UNIVERSITY OF CALIFORNIA SAN DIEGO

Elucidating Immune and Inflammatory Diseases on the Atomic and Microbiome Scales

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

in

Biomedical Sciences

by

Bryn Colleen Taylor

Committee in charge:

Professor Rommie E. Amaro, Co-Chair
Professor Robin Knight, Co-Chair
Professor Laurence Brunton
Professor Tracy Handel
Professor Larry Smarr

2020

The dissertation of Bryn Colleen Taylor is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

____

____

____

____

Co-Chair

____

Co-Chair

University of California San Diego

2020

DEDICATION

To my Mom: for cultivating my creativity and my courage, and for being my best friend.

To my Dad: whose drive and passion has molded me into the scientist and person I am today.

To my Grandma Jo: for always listening to me. You are my friend in my pocket.

To Chris: for being my inspiration and my motivation.

EPIGRAPH

Science makes people reach selflessly for truth and objectivity;

it teaches people to accept reality, with wonder and admiration.

—Dr. Lise Meitner

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This dissertation is the culmination of my graduate work and the product of an incredible support system. Here, I hope to convey my appreciation for those who have helped my dreams become a reality.

I would first like to acknowledge my advisors, Profs. Rob Knight and Rommie Amaro. I consider myself lucky to have overwhelmingly enjoyed my time earning my doctorate, and I owe that to you. My unusual cross-disiplinary graduate work would never have been possible without your limitless encouragement and flexibility. Thank you for providing me with the opportunity and resources to pursue my diverse interests. My committee members Profs. Tracy Handel and Larry Smarr also had a fundamental impact on my research and I will be forever thankful for their advice and critical insights.

Prof. Larry Brunton is the reason I am a "biomedical" scientist. Thank you for introducing me to my graduate program and, perhaps more importantly, a number of different tequilas. I am grateful that you are my mentor and my friend.

The work in this dissertation would be quite sparse without my coworkers in the Knight and Amaro labs. Together we talked through wild ideas and worked out problems on white boards; we traveled to conferences, ate lots of snacks, and drank a few happy hour beers. My life in graduate school was richer thanks to all of you. Thanks also to my network of friends I made during my graduate tenure. From you, I learned about more than just science, and I am the better for it.

I would not be where I am today without the overwhelming support from the communities who have raised me. To my "aunts" and "uncles", my Los Gatos and Saratoga "moms", my close family-friends, my Grandma Jo, my Carp team, and my best friends throughout the years: thank you for being my pillars. Special thanks my boyfriend Chris and his family (Serena, Tsengdar, and Gloria), for their unwavering personal and professional support.

Finally, I want to thank my parents, Tanya and Jim Taylor. Their abundance of love and the absolute force of their belief in my capabilities is the reason this dissertation exists at all. This is as much their achievement as it is my own.

Chapter 2, in full, is a modified reprint of the material as it appears in "B. C. Taylor, C. T. Lee, and R. E. Amaro. *Structural Basis for Ligand Modulation of the CCR2 Conformational Landscape*.

Proceedings of the National Academy of Sciences, U.S.A. (2019). DOI: 10.1073/pnas.1814131116." The dissertation author was the primary investigator and author of this work. The dissertation author designed the research, performed and analyzed the molecular dynamics simulations, built and analyzed the Markov state models, designed and created the figures, and wrote the manuscript.

Chapter 3, in full, is a modified reprint of the material as it appears in "R. Knight[†], A. Vrbanac[†], B. C. Taylor[†], A. Aksenov, C. Callewaert, J. Debelius, A. Gonzalez, T. Kosciolek, L. McCall, D. McDonald, A. V. Melnik, J. T. Morton, J. Navas, R. A. Quinn, J. G. Sanders, A. D. Swafford, L. R. Thompson, A. Tripathi, Z. Z. Xu, J. R. Zaneveld, Q. Zhu, J. G. Caporaso, and P.C. Dorrestein. *Best Practices for Analysing Microbiomes*. Nature Reviews Microbiology (2018). DOI: 10.1038/s41579-018-0029-9". The dissertation author was a primary coinvestigator and author of this work. With the three co-first authors, the dissertation author conceptualized the scope of the manuscript, researched the data for the manuscript, designed and created the figures, and wrote the manuscript.

Chapter 4, in full, is a modified reprint of the material as it appears in "M. Yazdani, B. C. Taylor, L. Weizhong, J. W. Debelius, R. Knight, and L. Smarr. *Using Machine Learning to Identify Major Shifts in Human Gut Microbiome Protein Family Abundance in Disease*. IEEE Big Data (2016). DOI: 10.1109/BigData.2016.7840731." The dissertation author was an investigator and coauthor of this work. The dissertation author contributed to research design, analysis of the Random Forest and natural language classifiers, biological interpretation of the classifier results, figure creation, and manuscript writing.

Chapter 5, in full, will be submitted for publication and may appear as "B. C. Taylor, K. C. Weldon, T. Groth, R. Ellis, E. Gentry, A. Tripathi, D. McDonald, G. Humphrey, M. Bryant, J. Toronczak, T. Schwartz, M. F. Oliveira, R. Heaton, S. Gianella, A. D. Swafford, P. C. Dorrestein, and R. Knight. *Depression in HIV and HCV Co-Infected Individuals is Associated with Systematic Differences in the Gut Microbiome and Metabolome*." The dissertation author was the primary investigator and author of this work, contributing substantially to experimental design. Sample collection, 16S rRNA extraction, and sequencing/mass spectrometric analysis were performed by others. The dissertation author computationally processed the 16S rRNA sequence data, designed the analysis of the 16S rRNA sequence data and mass spectrometric data, interpreted the results of the analyses, designed and assembled the figures, performed the statistical analyses, and wrote the manuscript.

Chapter 6, in full, is a modified reprint of the material as it appears in "B. C. Taylor, F. Lejzerowicz, M. Poirel, J. Schaffer, L. Jiang, A. Aksenov, G. Humphrey, C. Martino, S. Miller-Montgomery, P. Dorrestein, P. Veiga, S. J. Song, D. McDonald, M. Derrien, and R. Knight. *Consumption of Fermented Foods Is Associated with Systematic Differences in the Gut Microbiome and Metabolome*. mSystems (2020). DOI: 10.1128/mSystems.00901-19." The dissertation author was the primary investigator and author of this work, contributing substantially to experimental design. Sample collection, 16S rRNA extraction, and sequencing/mass spectrometric analysis were performed by others. The dissertation author computationally processed the 16S rRNA sequence data and shotgun metagenomics data; designed the analyses for and interpreted the 16S rRNA sequence data, shotgun metagenomics data, and mass spectrometric data; designed and assembled the figures; performed the statistical analyses; and wrote the manuscript.

† denotes joint authorship.

VITA

2012-2014	Undergraduate Research Fellow Mayfield Lab, Department of Biology, University of California San Diego
2014	Bachelor of Science in General Biology University of California San Diego
2015-2020	Graduate Research Fellow Knight and Amaro Labs, University of California San Diego
2020	Doctor of Philosophy in Biomedical Sciences University of California San Diego

PUBLICATIONS

† denotes joint authorship.

C. E. Jung, **B. C. Taylor**, J. Shin, K. Ferrante, E. Wasenda, Q. Lippmann, R. Knight, D. Pride, and E. S. Lukacz. “Impact of Vaginal Estrogen on the Urobiome in Postmenopausal Women with Recurrent Urinary Tract Infection”. In Preparation.

B. C. Taylor, K. C. Weldon, T. Groth, R. Ellis, E. Gentry, A. Tripathi, D. McDonald, G. Humphrey, M. Bryant, J. Toronczak, T. Schwartz, M. F. Oliveira, R. Heaton, S. Gianella, A. D. Swafford, P. C. Dorrestein, and R. Knight. “Depression in HIV and HCV Co-Infected Individuals is Associated with Systematic Differences in the Gut Microbiome and Metabolome”. In Preparation.

T. Hempel, M. J. del Razo†, C. T. Lee†, **B. C. Taylor**†, R. E. Amaro, and F. Noe. “De-composing Markov models: Practical Perspectives on Modeling Biomolecular Systems”. In Preparation.

B. C. Taylor, F. Lejzerowicz, M. Poirel, J. Schaffer, L. Jiang, A. Aksenov, G. Humphrey, C. Martino, S. Miller-Montgomery, P. Dorrestein, P. Veiga, S. J. Song, D. McDonald, M. Derrien, and R. Knight. Consumption of Fermented Foods Is Associated with Systematic Differences in the Gut Microbiome and Metabolome. mSystems (2020). DOI: 10.1128/mSystems.00901-19.

V. Gligorijecvi, P. D. Renfrew, T. Kosciolek, J. K. Leman, K. Cho, T. Vatanen, D. Berenberg, **B. C. Taylor**, I. M. Fisk, R. J. Xavier, R. Knight, and R. Bonneau. “Structure-Based Function Prediction Using Graph Convolutional Networks”. bioRxiv (2019). DOI: https://doi.org/10.1101/786236.

G. Sharon, N. J. Cruz, D.-W. Kang, M. J. Gandal, B. Wang, Y.-M. Kim, E. M. Zink, C. P. Casey, **B. C. Taylor**, C. J. Lane, L. M. Bramer, N. G. Isern, D. W. Hoyt, C. Noecker, M. J. Sweredoski, A. Moradian, E. Borenstein, J. K. Jansson, R. Knight, T. O. Metz, C. Lois, D. H. Geschwind, R. Krajmalnik-Brown, and S. K. Mazmanian. “Human Microbiomes from Autism Spectrum Disorder promote Behavioral Symptoms in Mice”. Cell (2019). DOI: https://doi.org/10.1016/j.cell.2019.05.004.

B. C. Taylor, C. T. Lee, and R. E. Amaro. “Structural Basis for Ligand Modulation of the CCR2 Conformational Landscape”. Proceedings of the National Academy of Sciences, U.S.A. (2019). DOI: 10.1073/pnas.1814131116.

R. Knight[†], A. Vrbanac[†], **B. C. Taylor**[†], A. Aksenov, C. Callewaert, J. Debelius, A. Gonzalez, T. Kosciolek, L. McCall, D. McDonald, A. V. Melnik, J. T. Morton, J. Navas, R. A. Quinn, J. G. Sanders, A. D. Swafford, L. R. Thompson, A. Tripathi, Z. Z. Xu, J. R. Zaneveld, Q. Zhu, J. G. Caporaso, and P.C. Dorrestein. "Best Practices for Analysing Microbiomes". Nature Reviews Microbiology (2018). DOI: 10.1038/s41579-018-0029-9.

M. Yazdani, **B. C. Taylor**, L. Weizhong, J. W. Debelius, R. Knight, and L. Smarr. "Using Machine Learning to Identify Major Shifts in Human Gut Microbiome Protein Family Abundance in Disease". IEEE Big Data (2016). DOI: 10.1109/BigData.2016.7840731.

R. A. Abagyan and **B. C. Taylor**. "Toward Complete Cellular Pocketomes and Predictive Polypharmacology". In Silico Drug Discovery and Design: Theory, Methods, Challenges, and Applications. Ed. by C. Cavasotto. Oxford: Taylor & Francis, 2015. Chap. 10, pp. 266290. DOI: 10.1201/b18799.

ABSTRACT OF THE DISSERTATION

Elucidating Immune and Inflammatory Diseases on the Atomic and Microbiome Scales

by

Bryn Colleen Taylor

Doctor of Philosophy in Biomedical Sciences

University of California San Diego, 2020

Professor Rommie E. Amaro, Co-Chair
Professor Robin Knight, Co-Chair

The driving purpose of biomedical science is to understand the molecular mechanisms of disease progression and design tools to enable disruption. In this dissertation, I leverage two disparate scientific fields – computational chemistry and the microbiome – to explore the environments that drugs interact with, with an eye toward improving therapeutic development prospects. On the atomic scale, I present our work using molecular dynamics simulations and Markov-state models to characterize the effect of small molecule inhibitors on CCR2, a critical protein target for the treatment of a number of immune or inflammatory diseases. On the microbiome scale, I present several studies investigating the interplay between health, immune and inflammatory diseases, diet, and the community of microbes that inhabit the human gut.

Chapter 1

Introduction

Each year, millions of Americans are affected by diseases caused by aberrant inflammation or immune system dysfunction. In the United States, trillions of dollars are spent each year on the research, management, and treatment of such diseases that include cancer, HIV/AIDS, diabetes, and inflammatory bowel disease (IBD). Recent advancements in treatment have yielded small molecule drugs that bind to a target protein to modulate its function. However, drug development is a long and costly process. Candidate drugs can fail at the final stage of clinical trials for unknown reasons, consuming an average of 10 years of development time and $2.6 billion per drug. To interrogate why drugs fail and gain insight into how we can develop better drugs, the research presented in this dissertation focuses on drug environments at two different scales. On the atomic scale, I use computational chemistry and mathematical models to explore how drugs perturb the dynamics of their target protein (Chapter 2). On the microbiome scale, I study cohorts of patients with these diseases to characterize their effect on the communities of bacteria that inhabit the gastrointestinal tract, also known as the gut microbiome (Chapter 3 to Chapter 6). A significant modulator of both health and disease, the gut microbiome has also been previously shown to affect drug efficacy. This research aims to understand these critical drug-environment interactions in order to leverage this knowledge for the long-term goal of reducing the cost of drug development.

