue) and cesarean (orange) birth modes (N sOTUs in each genus annotated on plot). Error bars represent standard error of the mean.

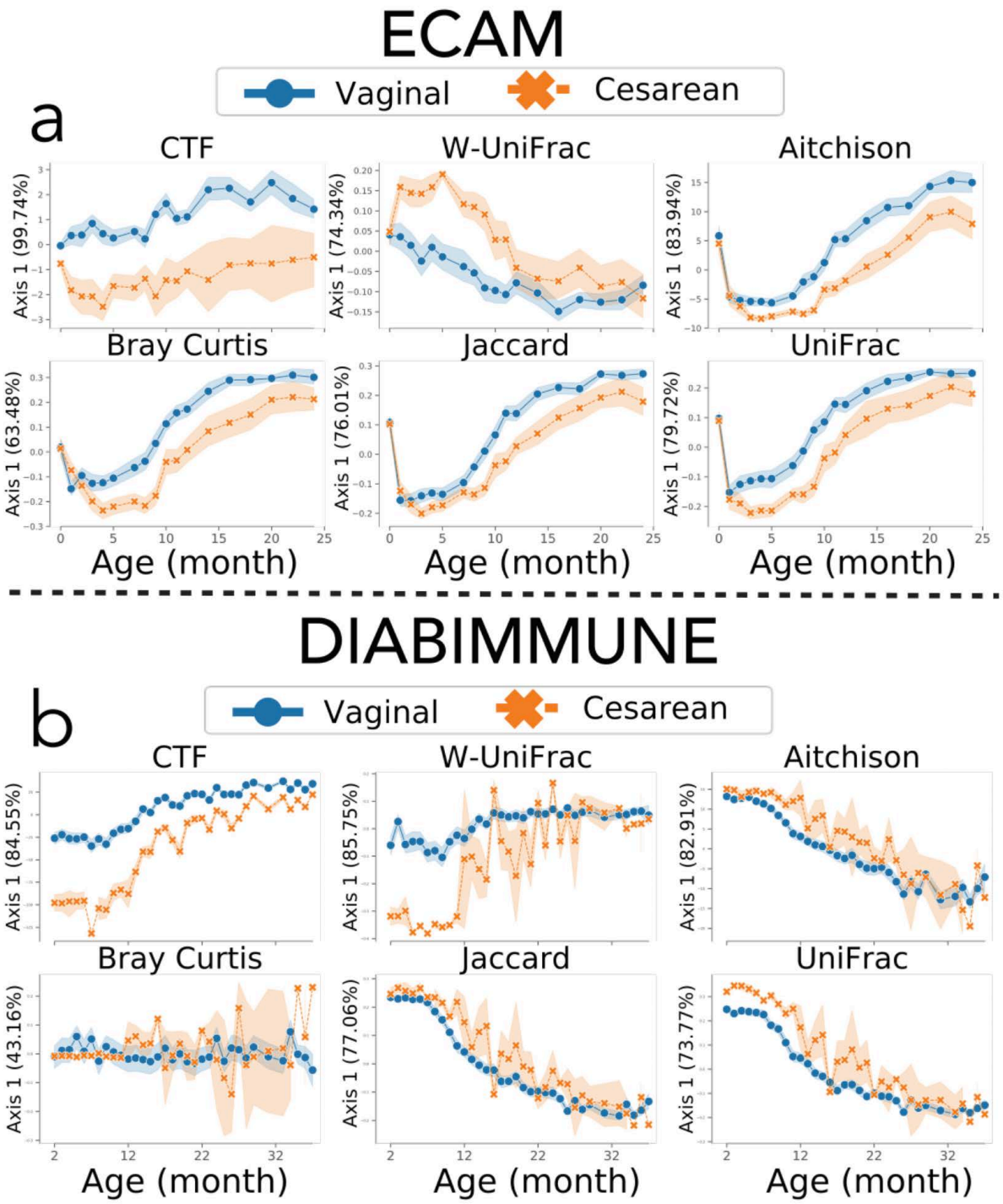

Figure AB.1.S3. CTF outperforms traditional distance metrics in distinguishing samples by birth-mode over time. (a & b) Comparison between the ECAM (top) and DIABIMMUNE (bottom) infant development studies with the first principal component (y-axes) of various distance metrics over time (x-axes) colored by vaginal (blue) and cesarean (orange) birth-modes. The relative percent explained variance is the fraction of the first component divided by the top 3 components to normalize eigenvalues among methods. Error bars represent standard error of the mean.