Understanding CCR2 inhibitor attrition and improving future prospects.

One component of the immune system which has been identified as a potentially critical drug target is the protein CC chemokine receptor 2 (CCR2). CCR2 promotes the metastasis of cancer cells and is also implicated in autoimmunity-driven type-1 diabetes, multiple sclerosis, asthma, atherosclerosis, neuropathic pain, and rheumatoid arthritis. Although promising, drugs that prevent the function of CCR2 have been unsuccessful. In Chapter 2 of my dissertation, I investigate the effect of two drugs on CCR2 dynamics by coupling long-timescale molecular dynamics simulations with Markov-state model theory. These methods act as a "computational microscope" that allows us to see how the drugs affect the protein's movement over time. I present preliminary evidence for why these drugs failed and the discovery of a cryptic drug-binding pocket which may be amenable to targeting with a third drug to improve protein inhibition.

The effect of the microbiome on health, disease progression, and therapeutic strategies.

A new paradigm in drug development is to examine not only the direct interaction between the drug and the patient, but also between the drug and the patient's microbiome. The human gut microbiome contains about 3.8 x 10^{13} microorganisms with which an orally-dosed drug must interact with prior to reaching its target. In Chapter 3, I present a review on best practices for analysing the microbiome.

In Chapter 4, I present our work on classifying major changes in microbiome protein family abundances between healthy individuals and IBD patients. IBD is an autoimmune condition that is associated with major alterations in the gut microbiome. In this work, we develop a machine learning pipeline to identify which proteins from the microbiome best separate between states of health and disease. As the field expands our analysis from 10,000 protein families to millions of proteins identified in the gut microbiome, scalable methods for quickly identifying such anomalies between health and disease states will be increasingly valuable for biological interpretation of sequence data and development of targeted therapeutics.

For the more than 1.1 million people who are living with HIV/AIDS and the 71 million people living with Hepatitis C (HCV), the interaction between orally dosed drugs and their environment is of critical importance to their quality of life. Despite suppressive combination antiretroviral therapy, HIV-infected individuals have persistent gut barrier dysfunction ('leaky gut'), gut microbiome dysbiosis, HIV-associated neurocognitive disorders, increased risk for major depressive disorders, and other behavioral impairments. Together, dysbiosis and leaky gut render HIV-monoinfected individuals more vulnerable to microbial antigen-driven effects on the central nervous system via pro-inflammatory bacterial antigens. The gut microbiota also may affect blood-brain barrier integrity. In Chapter 5, we explore the relationship between the gut microbiome and depression in HIV-mono-infected, HIV/HCV-coinfected, and uninfected individuals. We find that the gut microbiome and metabolome is altered in coinfected individuals who have suffered a major depressive disorder. These findings are of clinical importance, with implications for leveraging existing interventions that can restore normal gut flora and barrier integrity and may have the potential to improve central nervous system function and patient outcomes.

As we move toward understanding how the microbiome is associated in particular disease states, the next step is to understand how we can modulate the microbiome to address dysbiosis and potentially mitigate disease. One method of modulating the gut microbiome is by adding live microbes to our diet by consuming fermented foods. Fermented foods are historically and culturally significant, but limited studies have explored the association between fermented food consumption and the gut microbiome in large cohorts. In Chapter 6, we used a combination of three omics-based analyses to study the relationship between the microbiome and fermented food consumption in thousands of people using both cross-sectional and longitudinal data. We find that fermented food consumers have subtle differences in their gut microbiota structure, which is enriched in conjugated linoleic acid, thought to be beneficial. The results motivate further studies on how we can utilize specific kinds of fermented food as therapeutic strategies to impact the microbiome and human health.

In summary, the work in this dissertation aims to better understand the environments that drugs interact with in order to provide insight into drug attrition rates and improve upon future prospects of drug development and therapeutic interventions.

Chapter 2

Structural Basis for Ligand Modulation of the CCR2 Conformational Landscape

2.1 Abstract

CC Chemokine Receptor 2 (CCR2) is a part of the chemokine receptor family, an important class of therapeutic targets. These class A G-protein coupled receptors (GPCRs) are involved in mammalian signaling pathways and control cell migration toward endogenous CC chemokine ligands, named for the adjacent cysteine motif on their N-terminus. Chemokine receptors and their associated ligands are involved in a wide range of diseases and thus have become important drug targets. CCR2, in particular, promotes the metastasis of cancer cells and is also implicated in autoimmunity driven type-1 diabetes, diabetic nephropathy, multiple sclerosis, asthma, atherosclerosis, neuropathic pain, and rheumatoid arthritis. Although promising, CCR2 antagonists have been largely unsuccessful to date. Here, we investigate the effect of an orthosteric and an allosteric antagonist on CCR2 dynamics by coupling long timescale molecular dynamics simulations with Markov-state model theory. We find that the antagonists shift CCR2 into several stable inactive conformations that are distinct from the crystal structure conformation and disrupt a continuous internal water and sodium ion pathway, preventing transitions to an active-like state. Several metastable conformations present a cryptic drug binding pocket near the allosteric site that may be

amenable to targeting with small molecules. Without antagonists, the apo dynamics reveal intermediate conformations along the activation pathway that provide insight into the basal dynamics of CCR2, and may also be useful for future drug design.

2.2 Introduction

The signaling axis of CCR2 and its endogenous ligand, CCL2, is a notable therapeutic target due to its association with numerous diseases, including cancer, autoimmunity driven type-1 diabetes, diabetic nephropathy, multiple sclerosis, asthma, atherosclerosis, neuropathic pain, and rheumatoid arthritis[1–3]. Despite much effort that has been devoted to clinical and pre-clinical trials, a successful antagonist has yet to be developed[4–7]. Prior to a full-length crystal structure of CCR2, several studies used homology modeling and docking to gain insights into the structure and dynamics of the protein and its associated ligands or small molecule drugs[8–10]. However, recently CCR2 was crystallized for the first time[11], opening up new opportunities for rational drug design.

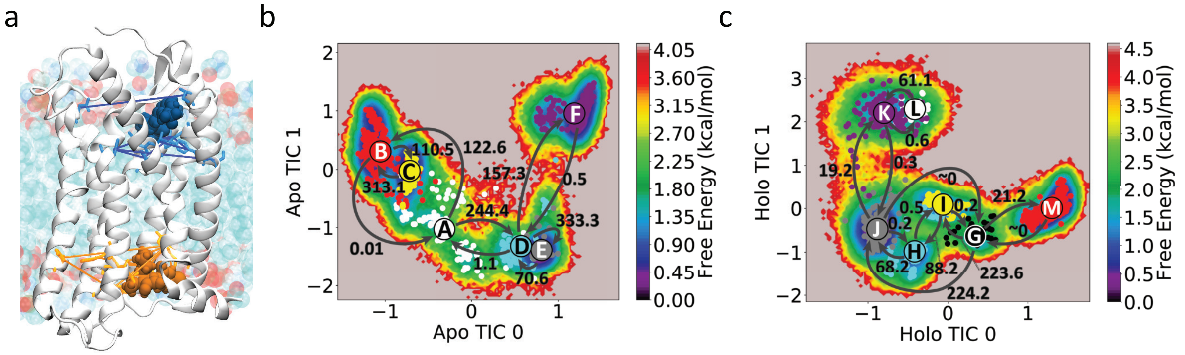


Figure 2.1: MD simulations of CCR2 in a lipid bilayer were performed apo and holo CCR2. A) Sets of residue pairs surrounding the two ligand binding pockets were used with TICA (SI Appendix). The protein is shown in white cartoon. Lipids are teal, red, and blue. The orthosteric and allosteric ligands are shown in blue and orange, respectively, with inter-residue pair distances denoted by similarly colored lines. The free energy and Maximum-Likelihood HMMs of B) apo and C) holo CCR2, projected onto the first two TICA components. Coarse-grained states are labeled and colored. Transition rates between macrostates are represented by arrows reported in units of ms^{-1}.

As with most GPCRs, chemokine receptors transmit signals across cell membranes by means of extracellular ligand and intracellular G-protein binding. Distinct conformational states of the receptor are necessary for chemokine/ligand binding, G-protein binding, activation, inactivation, and signal transmission

[12–14]. GPCRs are no longer considered to be simple on/off molecular switches – instead, they assume a wide range of conformational states, including ligand-specific states, intermediate states, and states that allow for basal (apo) signaling without ligands bound [13, 15–21]. Ligands and small molecule drugs may shift the equilibrium of the receptor's conformational states towards particular states. Effective small molecule antagonists that inhibit CCR2 signaling, potentially by shifting the receptor equilibrium toward inactive conformational states, are desired for treatment of diseases that involve the CCR2/CCL2 axis. Key challenges are to characterize the basal dynamics of CCR2 and to understand how current antagonistic small molecule drugs modulate these dynamics. While crystal structures provide valuable snapshots of proteins and protein complexes, they lack the ability to reveal dynamics at the atomic level. Starting with the newly resolved crystal structure of CCR2 (PDB ID: 5T1A) we performed multi-microsecond all-atom explicitly solvated molecular dynamics (MD) simulations of the receptor in a lipid bilayer in unbound (apo) and dual-antagonist-bound (holo) states (Fig. 2.1). The two antagonists were co-crystallized with CCR2: the orthosteric antagonist, BMS-681, and the allosteric antagonist, CCR2-RA-[R]. Each system was simulated in triplicate on Anton 2 [22, 23] for a total of 260 microseconds (SI Appendix, Table S1, Fig. 2.5).

While long timescale (tens of microseconds) simulations are useful for analyzing sequential conformational changes, simulations are generally unable to directly probe timescales of biological interest (milliseconds - seconds)[24]. One way to bridge this timescale gap is to couple MD simulations with Markov state model theory[25–33] (MSM, described in SI Materials and Methods). Integrating MD simulations with MSMs allowed us to extend the reach of simulated timescales, and identify key differences in the conformational ensembles and dominant slow motions of apo and holo CCR2 (Fig. 2.1). We find that the antagonists disrupt a continuous internal water and sodium ion pathway preventing transitions to an active-like state, and shift CCR2 into several stable states that are distinct from the crystal structure conformation, three of which present a cryptic druggable pocket. Without antagonists, intermediate conformations with active-state conformational signatures shed light on the apo dynamics of CCR2 and may also be useful for future drug design.

2.3 Results and Discussion

To compare the conformational landscapes of apo and holo CCR2 we ran all-atom MD simulations totalling 260 μs on Anton 2[22]. For one MD replicate of the holo system, we observed the orthosteric drug dissociate from the protein. Analyzing the conformations before and after ligand dissociation yields a first glimpse of the allosteric effect of the remaining antagonist on the protein dynamics, and provides a starting point for future rounds of adaptive sampling to obtain robust dissociation statistics (not pursued here). In order to extend the analysis beyond a dissociation event and connect to longer timescale phenomena, MSMs were constructed from the trajectories (Fig. 2.1B,C). The variational approach for conformation dynamics[34], specifically Time-structure-based independent component analysis (TICA)[35, 36] was used to perform dimensionality reduction for the MSMs and identify the features and collective variables (time-structure based independent components (TICs)) that best represent the dominant slow motions. The MSMs create human interpretable models that we use to interrogate the conformational and kinetic differences between the two ensembles to derive new understandings about the mechanisms underlying effects of CCR2 antagonism. Further methodological details are provided in Methods and in the SI Appendix.

2.3.1 Comparison of the CCR2 conformational ensemble with other class A GPCRs

We compare representative states from the apo and holo conformational ensemble with other class A GPCRs to establish similarities within the class. We find that the states of apo CCR2 have conformational signatures found in the active or intermediately-active states of GPCRs, suggesting that these states are on a pathway toward activation. Holo CCR2 diverges from the crystal structure to form distinct states that expose putative drug binding pockets and reveal the effect of antagonists on receptor dynamics. The most populated holo macrostate, J, is not representative of the crystal structure as it deviates 10.8 from the crystal structure conformation (Fig. 2.1C).

We evaluate the metastable states by comparing helical conformational signatures and conserved groups of structurally neighboring amino acids called 'microswitches'. These include NPxxY (Tyr $305^{7.53}$), DRY (Arg $138^{3.50}$), Tyr $222^{5.58}$, sets of residues in the orthesteric and allosteric binding sites, and the chemokine and G-protein binding pockets to the inactive crystal structure of CCR2 that we used in this study (PDB ID: 5T1A), to an intermediately-active crystal structure of a class A GPCR, A_{2A}AR(PDB ID: 2YDO[37]; 25% sequence identity to CCR2), the active crystal structure of a class A GPCR, US28 (PDB ID: 4XT3[38]; 30% sequence identity to CCR2), and three other chemokine receptors: CCR5 (PDB ID: 4MBS[39], CCR9 (PDB ID: 5LWE[40]), and CXCR4 (PDB ID: 4RWS[41]). Signatures of an active GPCR state include: 1) the inward shift of the intracellular part of helix VII toward the helical bundle, 2) the outward shift of the intracellular part of helix VI in concert with helix V, 3) the upward shift and lateral movement of helix III, and 4) the rearrangements of conserved microswitches [15]. According to these metrics, the starting crystal structure of CCR2 is in an inactive conformation[11], the crystal structure of A_{2A}AR is an an intermediately-active conformation, and the crystal structure of US28 is in the active conformation.