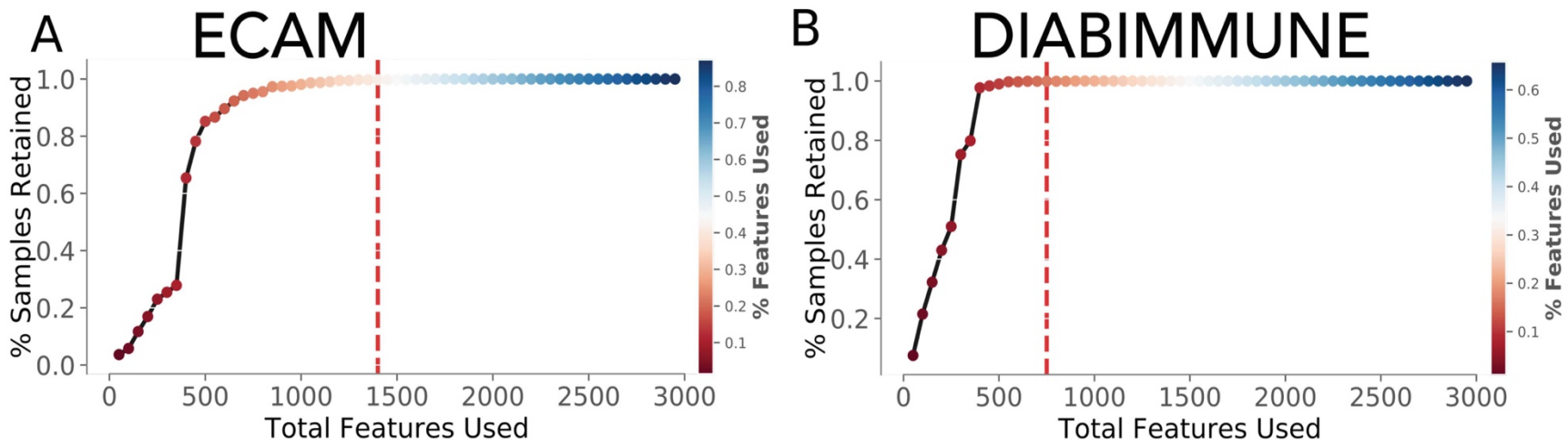

Figure AB.1.S4. Selecting the number of features used in the log-ratio to prevent sample dropouts from zeros. The percent of samples retained (y-axis) when including a number of ranked features (x-axis) in the log-ratio colored by the percent of features used from the total dataset for ECAM (A) and DIABIMMUNE (B). The red line represents the number of features used in the final log-ratio for each dataset.