1) Apo macrostates show an active-like inward shift of the intracellular part of helix VII toward the helical bundle

All of the apo macrostates exhibit an active state hallmark (Fig. 2.2A): the intracellular end of helix VII tilts slightly inward toward the center of the helical bundle. More prominently, the extracellular end of helix VII tilts outward, resembling the active conformation of US28. The holo macrostates show the opposite: the intracellular end of helix VII tilts slightly outward and the extracellular end of helix VII tilts inward, remaining in the crystal structure conformation.

2) Holo macrostates, not apo, show an active-like outward shift of the intracellular part of helix VI in concert with helix V

Helix V and VI in the apo macrostates are not in an active conformation. Instead, it is the holo macrostates that have the intracellular end of helix V and VI tilting outward to resemble the active conformation (Fig. 2.2C,D), suggesting that neither apo nor holo macrostates are in an exclusively inactive or active conformational state, despite starting from a particularly inactive crystal structure. Due to this outward motion of helix VI, holo macrostates K and L exhibit a more open G-protein binding site compared to holo macrostate G which is more closed (Fig. 2.1B,C). The RMSF of the allosteric ligand is larger in

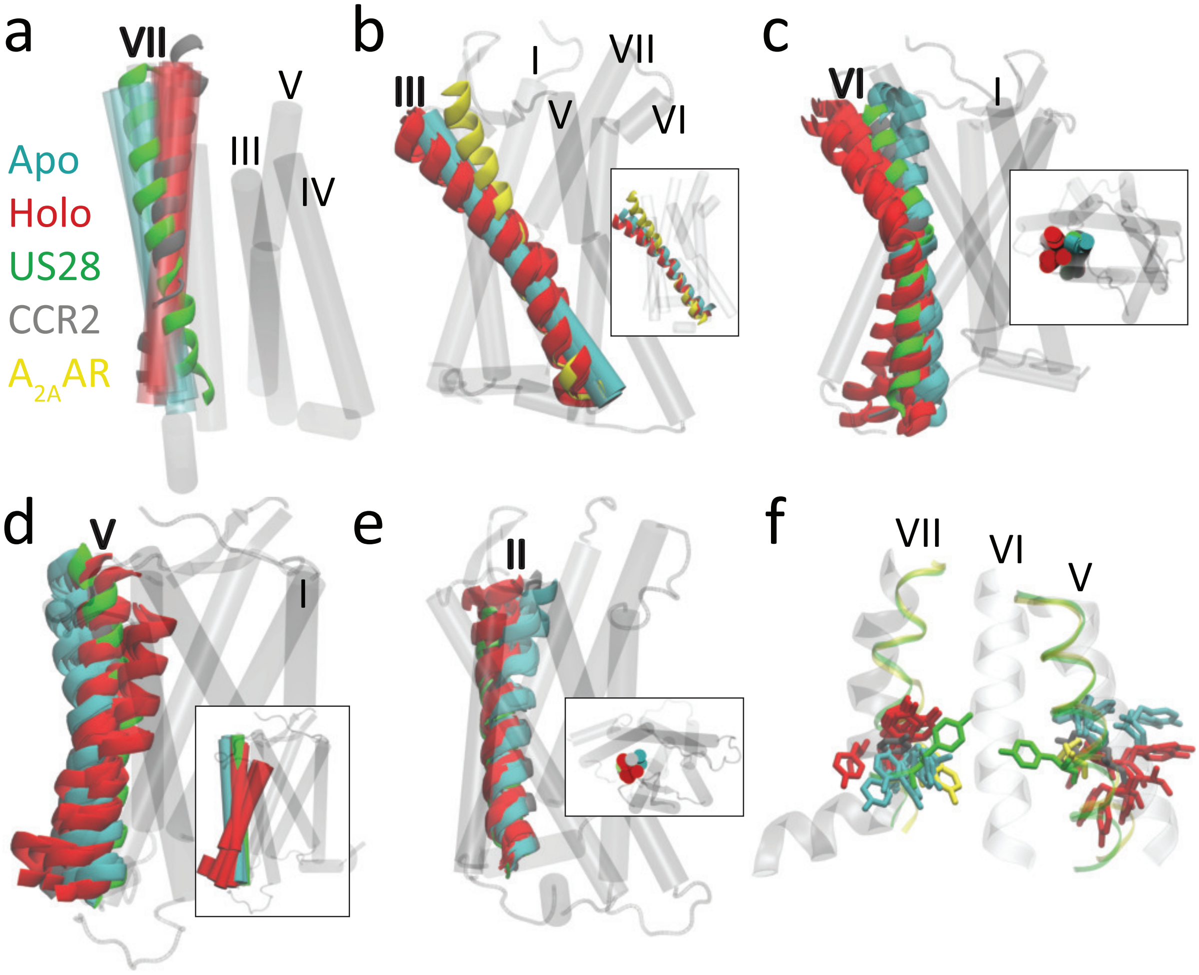


Figure 2.2: Apo and holo macrostates are compared to the active crystal structure of US28 (green, PDB ID 4XT3) or the active crystal structure of A_{2A}AR (yellow, PDB ID 2YDO) and the inactive crystal structure of CCR2 (grey, PDB ID 5T1A). A) Helix VII of apo macrostates resemble the active conformation of US28; holo macrostates resemble the CCR2 crystal structure. B) The conformation of helix III in apo macrostates subtly resemble active A_{2A}AR; holo macrostates are tilted away from the center of the helical bundle. C,D) Helix V and VI of apo macrostates straighten or tilt in toward the center of the protein, similar to the active conformation of US28 and the inactive CCR2 crystal structure; helix V and VI of several holo macrostates tilt away from the binding sites, accessing more active-like conformations than the apo macrostates or even the active state of US28. E) Helix II in the apo macrostates shifts inward; in the holo macrostates it shifts outward. F) In licorice are conserved motifs TYR $305^{7.53}$ and TYR $222^{5.58}$. All six apo metastable state assume a new conformation for TYR $305^{7.53}$, pointing intracellulary and in a similar conformation to active A_{2A}AR. Six out of the seven holo metastable states have TYR $305^{7.53}$ in the same conformation as the equilibrated crystal structure. Post-ligand-dissociation holo state L assumes a new position of TYR $305^{7.53}$, more similar to the dominant apo conformation. Apo metastable states sample a narrower range of conformations for TYR $222^{5.58}$ than holo.

macrostates K and L, indicating that the inactive (inward) conformation of helix VI may play a critical role in stabilization of the allosteric ligand (SI Appendix, Fig. 2.6).

3) Apo macrostates show an active-like upward shift and lateral movement of helix III

Apo macrostates also resemble the active conformation by the slight upward shift of helix III; unlike holo macrostates, which remain in a position similar to the inactive crystal structure (Fig. 2.2B).

4) The rearrangements of conserved microswitches suggest that apo macrostates resemble active states, and holo macrostates resemble inactive states

a) NPxxY motif (Tyr $305^{7.53}$). In the inactive conformation of GPCRs, Tyr $305^{7.53}$ points towards helices I, II, or VIII (in CCR2, it points toward II), and in the active state Tyr $305^{7.53}$ points toward middle axis of helical bundle[15]. Each apo macrostate shows Tyr $305^{7.53}$ pointing downward into the intracellular (G-protein) binding pocket (Fig. 2.2F). This positioning of Tyr $305^{7.53}$ matches the intermediately-active conformation of A_{2A}AR, which also points down. It does not match the active conformation in US28 that points up, and is also distinct from the inactive crystal structure of CCR2. In six out of the seven the holo macrostates, Tyr $305^{7.53}$ is stabilized in the inactive state and matches the inactive CCR2 crystal structure conformation as well as the inactive chemokine receptor crystal structures of CCR5 and CCR9.

The holo macrostate in which Tyr $305^{7.53}$ is not stabilized in the inactive conformation is accessed after the orthosteric ligand dissociates (State L, Fig. 2.1C, 2.2F); the allosteric pocket residues rearrange and Tyr $305^{7.53}$ assumes a downward conformation similar to the apo states and CXCR4. These concerted events may indicate allosteric cross-talk between the chemokine binding site and the G-protein binding site.

b) The microswitch residue Trp $256^{6.48}$, and the interaction of the DRY motif (Arg $138^{3.50}$) with Tyr $222^{5.58}$. Apo and holo macrostates both maintain the same chi angle of the conserved microswitch residue Trp $256^{6.48}$ which describes an active GPCR when it switches from gauche to trans conformation and facilitates the interaction of Tyr $222^{5.58}$ and Tyr $305^{7.53}$. In the CCR2 crystal structure and the crystal structures of chemokine receptors CCR5 and CXCR4, Trp $256^{6.48}$ points upward and extends into the helical core. Each apo macrostate shows Trp $256^{6.48}$ in a single conformation pointing toward helix III

(SI Appendix, Fig. 2.7). In the holo macrostates, Trp 256$^{6.48}$ access three distinct conformations: one resembling the crystal structure but with the helix shifted slightly outward from the helical core, another that laterally twists toward helix V, and one conformation that points down into the helical core toward the G-protein binding site (SI Appendix, Fig. 2.7). This third conformation, in which the orthosteric ligand has dissociated, represented by the holo macrostate L, further suggests cross-talk between the chemokine binding site and the rest of the protein.

The interaction of these two tyrosines and Arg 138$^{3.50}$ also characterizes an active state GPCR [42]. In the inactive crystal structure of CCR2, Tyr 222$^{5.58}$ points toward lipids, sterically blocked by Phe 246$^{6.38}$ from interaction with Arg 138$^{3.50}$ and Tyr 305$^{7.53}$ [11]. In apo macrostates, Tyr 222$^{5.58}$ remains pointed toward the lipids, never swiveling around to interact with Arg 138$^{3.50}$ or Tyr 305$^{7.53}$ as occurs in activated GPCR states (Fig. 2.2F). Holo macrostates actually show increased range of motion of Tyr 222$^{5.58}$, diverging from the crystal structure and stabilizing in unique intermediate conformations. The steric obstruction from Phe 246$^{6.38}$ is alleviated in both apo and holo macrostates, as Phe 246$^{6.38}$ swings outward and points toward the lipids. The conformations of these microswitch residues indicate that both apo and holo macrostates are sampling different conformations.

5) Formation of continuous water pathway suggests movement of apo towards activation

Internal water molecules, which may influence conformational changes in GPCRs by interfering with hydrogen bonding networks of the receptor's backbone and side chains, are postulated to be an integral part of receptor activation in GPCRs[43–46]. Work in other GPCRs has additionally shown that activation can allow water and sodium ion flow through GPCR core[47]. Furthermore, it has been shown that the activation of GPCRs is voltage sensitive[48]. Our simulations enable the direct visualization of water and sodium ion density in both CCR2 ensembles.

A continuous internal water pathway forms in apo CCR2 (Fig. 2.3A). The antagonists disrupt this water pathway, slowing the rate of water entry to and egress from the protein core (Fig. 2.3A). An analysis of the water occupancy per residue (SI Appendix, Fig. 2.8A) indicates that several of the high water occupancy residues (e.g. Asp 36$^{1.26}$, Ser 50$^{1.40}$, Glu 235$^{6.27}$, Lys 236$^{6.28}$, Glu 310$^{8.48}$, Lys 311$^{8.49}$) may be exposed to more water in the apo simulations than in the holo simulations simply because the

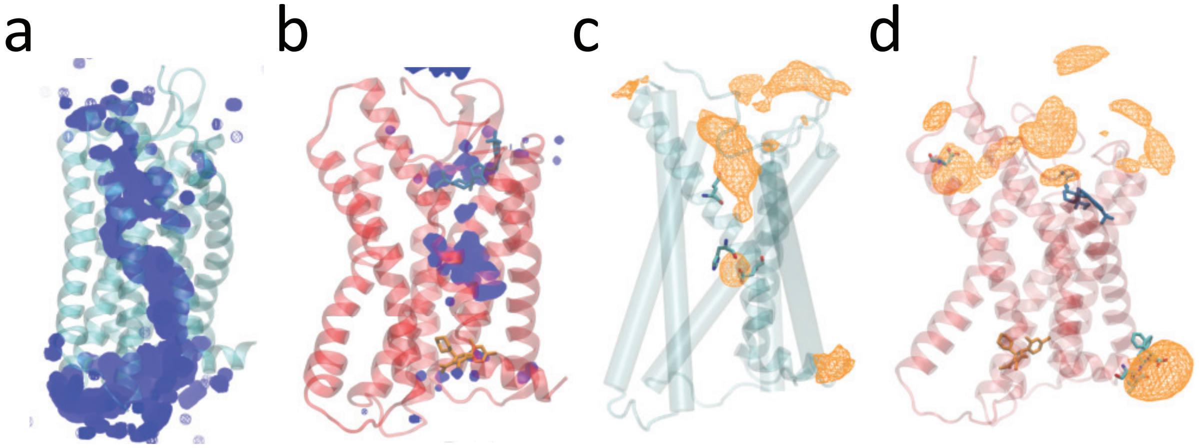


Figure 2.3: Ligands disrupt a continuous internal water and sodium ion pathway. Average water density over a 50 microsecond simulation of A) apo (teal) and B) holo (red). The orthosteric ligand is shown in blue and the allosteric ligand is shown in orange. Total average sodium ion density in C) apo and D) holo. Highest occupancy residues are depicted in cyan licorice and plotted in SI Appendix Fig. 2.8.

ligands have been removed and the water has access to the binding pockets. The other residues (e.g., Asp $78^{2.40}$, Tyr $80^{2.42}$, Asp $88^{2.50}$, Leu $92^{2.54}$, Ile $93^{2.55}$, Gly $127^{3.39}$, Ile $128^{3.40}$, Glu $291^{7.39}$, and Phe $312^{8.50}$) reside in the protein core, along the continuous pathway (Fig. 2.3A).