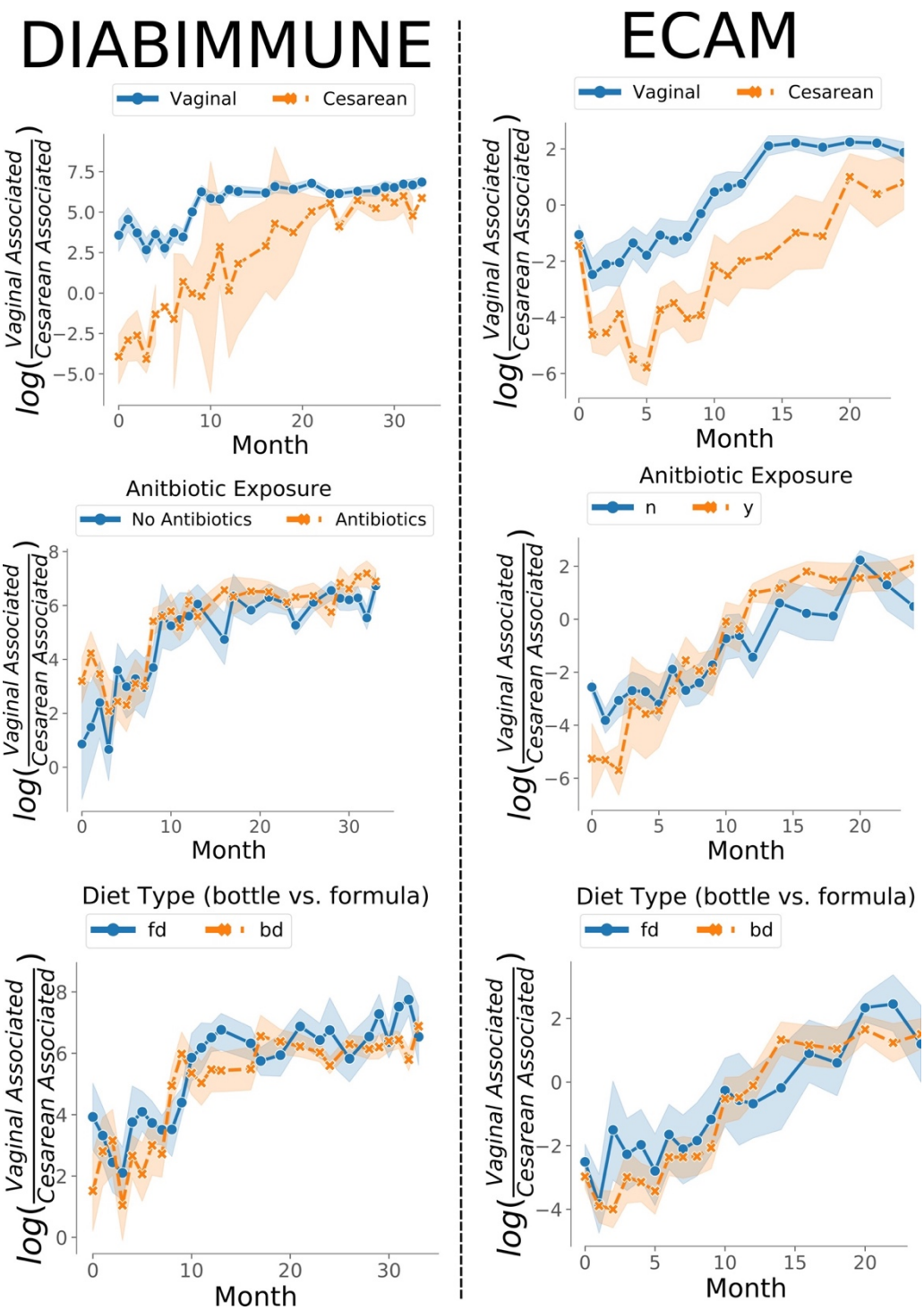

Figure AB.1.S5. Birth-mode ratios designed from CTF feature rankings distinguish samples by birth-mode over time. The log-ratios for the ECAM (1400 sOTUs) and DIABIMMUNE (750 studies are plotted on the y-axis over time (x-axis) showing separation by birth-mode using these ratios. This grouping of subjects is not confounded by antibiotics exposure (yes/no) or by diet. Error bars represent standard error of the mean.