Class A GPCRs possess a conserved sodium binding site at $\text{Asp}^{2.50}$ corresponding to Asp 88 in CCR2[49]. The role of sodium is thought to contribute to the mechanism of receptor activation[50–52]. In particular, dynamics of activation were previously hypothesized to impinge upon the sodium binding pocket, eventually leading to ion permeation from the sodium binding site into the cytosol[51]. A sodium ion occupancy per residue analysis (2.3C, SI Appendix, Fig. 2.8B) indicates that, while no sodium permeation events into the cytosol were observed in the apo trajectories, ions interact with sodium binding site residues Asp $88^{2.50}$, Glu $291^{7.39}$, and His $297^{7.45}$. In holo CCR2, sodium does not interact with binding site residues, preventing the possibility of a permeation event (2.3D, SI Appendix, Fig. 2.8B).

2.4 Effects of antagonist binding on CCR2 dynamics

Comparisons of the apo and holo MSMs elucidate the effects of antagonists on CCR2 dynamics. Notably, apo relaxation timescales are an order of magnitude less than holo relaxation timescales (SI Appendix, Table S2), indicating that the antagonists greatly perturb CCR2 dynamics. The motions described by apo and holo TIC 0 represent the most striking difference between the two systems' dynamics. In the apo MSM, the inter-residue distances most closely correlated with apo TIC 0 are all a part of the allosteric (G-protein) binding pocket, whereas in the holo MSM, the inter-residue distances most closely correlated with holo TIC 0 are all a part of the orthosteric (chemokine) binding pocket (SI Appendix, Fig. 2.9).

Apo TIC 1 represents the flipping of Trp $98^{2.60}$ into the orthosteric drug binding site (SI Appendix, Fig. 2.11A,B, 2.9C). In the crystal structure Trp $98^{2.60}$ packs with the tri-substituted cyclohexane of the orthosteric antagonist, BMS-681[11]. Without the presence of this ligand, Trp $98^{2.60}$ assumes three distinct positions. The Trp $98^{2.60}$ conformation in the cluster at the neutral TIC (boxed in yellow, SI Appendix Fig. 2.11A,B) most closely resembles the conformation of Trp $98^{2.60}$ in the active GPCR US28, which is shifted slightly up and in towards the helical core in comparison to the CCR2 crystal structure. The two other conformations are found at the extreme ends of apo TIC 1 in densely populated free energy wells. Of these two conformations, state F assumes the most dramatic conformation and protrudes into the chemokine binding site (SI Appendix, Fig. 2.10). In the holo macrostates there is markedly less intrusion into the binding pocket due to the presence of the ligand.

Holo TIC 1 represents the concerted movement of 5 pairs of residues in the orthosteric ligand binding site during orthosteric ligand dissociation (SI Appendix, Fig. 2.11C,D, Fig. 2.9D). The separation projected in the first two TICs (SI Appendix, Fig. 2.11D) is divided into clusters of frames that occur before (white clusters), during (grey), and after (black) dissociation. The residue pairs identified by TICA that contribute to holo TIC 1 and this ligand dissociation (SI Appendix, Fig. 2.9D) were confirmed by analyzing the original simulation data. The key changes are the change in distance between Tyr $49^{1.39}$ - Thr $292^{7.40}$, Trp $98^{2.60}$ - Tyr $120^{3.32}$, Ser $50^{1.40}$ - Tyr $259^{6.51}$, and the chi angle of Glu $291^{7.39}$. Notably, four of these residues (Tyr $49^{1.39}$, Trp $98^{2.60}$, Tyr $120^{3.32}$, and Thr $292^{7.40}$) are not only involved in binding to the co-crystallized orthosteric antagonist BMS-681 and/or CCL2 binding, but are also critical for GPCR activation [53, 54].

The positioning of the orthosteric ligand and the conformation of Trp $98^{2.60}$ are closely linked (SI Appendix, Fig. 2.11C, Fig. 2.12). After ligand dissociation, in holo states K and L (purple and white, respectively), Trp $98^{2.60}$ turns towards helix III, bending slightly inward toward the chemokine binding site.

Prior to ligand dissociation, Trp $98^{2.60}$ has two distinct conformations. In the first conformation (states I and G, yellow and black), the ligand positions itself between helices I and VII, in the same conformation as the crystal structure. Trp $98^{2.60}$ is constrained in a downward position, pointing intracellularly, also resembling the CCR2 crystal structure conformation and the crystal structure of chemokine receptor CCR5. In the second conformation (States H and J, cyan and grey), Trp $98^{2.60}$ flips up and out of the binding pocket, pointing extracellularly, and the ligand moves between helices I and II. This conformation of Trp $98^{2.60}$ more closely resembles CCR9 (SI Appendix, Fig. 2.13). The third conformation of Trp $98^{2.60}$ is found in State M (red), and is the most prominent position of the residue as it extends deeper into the chemokine binding site toward helix III. In this case, the ligand interacts with helices II, IV, and V, and there are no transitions from this state to a dissociated state.

As in the apo MSM, the absence of the orthosteric ligand causes a shift in the position of Trp $98^{2.60}$. In the holo simulations shown in SI Appendix, Fig. 2.14, the dissociation event is preceded by a doubling of the distance between Trp $98^{2.60}$ and Tyr $120^{3.32}$, and 3 μs after dissociation the distance returns to its previous 0.4 nm. This increase in distance may be required for the ligand to begin the process of dissociating. Another drastic change during the dissociation event is the switch of Glu $291^{7.39}$ from a constrained chi angle of -50 to -100 degrees to an unconstrained chi angle (SI Appendix, Fig. 2.15). After dissociation, this angle more closely resembles the conformation in all apo simulations. Glu $291^{7.39}$ is a key mediator of many CCR2 antagonists[55], but there is no direct interaction between Glu $291^{7.39}$ and the orthosteric antagonist in the CCR2 crystal structure[41]. That the conformation of Glu $291^{7.39}$ switches after dissociation suggests that Glu $291^{7.39}$ is involved in ligand stabilization despite not directly interacting with the ligand.

In the CCR2 crystal structure, there is a hydrogen bond between Tyr $49^{1.39}$ and Thr $292^{7.40}$. The gamma-lactam secondary exocyclic amine of the orthosteric ligand forms a hydrogen bond with the hydroxyl of Thr $292^{7.40}$, and the carbonyl oxygen of the gamma-lactam forms a hydrogen bond with Tyr $49^{1.39}$. During simulation, the distance between Tyr $49^{1.39}$ and Thr $292^{7.40}$ remains stable until 3 μs after the ligand dissociates, when it begins fluctuating (SI Appendix, Fig. 2.16). This suggests that the orthosteric ligand dissociation breaks the hydrogen bond between these key ligand binding residues. This motion is captured in the holo MSM: the separation of the two residues is exemplified between States H (pre-dissociation) and L (post-dissociation) in SI Appendix, Fig. 2.16.

Finally, the faster dominant motions (TICs 2, 3, 4) in the holo MSM consist of rearrangements in the allosteric ligand binding site, suggesting that an allosteric rearrangement must first happen in order for the orthosteric ligand to dissociate. Further evidence for this is the observed correlated motion of the downward flip of the conserved residue Tyr $305^{7.53}$ in the G-protein binding site with the dissociation of the orthosteric ligand from the chemokine binding site.

Overall, holo macrostates show more helical tilting and binding site expansion, which increases the solvent-accessible surface area (SASA) when compared to the crystal structure and the apo macrostates. However, the apo simulations overall have greater residue fluctuation, suggesting that the antagonist ligands dampen CCR2 dynamics (SI Appendix, Fig. 2.17).

2.4.1 Opening of a cryptic druggable pocket

A dramatic expansion of the extracellular (chemokine) binding site is exhibited in the holo macrostates. The expansion is caused by a pronounced outward tilting of helix VI and slight outward tilting of helix II in the holo macrostates, whereas the apo macrostates show the opposite, with a slight inward tilting of both helices VI and II (Fig. 2.2C,E). The intracellular (G-protein) binding site also enlarges in the holo macrostates due to the outward shift of the intracellular ends of helices V and VI, but remains obstructed in all apo and holo states. In the crystal structure, this obstruction occurs by the interaction of Arg $138^{3.50}$ with Asp $137^{3.49}$ and with Thr $77^{2.39}$ [11], which are maintained throughout all the simulations. The outward movement of the intracellular end of helix VI and the movement of helix V toward helix VI in states L, J, and H in the holo MSM create a putative site for novel allosteric antagonists; this pocket also transiently appears in the apo simulations (Fig. 2.4). Computational solvent mapping[56] of this novel site indicates that the pocket presents surfaces that are amenable to ligand binding due to its ability to bind clusters of multiple different drug-like probes (Fig. 2.4C). The pocket can be accessed through the lipid bilayer between helices IV and V, or from the G-protein binding site, as a deeper extension of the current allosteric binding site of CCR2-RA-[R], and may be useful for rational drug design or modification of current antagonists.

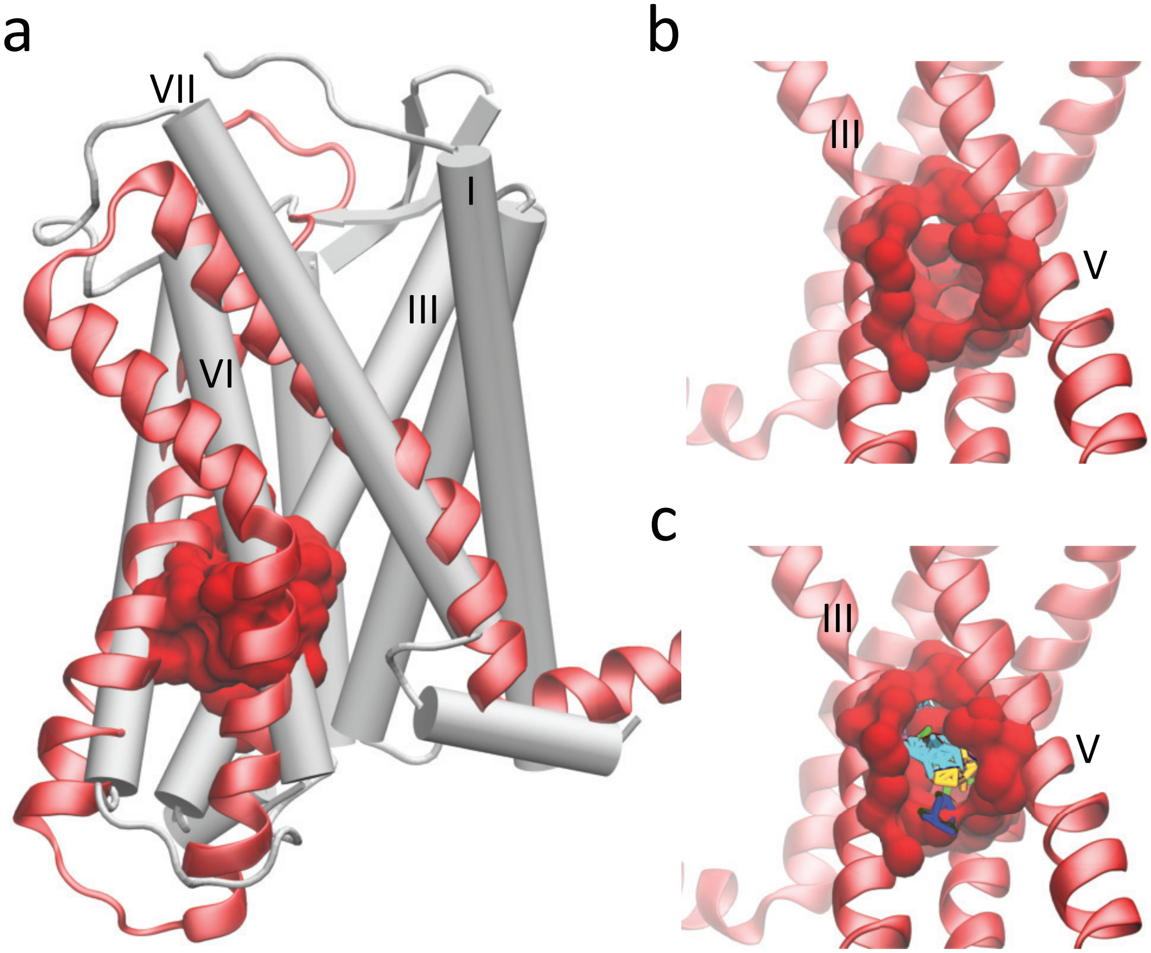


Figure 2.4: A putative allosteric drug binding pocket is revealed by three holo macrostates. A) A comparison of the CCR2 crystal structure (white cartoon) with helices V, VI, and VII (red new cartoon) of one holo macrostate. The pocket is shown in red surface. B) A closer view of the pocket from the other side of the protein, between helices III and V. C) Small organic probes used for computational fragment mapping are multicolored.