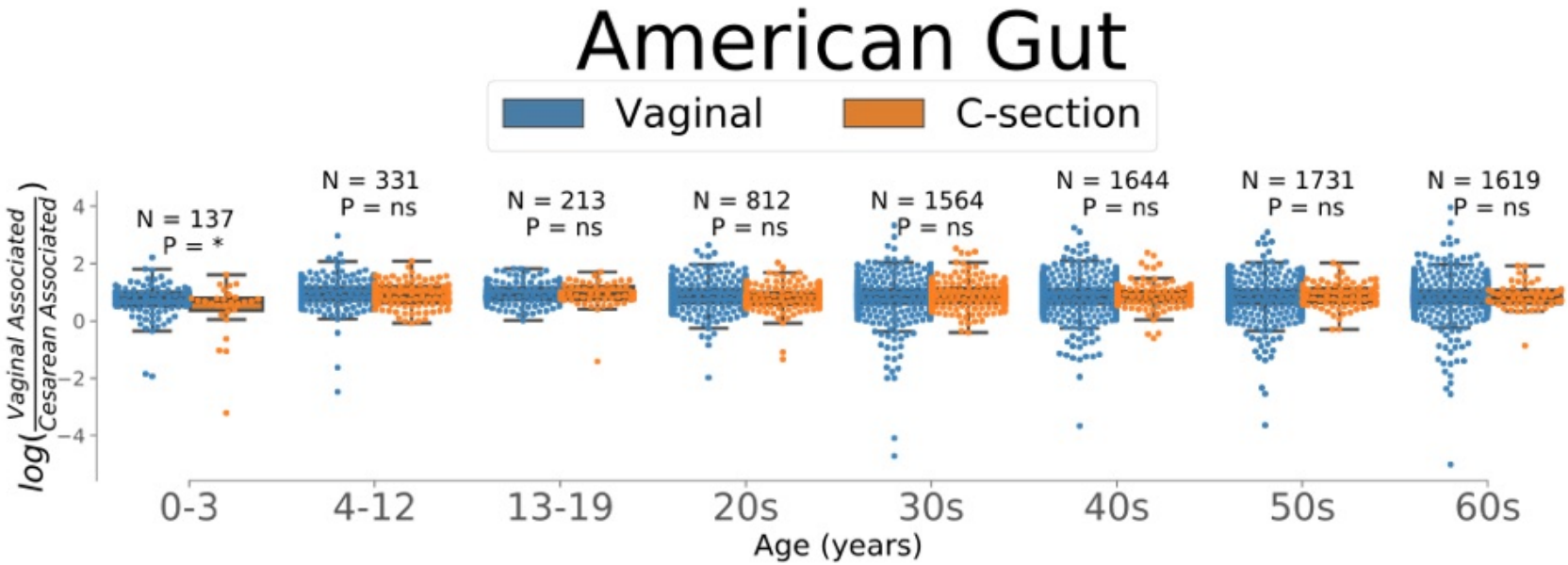

Figure AB.1.S6. Birth-mode microbial signature in AGP dataset.
Birth-mode microbial signature (log-ratio of the 532 most highly and 532 most lowly ranked features that were all identified in the DIABIMMUNE, ECAM, and AGP datasets) is plotted on the y-axis in all sub-panels. Birth-mode microbial signature over age groups (years; x-axis) colored by cesarean (orange) and vaginal (blue) birth modes in the AGP dataset. Significance was evaluated by a t-test and * denotes p=<0.01, ** denotes p=<0.001 and *** denotes p<=0.0001.

AB.2. Supplemental Tables

Table AB.2.S1. CTF shows improvement over traditional distance metrics in simulations across different sequencing depth.
Fold increase in PERMANOVA f-statistic (left) or percent increase in K-Nearest Neighbor classification (right) by CTF over other distance metrics in simulated dataset.
AUC and AUPR percent increase mean ± s.d. across all time points for mean 100-fold cross-validation at each time point. PERMANOVA F-statistic fold-increase mean ± s.d. across all time points.

		CTF Fold-Increase	CTF Percent-Increase	
Sequencing Depth (seq/sample)	**Comparison Method**	**F-stat Fold Increase**	**AUPR**	**AUC**
500	**Aitchison**	3.90 ± 1.93	23.04 ± 6.86	33.48 ± 8.92
	Bray-Curtis	4.78 ± 2.42	24.49 ± 6.66	42.61 ± 8.32
	Jaccard	4.69 ± 2.36	18.55 ± 7.24	26.74 ± 12.31
	UniFrac	3.01 ± 1.48	18.99 ± 4.94	28.48 ± 5.89
	W-UniFrac	5.55 ± 3.08	18.91 ± 7.63	38.70 ± 15.35
1000	**Aitchison**	3.20 ± 1.28	23.62 ± 6.50	33.48 ± 10.28
	Bray-Curtis	4.23 ± 1.80	33.04 ± 7.58	51.09 ± 12.69
	Jaccard	3.66 ± 1.61	23.77 ± 7.24	33.48 ± 10.14
	UniFrac	2.23 ± 0.85	17.39 ± 1.73	24.57 ± 2.63
	W-UniFrac	5.69 ± 2.12	26.01 ± 4.76	46.74 ± 8.80
10000	**Aitchison**	3.67 ± 1.94	22.97 ± 2.90	30.22 ± 3.60
	Bray-Curtis	6.80 ± 4.11	24.49 ± 7.02	42.17 ± 12.89
	Jaccard	3.54 ± 1.97	14.64 ± 3.66	19.57 ± 5.79
	UniFrac	1.91 ± 1.02	11.59 ± 4.54	18.48 ± 2.48
	W-UniFrac	9.94 ± 6.33	12.90 ± 5.55	20.00 ± 10.35

Table AB.2.S2. CTF improves over existing methods across all time and increases the number of significant time points.
Comparison of KNN-classification and PERMANOVA quantitative benchmarking between CTF and existing methods for DIABIMMUNE and ECAM datasets. AUPR mean ± s.d. across all time points for mean 100-fold cross-validation at each time point. PERMANOVA F-statistic fold-increase mean ± s.e. across all time points.

	AUPR		PERMANOVA F-statistic CTF Fold-Increase	
Comparison Method	**DIABIMMUNE**	**ECAM**	**DIABIMMUNE**	**ECAM**
CTF	0.983 ± 0.001	0.768 ± 0.007	1.0 ± 0.0	1.0 ± 0.0
Aitchison	0.885 ± 0.003	0.552 ± 0.004	6.13 ± 0.39	8.11 ± 1.17
Bray-Curtis	0.87 ± 0.002	0.589 ± 0.006	5.00 ± 0.24	8.88 ± 2.53
Jaccard	0.88 ± 0.002	0.592 ± 0.006	6.40 ± 0.48	8.66 ± 1.05
UniFrac	0.874 ± 0.001	0.552 ± 0.005	5.32 ± 0.22	7.79 ± 1.09
W-UniFrac	0.864 ± 0.003	0.582 ± 0.007	3.94 ± 0.45	10.41 ± 4.94

Table AB.2.S3. Linear mixed-effects model results on birth mode associated log-ratios is significant by birth mode for both ECAM and DIABIMMUNE.