2.5 Conclusions

To characterize the basal dynamics of CCR2 and understand how small molecule antagonists modulate these dynamics, we coupled long timescale atomic simulations and MSM theory to compare the metastable states accessed by apo and holo CCR2 in its native membrane-embedded form.

Antagonists perturb CCR2 dynamics and kinetics, and are associated with distinct residue rearrangement and key motions. Several intermediate states reveal a novel cryptic binding site that could be targeted with small molecule inhibitors. In a previous study[57], cryptic pockets predicted with MSM theory have been experimentally confirmed and suggest that this methodology can successfully be used to guide drug discovery efforts.

Without antagonists, CCR2 is able to access other distinct metastable states that are likely sampling along an activation pathway. These intermediate states inform on the basal dynamics of CCR2 and may be useful for modification of previously unsuccessful drugs.

2.6 Abriged Materials and Methods

See the SI Appendix for full Materials and Methods. MD trajectories and MSM construction scripts are available for download[23].

2.6.1 System Preparation and Molecular Dynamics Simulations

Two systems were simulated for a total of 260 μs: CCR2 holo, with both co-crystallized antagonist ligands bound, and CCR2 apo, without ligands bound. CCR2-RA-[R] and BMS 681[11] were removed to build the apo system. Each all-atom system is embedded in a POPC bilayer, explicitly solvated with TIP3P, and simulated with 150mM NaCl, at pH 7.4, at 310K and 1 bar. The initial coordinates were taken from the experimental crystal structure[11].

2.6.2 Building the Markov State Models

The MSMs were built with PyEMMA version 2.5.4 [58] and selected based on implied timescale plots (SI Appendix, Fig. 2.18) and Chapman-Kolmogorov tests (SI Appendix, Figs. 2.19, 2.20) and coarse-grained with hidden Markov models (HMMs). Representative structures were selected from each macrostate by taking the centroid of the most populated microstate (SI Appendix, Figs. 2.21,2.22,2.23).

2.7 Acknowledgements

Chapter 2, in full, is a modified reprint of the material as it appears in "B. C. Taylor, C. T. Lee, and R. E. Amaro. *Structural Basis for Ligand Modulation of the CCR2 Conformational Landscape*. Proceedings of the National Academy of Sciences, U.S.A. (2019). DOI: 10.1073/pnas.1814131116." The dissertation author was the primary investigator and author of this work. The dissertation author designed the research, performed and analyzed the molecular dynamics simulations, built and analyzed the Markov state models, designed and created the figures, and wrote the manuscript.

The authors thank Tracy Handel and Irina Kufareva for their valuable input and initial modification of CCR2 coordinates. We thank Frank Noe and Cecilia Clementi for their useful discussions regarding MSM construction. We also thank the organizers and participants of the 2018 Workshop on Free energy Methods, Kinetics and Markov State Models in Drug Design for helpful input. This work was supported in part by the Director's New Innovator Award Program NIH DP2-OD007237, the National Biomedical Computation Resource (NBCR) NIH P41-GM103426, and the National Science Foundation through XSEDE supercomputing resources provided via TG-CHE060073 to R.E.A. C.T.L. also acknowledges support from the NIH Molecular Biophysics Training Program (T32-GM008326). Anton 2 computer time was provided by the Pittsburgh Supercomputing Center (PSC) through Grant R01GM116961 from the National Institutes of Health. The Anton 2 machine at PSC was generously made available by D.E. Shaw Research.

2.8 SI Appendix

2.9 Methods

2.9.1 System Preparation

Two systems were simulated: holo CCR2, with both co-crystallized antagonist ligands bound, and apo CCR2, without the ligands bound. Each all-atom system is embedded in a biologically similar POPC bilayer and explicitly solvated with TIP3P. The initial coordinates were taken from the experimental crystal structure[11] and simulated for 50ns MD simulations on local resources before simulation on Anton2. All simulations are in a POPC lipid bilayer and cubic water box with 150mM NaCl, at pH 7.4, at 310K and 1 bar.

The two small molecules CCR2-RA-[R][11, 59] and BMS 681[11, 60] were deleted to build the apo system. Both systems were protonated at pH 7.4 in Maestro-integrated PROPKA. A POPC lipid bilayer was added to each system and solvated with TIP3 waters and 0.15 M NaCl using CHARMM-GUI[61]. The small molecules were parameterized with CGenFF[62]. System coordinates were parameterized with the CHARMM36[63] force field. No restraints were added.

2.9.2 Modification of CCR2 coordinates

The coordinates for CCR2 were taken from the experimental crystal structure [11] and modified to build the apo (unbound) and holo (dual-antagonist-bound) systems. For both systems, the T4 Lysozyme was removed from the crystal structure and intracellular loop 3 and part of the ECL3 and N terminus was constructed. For ICL3, a peptide containing residues 223:243 was built ab-initio, the backbones of residues 223:231,236:243 and the side chains of residues 223:226,241:243 were tethered to their respective positions, the receptor represented as a set of potential grid maps representing vw, el, hb, and sf "potentials" and, and the peptide was sampled in these maps. For NT/ECL3, the protocol is similar except that 2 separate peptides are built (31:41 and 276:285), a disulfide bond is imposed between 21 and 277, and the entire thing is sampled as above. There are several zero-occupancy side chains whose conformations are predicted as a part of this simulation.

Best scoring conformations of the two fragments are merged with the rest of receptor coordinates and the system is minimized in its full-atom representation: first by exhaustively sampling polar rotatable hydrogens, then by minimizing the side-chain conformers, then by Monte-Carlo sampling of side-chain conformers, then by minimizing everything. During these steps, harmonic restraints of gradually decreasing strength are imposed between the model and either the X-ray coordinates or the best prediction conformations of the built regions. Towards the end of the optimization, the restraints were released almost entirely and the complex remained stable. This was done in the presence of ligands. Zn ion and water molecules were removed.

2.9.3 Molecular Dynamics Simulations

Both systems were minimized and equilibrated using the GPU version of AMBER12. The systems were minimized at NPT for 15,000 total steps and were equilibrated for 2 sequential 25 nanosecond runs. The systems were then simulated for 50 nanoseconds in the NPT ensemble at 310K and 1 bar with 2-fs time-step and particle mesh Ewald electrostatic approximation. The additional replicates were made from the final production output by simulating for three additional nanoseconds to scramble the input velocities.

MD simulations on Anton2 were performed on the final coordinates from the short GPU-enabled AMBER12 simulations. The Anton 2 simulations were run in the NPT ensemble, using Anton's Nose-Hoover thermostat-barostat, at 310K and 1 bar with a 2.5-fs time-step and particle mesh Ewald electrostatic approximation. The two systems were simulated for an aggregate total of 260 microseconds (SI Appendix, Table 2.1, Fig.2.5).

- Ensemble and Constraints: After minimization, equilibration, and initial production runs, both simulations were run as NPT ensembles using Anton's Multigrator framework. No constraints were used in the simulations.

- Boundaries: The fully solvated systems are cubic with X, Y, Z unit lengths of 72 Å, 72 Å, and 103 Å respectively.

- Force Fields: CHARMM36 force field with TIP3P waters; small molecule ligands were parameterized with CGenFF.

- Atom Count and Types: Each system contains 55,000 atoms. The systems are composed of the protein, POPC lipids, water, small molecule drugs, Na+, and Cl-.

2.9.4 Trajectory Preparation

MD trajectories were processed using VMD[64]. All frames were aligned to the first frame using all residues of the protein. The frame rate for trajectories was 240ps, the standard for Anton2 simulations. The trajectories were converted into NAMD's .dcd trajectory format for analysis with PyEMMA[58], MSMBuilder[65], and in-house scripts.

2.9.5 Markov State Models

A Markov State Model (MSM) is a time-dependent master equation that describes the probability of transitioning between discrete states at a fixed time interval. These models are required to have the Markovian property (i.e., the probability of transitioning between discrete states is independent of previous transitions). By clustering protein structures extracted from an MD trajectory, discrete conformational states can be identified for use in MSMs[32, 65–67]. Transitions between conformational clusters observed over the course of an MD trajectory are tallied, and the MSM is then built from the transition probabilities between these distinct clusters. MSM/MD analysis provides access to the thermodynamic, kinetic, and structural characteristics of the protein conformational ensemble (i.e., a robust description of the free-energy landscape of the protein)[32, 65–69]. The thermodynamics of the various conformational states can be calculated from the equilibrium distribution. It is also possible to resolve the transition kinetics between individual states, the concerted or principal protein motions, metastable states, and the transition pathways between discrete states[65, 67, 68]. Lastly, the source molecular dynamics simulations provide representative cluster structures for use in structure-based drug design[69].

2.9.6 Building the Markov-State Models

We used time-structure independent component analysis (TICA)[35, 36, 70] starting with all pairwise inter-residue distances to perform dimensionality reduction and identify the features and collective variables (time-structure based independent components (TICs)) that best represent the dominant slow motions in the apo and holo simulations. To reduce the number of pairwise distances to a manageable number, we employed an iterative TICA feature selection approach. First, TICA was run on a curated starting set of hundreds of distances between transmembrane helices. Features with low TIC correlation were then removed from the set and TICA run on the resultant new basis. In this iterative fashion, we selected 22 representative features, listed below. These resultant features are the 22 sets of distances between residues in the orthosteric and allosteric ligand-binding pockets (Fig. 2.1A).

The apo and holo systems were clustered separately using K-means. Two MSMs were built: one MSM was built on the apo data and a second MSM was built on the holo data. In each case, the data was projected separately into TICA space and the trajectory frames were clustered using K-means clustering implemented in PyEMMA[58]. The MSMs were selected based on implied timescale plots (SI Appendix, Fig. 2.18) and the Chapman-Kolmogorov test[32] was used to test the consistency between the MSMs and the MD simulations (SI Appendix, Fig. 2.19, 2.20). The apo MSM has a lag time of 14.4 nanoseconds and 665 clusters; and the holo MSM has a lag time of 48 nanoseconds and 790 clusters. The MSMs were coarse-grained using hidden Markov models (HMMs)[71] to identify metastable macrostates and transitions between those states. The apo MSM is coarse-grained into six macrostates; the holo MSM into seven macrostates (Fig. 2.1B,C). A representative structure for each macrostate was selected by taking the centroid of the most populated microstate, using MSMBuilder[65] (SI Appendix, Fig. 2.21,2.22, 2.23).

Set of 22 Features:

Ile 40, Asn 199; Tyr 222, Arg 138; Tyr 305, Arg 138; Tyr 305, Tyr 222; Tyr 49, Thr 292; Tyr 120, Glu 291; Tyr 120, Tyr 259; Glu 291, Tyr 259; Tyr 49, Trp 98; Trp 98, Tyr 120; Trp 98, Glu 291; Trp 98, Thr 292; Tyr 49, Tyr 259; Phe 246, Leu 81; Ile 245, Leu 134; Ile 245, Leu 81; Val 244, Tyr 305; Tyr 305, Leu 81; Tyr 305, Val 63; Tyr 305, Leu 134; Tyr 305, Leu 67; Leu 67, Val 244.