		Intercept	birth-mode	month	birth-mode:month	Group Var
DIABIMMUNE	**Coef.**	-2.491	6.362	0.306	-0.204	1.791
	Std.Err.	0.785	0.832	0.023	0.025	0.215
	z	-3.173	7.644	13.306	-8.259	-
	P>|z|	0.002	<.001	<.001	<.001	-
	[0.025	-4.03	4.731	0.261	-0.252	-
	0.975]	-0.952	7.993	0.351	-0.156	-
ECAM	**Coef.**	-4.362	2.097	0.16	0.067	3.279
	Std.Err.	0.483	0.641	0.025	0.032	0.335
	z	-9.037	3.272	6.395	2.131	-
	P>|z|	0	0.001	0	0.033	-
	[0.025	-5.308	0.841	0.111	0.005	-
	0.975]	-3.416	3.353	0.209	0.129	-

AB.3. Supplemental References

1. Lek-Heng Lim. Singular values and eigenvalues of tensors: a variational approach. in 1st IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, 2005. 129–132 (2005).

2. Anandkumar, A., Ge, R. & Janzamin, M. Guaranteed Non-Orthogonal Tensor Decomposition via Alternating Rank-1 Updates. arXiv [cs.LG] (2014).

3. Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M. & Telgarsky, M. Tensor Decompositions for Learning Latent Variable Models (A Survey for ALT). Lecture Notes in Computer Science 19–38 (2015) doi:10.1007/978-3-319-24486-0_2.

4. Jain, P. & Oh, S. Provable Tensor Factorization with Missing Data. in Advances in Neural Information Processing Systems 27 (eds. Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D. & Weinberger, K. Q.) 1431–1439 (Curran Associates, Inc., 2014).

Appendix C. Supplemental Information for Naturalization of the microbiota developmental trajectory of Cesarean-born neonates after vaginal seeding

AC.1. Supplementary Methods

Supplemental Methods 4.S1-S16 can be found at http://dx.doi.org/10.1016/j.medj.2021.05.003

AC.2. Supplemental Figures

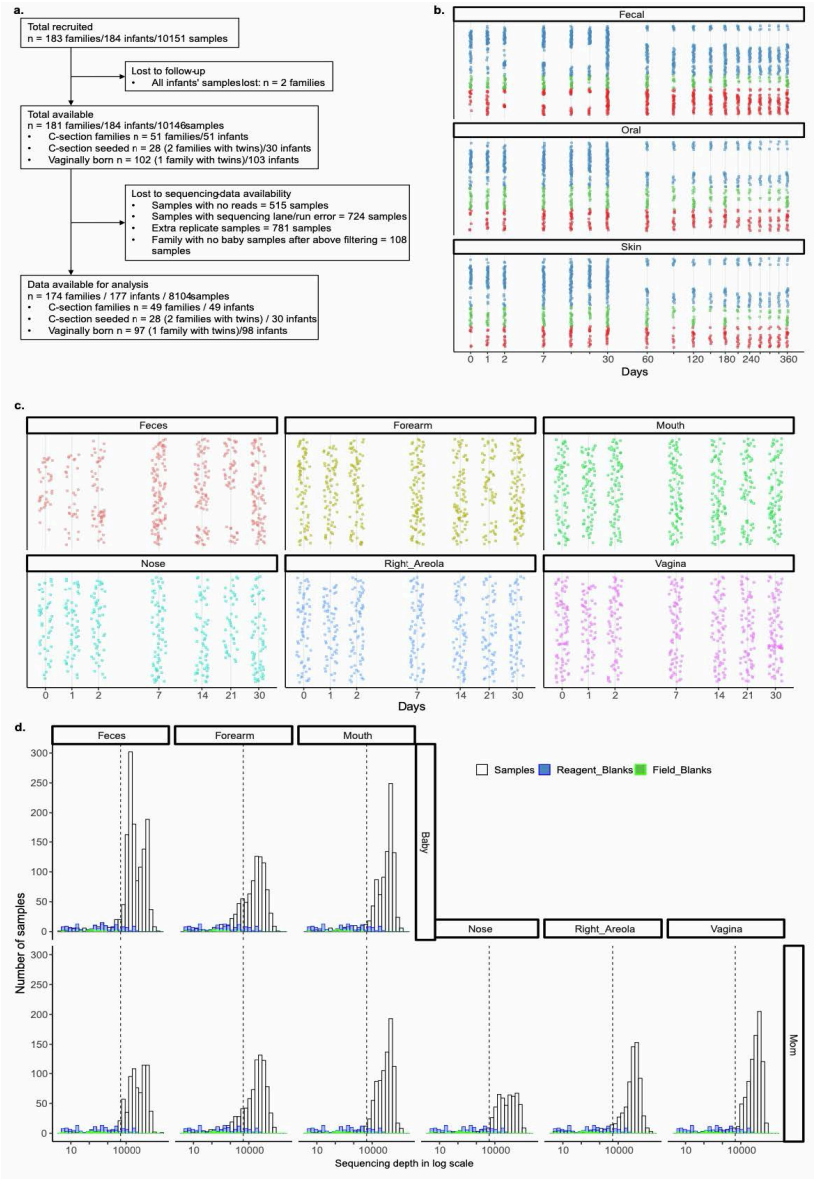

Figure AC.2.S1. Longitudinal sampling of mother-infant pairs. (a) Number of families, infants and samples from the current study. (b) Longitudinal sampling of infant samples by birth modes and body sites. Sampling with sterile swabs in different body sites took place within the first hours after birth in all babies (including the vaginal gauze exposed CS group, who were sampled after the gauze swabbing procedure), then at day 1-3, weekly for the first month and monthly up to the first year. Each row along y-axis is an individual baby. Each point represent a sample for a baby. The points are colored by birth modes, vaginal (blue), cesarean-seeded (green), and cesarean (red). On average, each baby contributed 18, 17, and 21 samples (across three body sites and multiple time points for the first year) for vaginal, cesarean, and cesarean-seed groups. (c) Longitudinal sampling of maternal samples by body sites within the first month after delivery. Each row along y-axis is an individual mom. On average each mom contributed 17 samples (across six body sites and multiple time points for the first month). (d) Distribution of number of reads per sample by different body sites in moms or babies Reagent blanks (blue), and field blanks (green) presentation were overplayed on each panel, and show much lower depth than the samples, indicating good overall quality of the sequences and lack of contamination. Dashed line marked the 5000 reads per sample position.