Table 2.1: System Information

	Apo			Holo		
Simulation Number	1	2	3	1	2	3
Simulation Time in microseconds	50	50	10	50	50	50
Ligands Bound	None			BMS-681 and CCR2-RA-[R]		
Number of Atoms	53,097			53,077		
Membrane Lipids	POPC					
Water Model	TIP3P					
Force Field	CHARMM36 FF					
Box Dimensions	72 Å, 72 Å,103 Å					

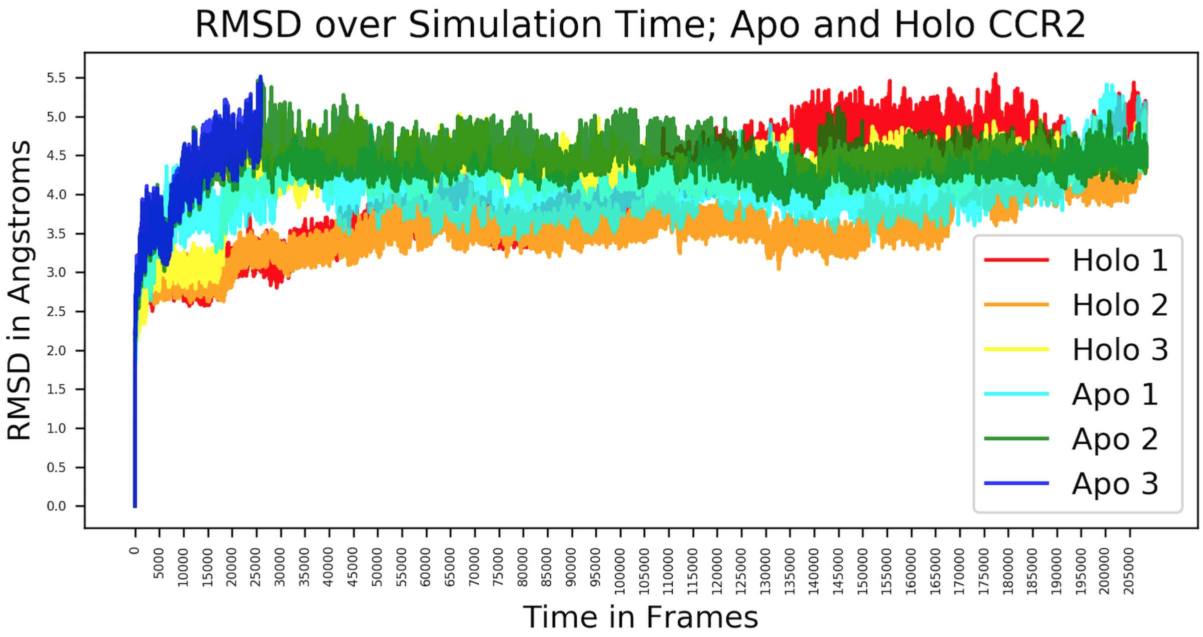


Figure 2.5: RMSD plot of all trajectories over simulation time.

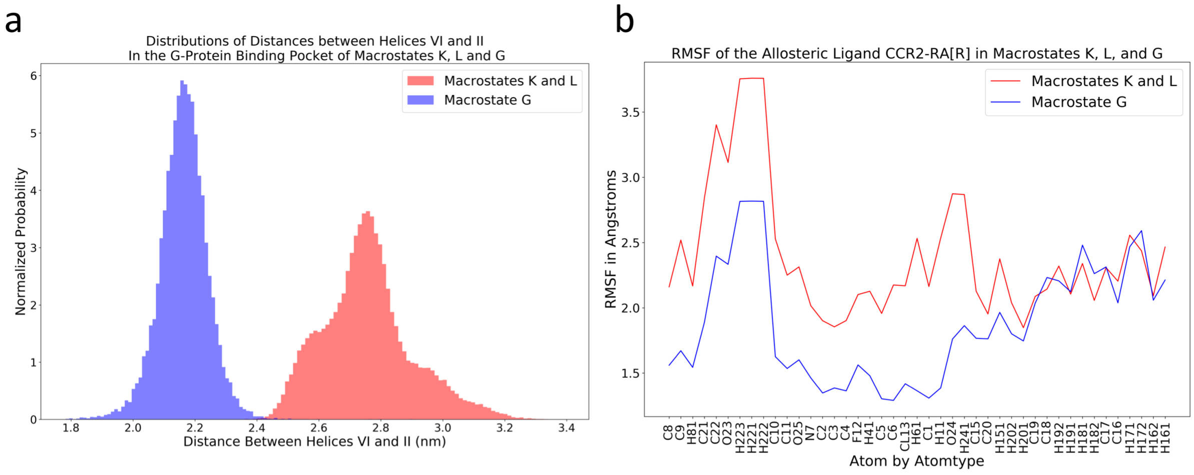


Figure 2.6: The RMSF of the allosteric ligand is larger in macrostates with more open G-protein binding sites. A) Distribution of distances in macrostates K, L and G between residues Gly 224 and Asp 78 in the intracellular regions of helices VI and II. B) RMSF of the allosteric ligand in holo macrostates K, L, and G.

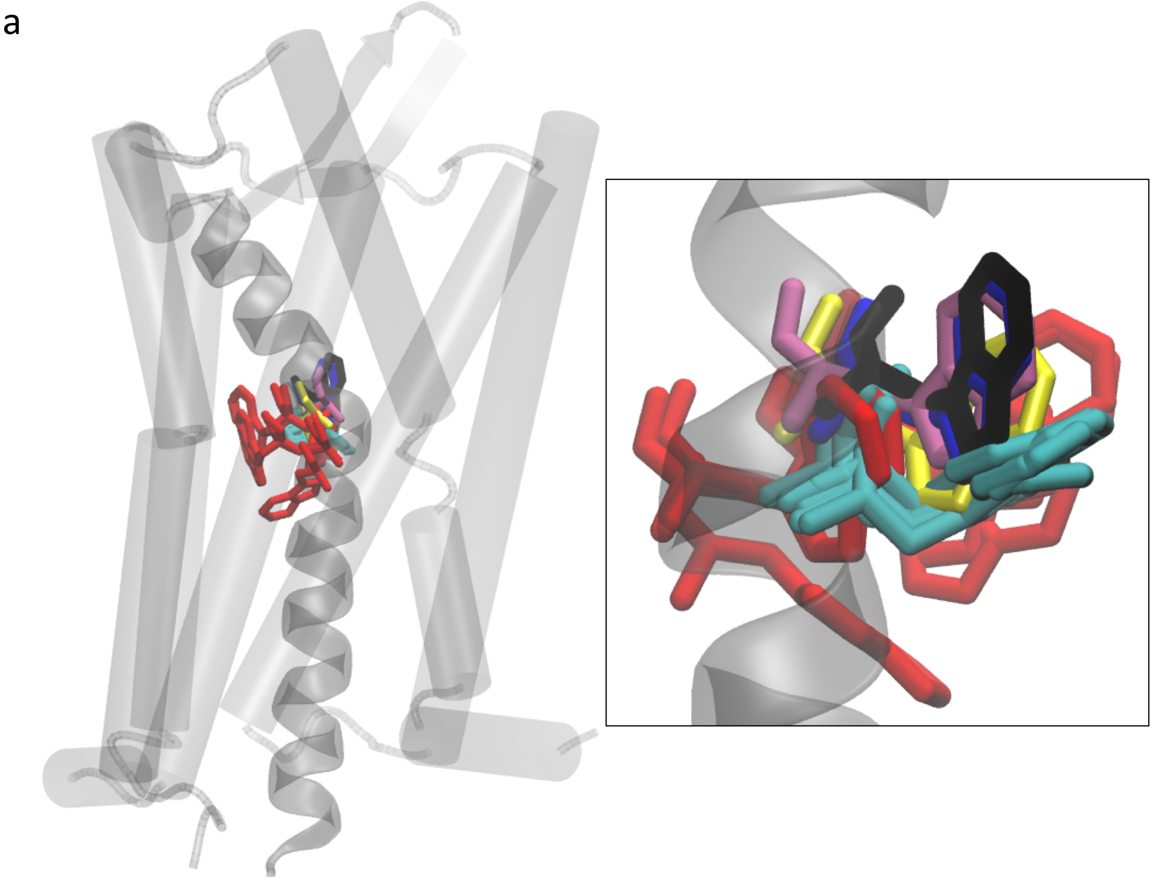


Figure 2.7: Trp 256$^{6.48}$ compared to the apo and holo macrostates and the crystal structures of CCR2 (PDB ID: 5T1A; grey, black) CCR5 (PDB ID: 4MBS; blue) and CXCR4 (PDB ID: 4RWS; mauve). A) Each apo macrostate shows Trp 256$^{6.48}$ in a single conformation pointing toward helix III. In the holo macrostates, Trp 256$^{6.48}$ access three distinct conformations: one resembling the crystal structure but with the helix shifted slightly outward from the helical core, another that laterally twists toward helix V, and one conformation that points down into the helical core toward the G-protein binding site.

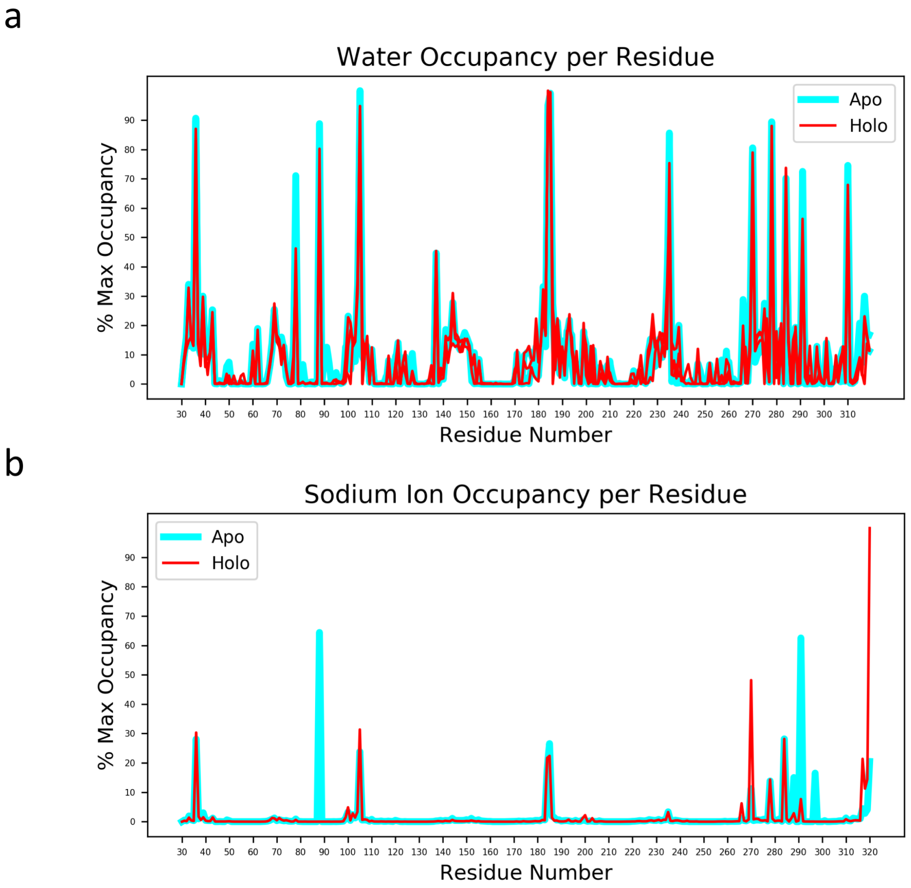


Figure 2.8: Ligands disrupt a continuous internal water and sodium ion pathway. A) Water occupancy per residue in all apo (teal) and holo (red) simulations. B) Sodium ion occupancy per residue in all apo and holo simulations.

Table 2.2: Relaxation timescales for the apo and holo MSMs. Units are in microseconds.

Apo	321.6 μs	57.7 μs	11.1 μs	9.0 μs	4.2 μs	2.8 μs	2.4 μs	1.7 μs	1.3 μs	1.3 μs
Holo	2246.9 μs	286.8 μs	84.9 μs	59.1 μs	44.3 μs	21.6 μs	8.2 μs	7.7 μs	4.4 μs	3.9 μs

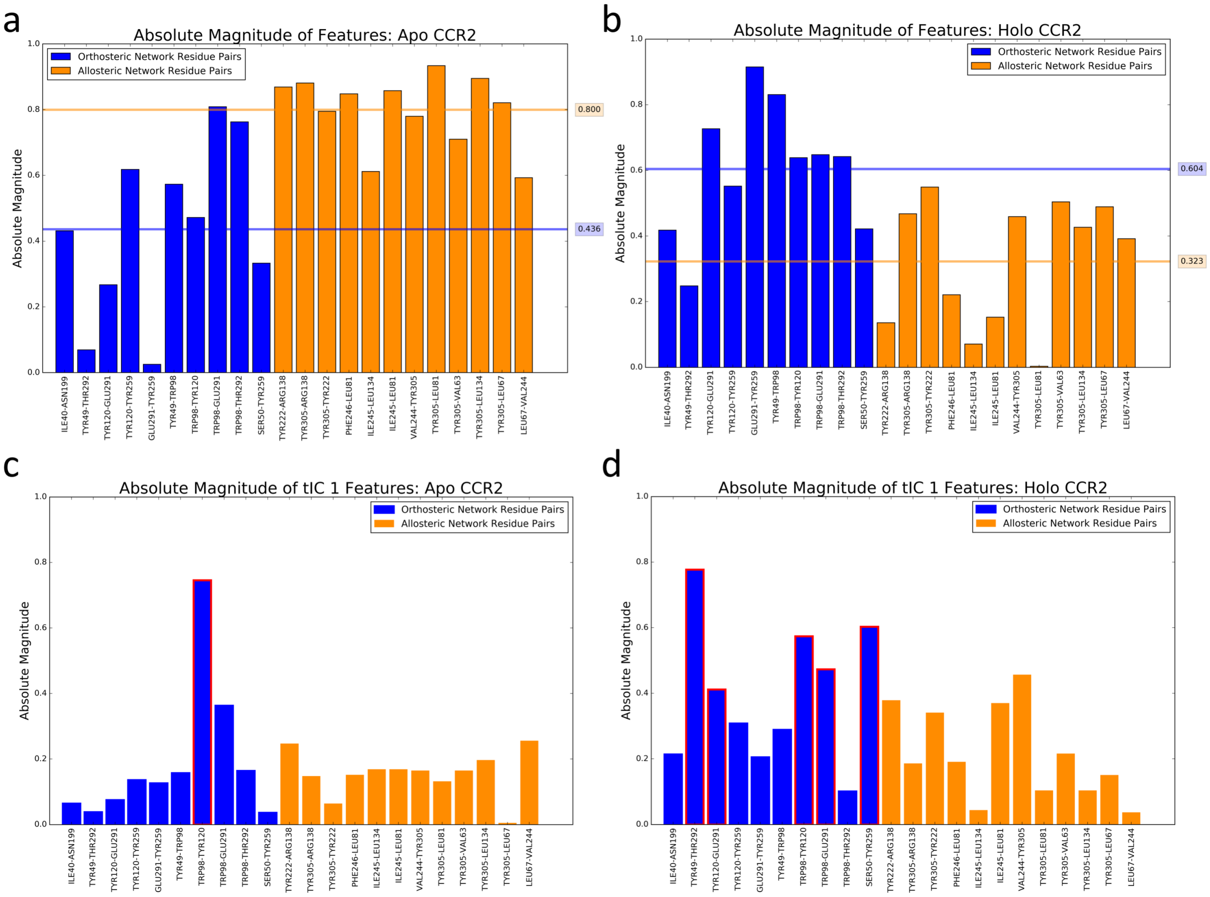


Figure 2.9: The absolute magnitude of each input feature in A) apo TIC 0, B) holo TIC 0, C) apo TIC 1, and D) holo TIC 1. Each bar represents the absolute magnitude of one inter-residue distance. Blue bars are distances between residue pairs in the orthosteric pocket, and orange bars are distances between residue pairs in the allosteric pocket.

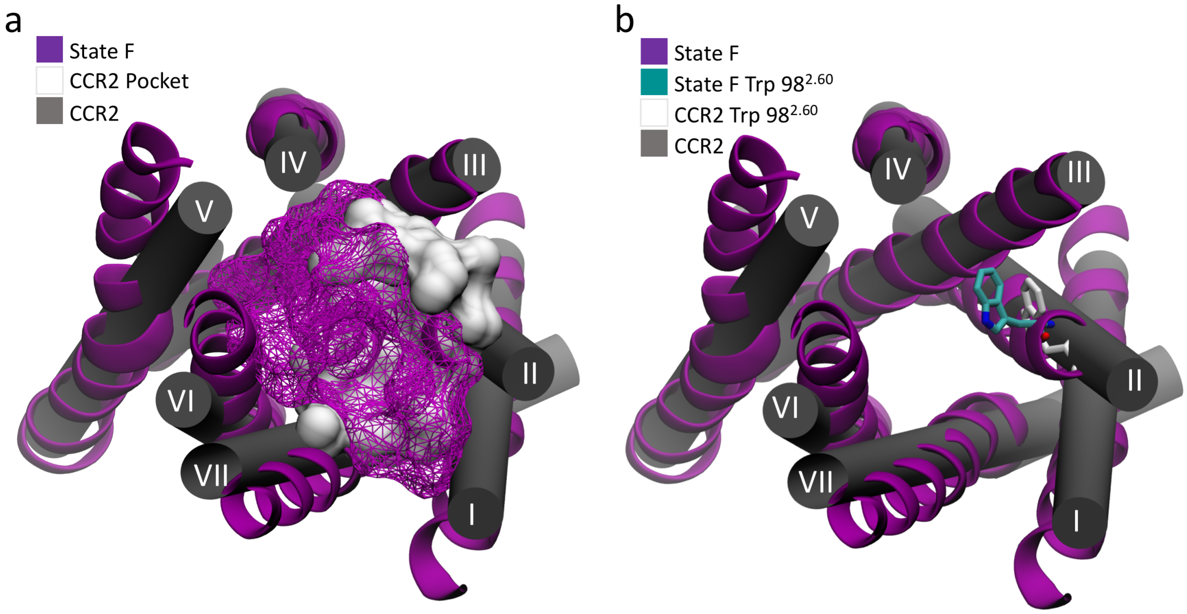


Figure 2.10: A) The shape of the chemokine binding site of apo state F in comparison to the crystal structure. Without the ligand, the binding site expands and rotates toward helices IV, V, and VI, and extends between helices I and VII. B) The conformations of Trp98$^{2.60}$ in apo state F and the crystal structure. Trp98$^{2.60}$ protrudes into the pocket in the absence of ligands.

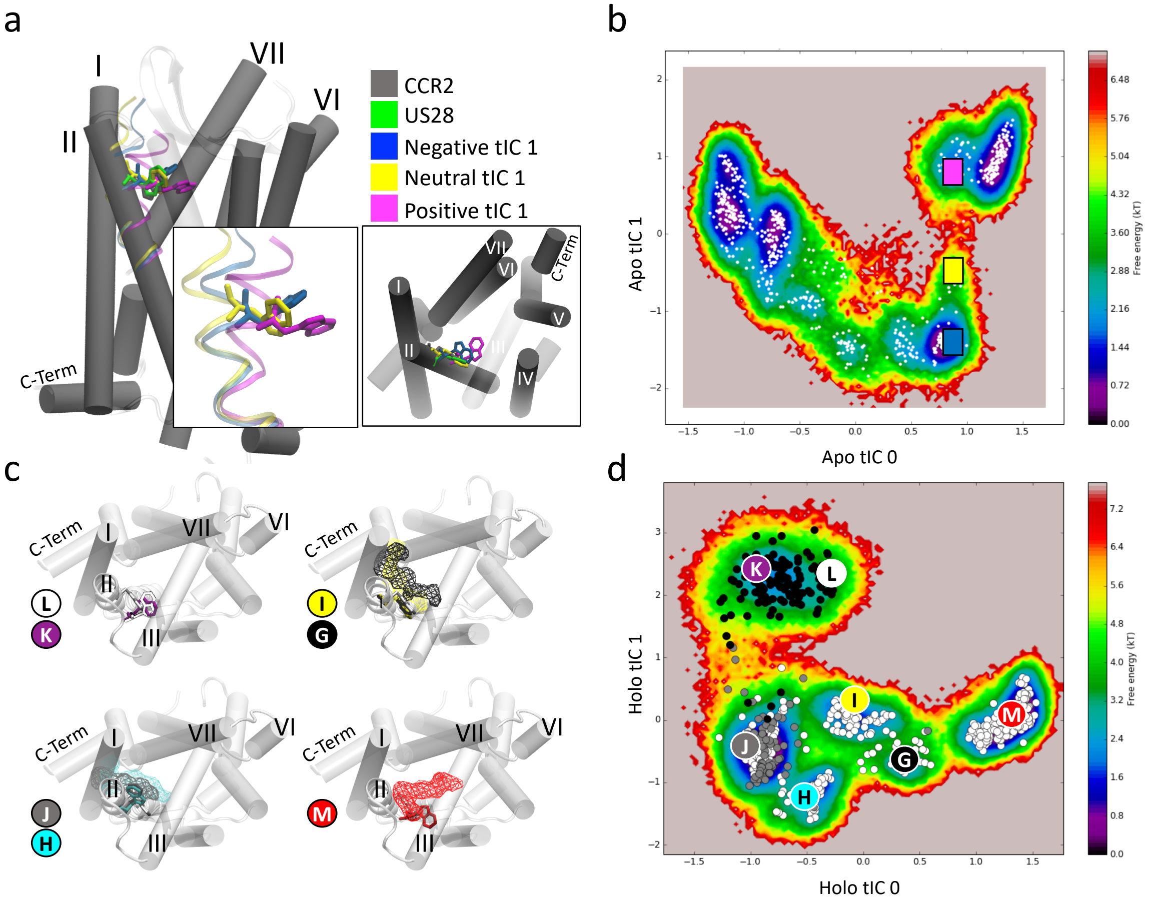


Figure 2.11: A) In apo CCR2, TIC 1 represents Trp 98$^{2.60}$ in three distinct positions. In gray is the crystal structure; in green is the active crystal structure of US28; in blue and yellow are transitions, and in magenta is the most dramatic conformation. Each conformation is plotted on the free energy in TICA space in B). C) In holo CCR2, the positioning of the orthosteric ligand and the conformation of Trp 98$^{2.60}$ is closely linked. Shown in light silver cartoon is CCR2 5T1A; Trp 98$^{2.60}$ is displayed as purple in state K, white in state L, gray in state J, cyan in state H, yellow in state I, black in state G, and red in state M. D) Holo CCR2. White circles are clusters of frames before any ligand dissociation. Grey circles are clusters of frames during the event. Black circles are clusters of frames after the event.

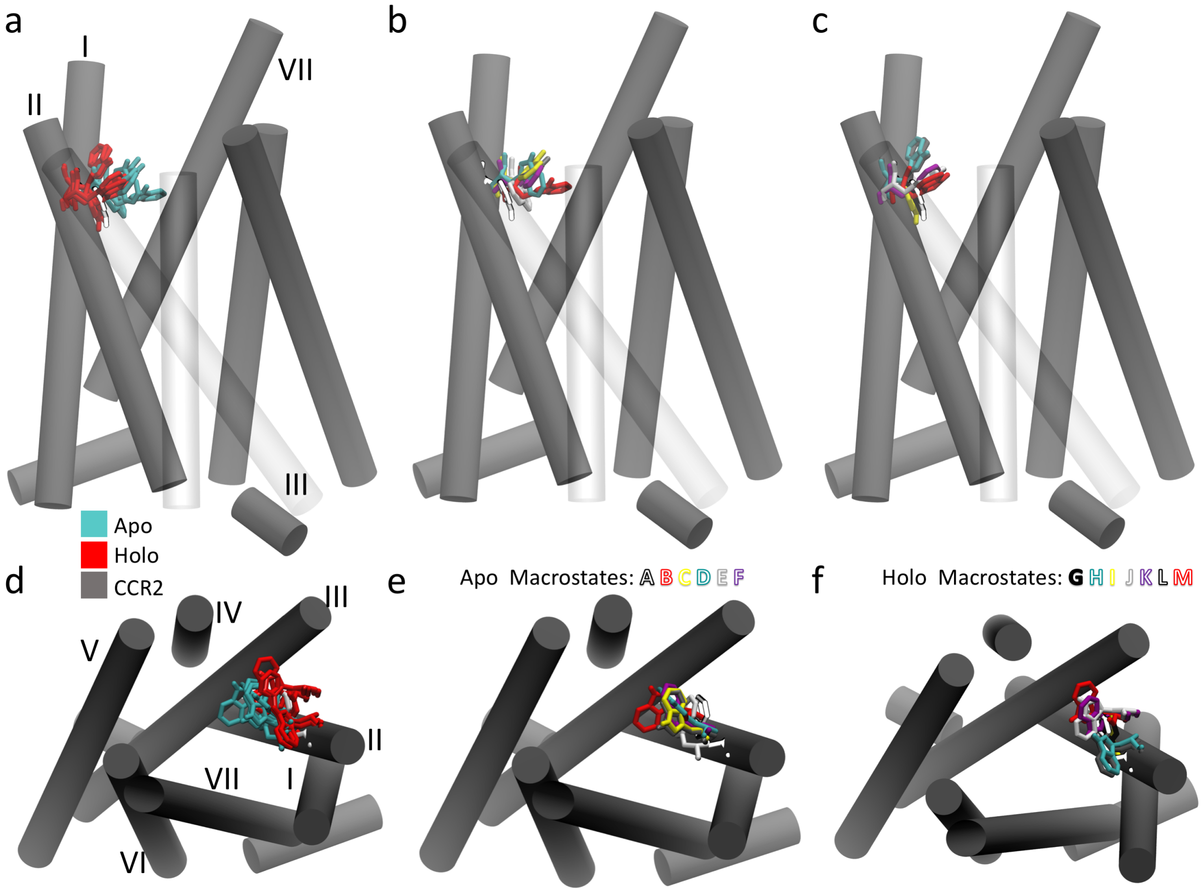


Figure 2.12: Trp $98^{2.60}$ extends farther into the chemokine binding site in apo CCR2 than in the holo crystal structure and holo macrostates. A side view of the CCR2 crystal structure (grey) and Trp $98^{2.60}$ (bright white) compared to A) the apo (cyan) and holo (red) macrostate conformations of Trp $98^{2.60}$ B) the holo conformations, and C) the apo conformations. The extracellular-to-intracellular view of A)-C).