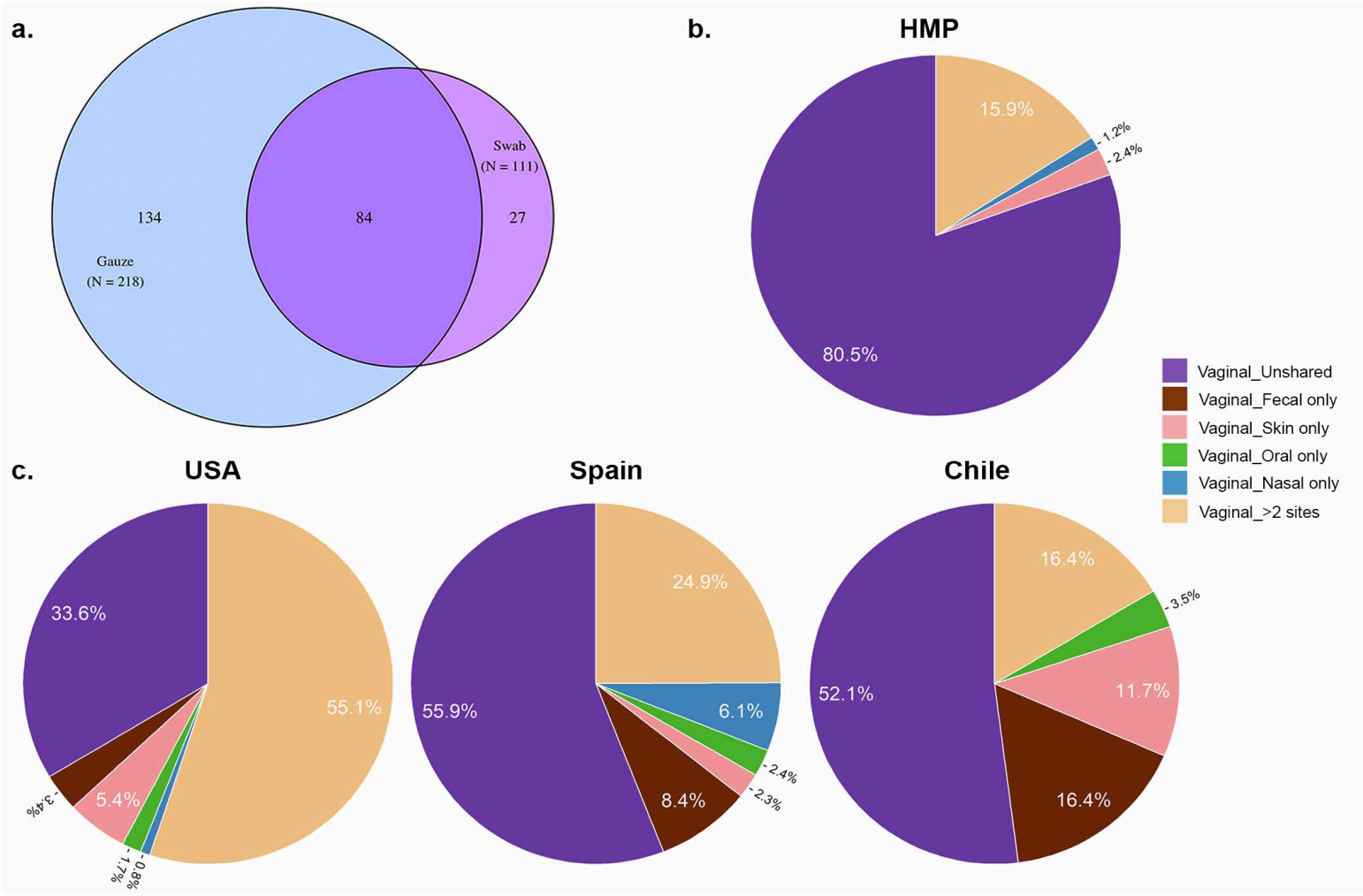

Figure AC.2.S2. Pluripotential nature of perinatal vaginal microbiome.(a) Number of ASVs shared between vaginal swabs and vaginal gauzes. Gauzes show higher ASVs richness than vaginal swabs. Proportions of bacterial vaginal ASVs shared with other body sites. (b) in HMP data of non-pregnant women (105 women), (c) in parturient mothers at the day of delivery from USA (53 mothers), Spain (24 mothers) and Chile (20 mothers). HMP data was reprocessed by extracting V4 sequences and analyzed using the same pipeline.

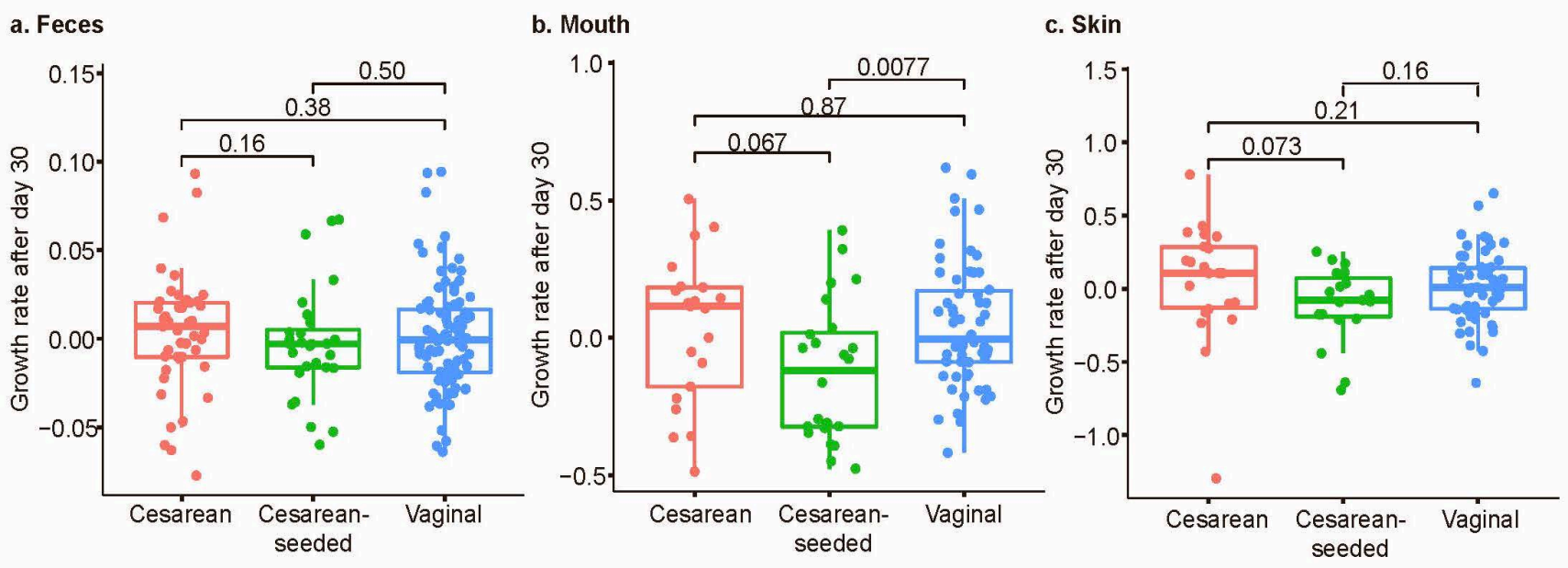

Figure AC.2.S3. Bayesian Sparce Functional PCA (SFPCA) analyses on Shannon alpha diversity from 1 to 12 months of age. Bayesian Sparse Functional Principal Components Analysis (SFPCA) performed on Shannon alpha diversity across time did not differ by birth mode using Wilcoxon rank-sum test. The rate of growth of the Shannon diversity after day 30 (y-axes) is shown across birth modes (x-axis) for fecal (left), oral (middle), and skin (right) samples.

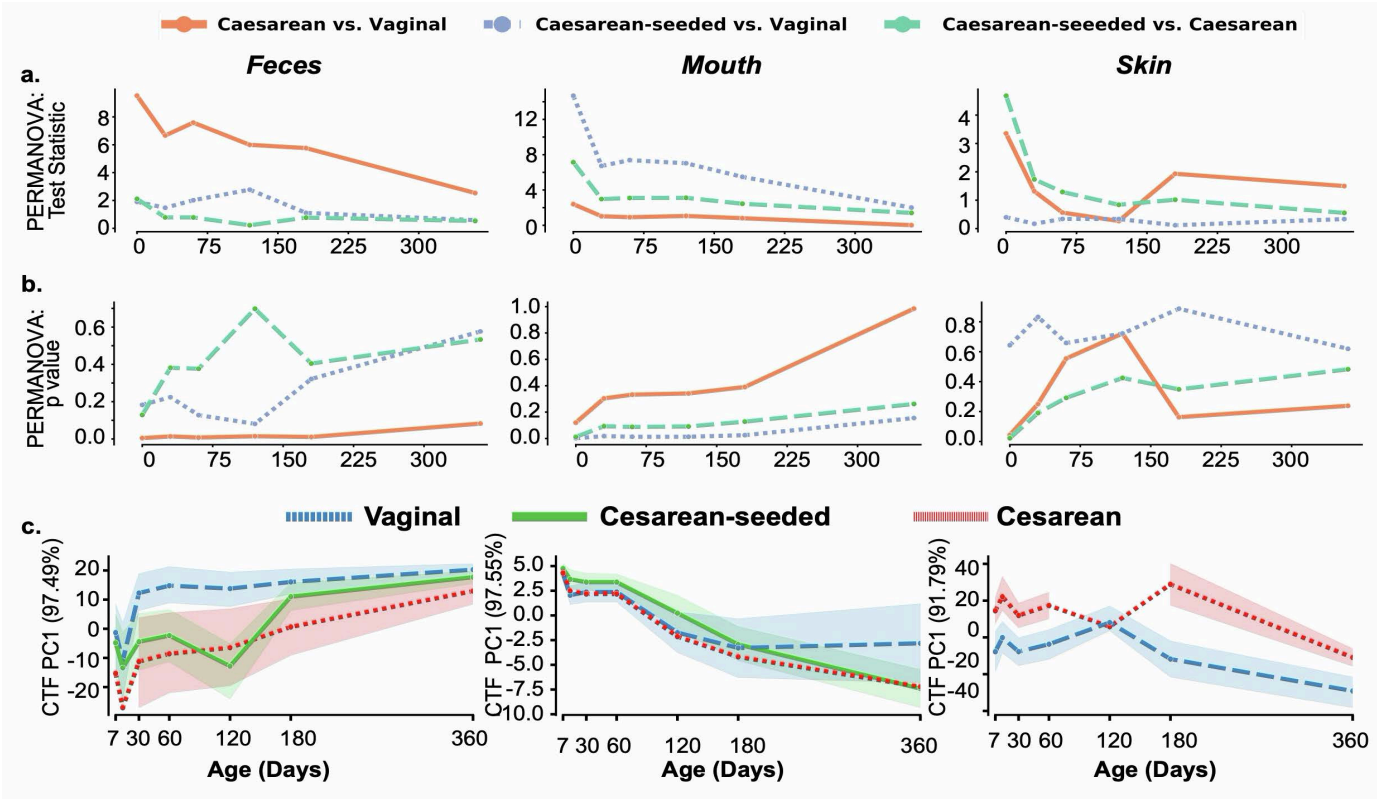

Figure AC.2.S4. Compositional Tensor Factorization identifies the partial restoration of microbiome among cesarean-seeded babies. PERMANOVA of Aitchison distances from Compositional Tensor Factorization (CTF) during the first year of life. (a) PERMANOVA test statistics and (b) Bonferroni corrected p-values are plotted across age in days (x-axes). PERMANOVA plots are colored by compared pairs, Caesarean vs. Vaginal, Caesarean-seeded vs. Vaginal, and Caesarean-seeded vs. Caesarean. (c) Compositional Tensor Factorization (CTF) in the USA cohort. CTF ordination plot as in Figure 4.1a but only with the 101 US infants shows the same trends as the whole dataset, with vaginally born and seeded babies clustering together and separately from Cesarean-born infants. Comparison of Vaginal (blue; n=62), Cesarean (red; n=23), Cesarean-seeded (green; n=16) with CTF first principal component (y-axes) of infant samples over age in days (x-axes); error bars show the standard error of the mean. There were not enough sequences after filtering the samples in the skin of Cesarean-seeded babies. Related to Figure 4.1, 4.2, 4.3.