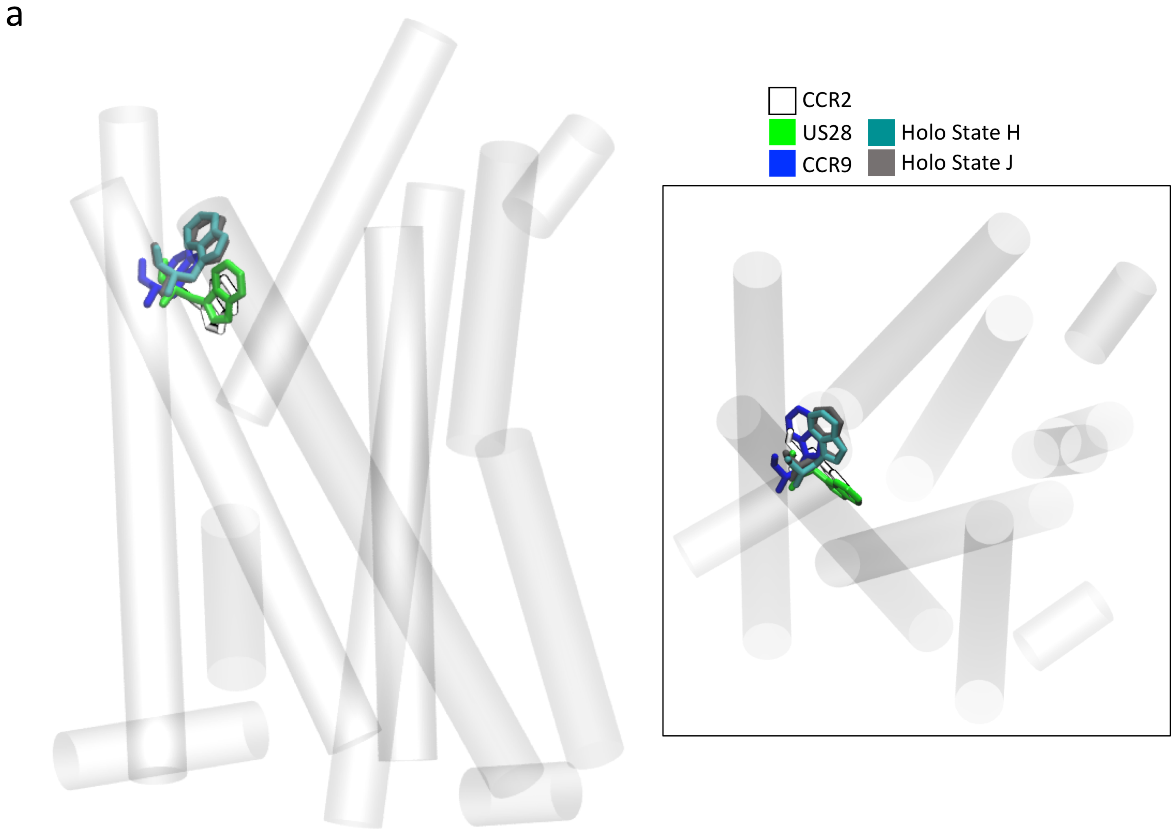


Figure 2.13: A) Trp 98$^{2.60}$ in holo states H and J (cyan and grey) compared to CCR9 (dark blue) and the CCR2 crystal structure (white). Inset depicts the extracellular-to-intracellular view of A).

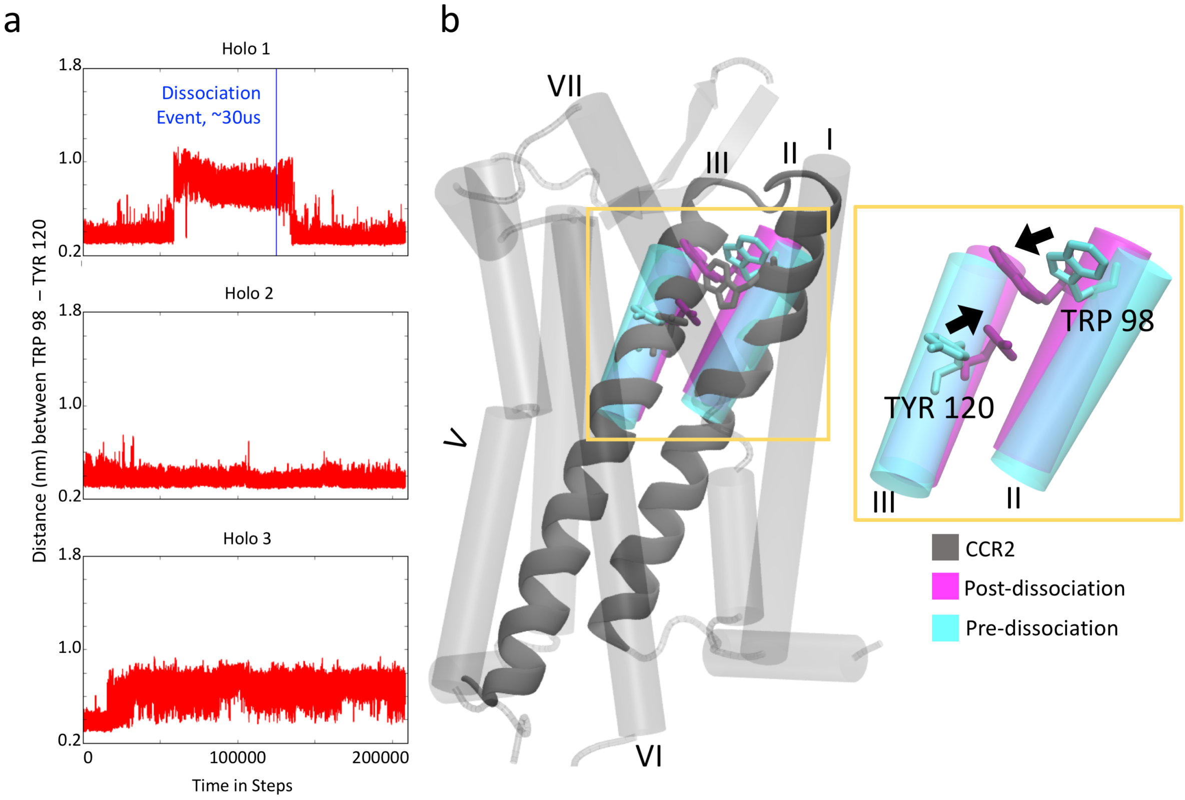


Figure 2.14: A) Distance between residues Trp $98^{2.60}$ and Tyr $120^{3.32}$ over simulation time. Dissociation event noted by blue line. B) Conformation of CCR2 before (teal) and after (purple) ligand dissociation.

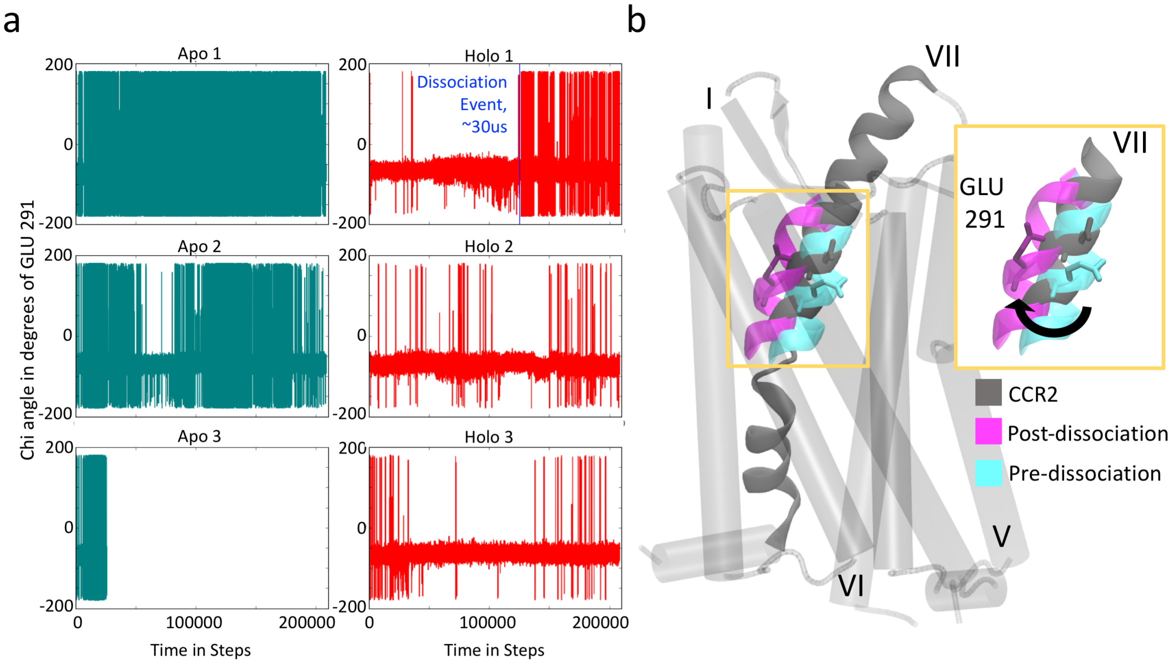


Figure 2.15: A) Chi angle of Glu 291 over simulation time. Dissociation event noted by blue line. B) Conformation of CCR2 before (teal) and after (purple) ligand dissociation.

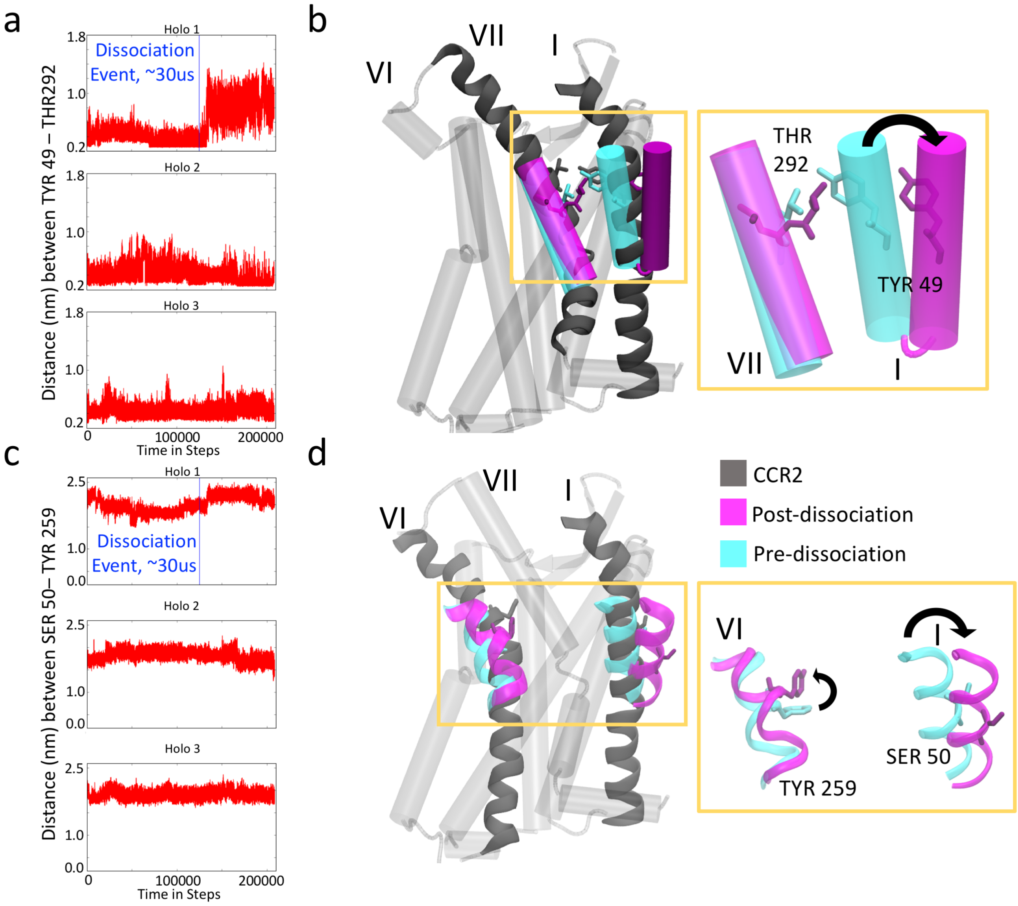


Figure 2.16: Orthosteric ligand dissociation breaks the hydrogen bond between key ligand binding residues. A) Distance between residues TYR 49$^{1.39}$ - Thr 292 over simulation time. Dissociation event noted by blue line. B) Conformation of CCR2 before (teal) and after (purple) ligand dissociation. C) The distance between Ser 50$^{1.40}$ and Tyr 259$^{6.51}$ over simulation time. This distance is also a contributor to holo TIC 1, and shows the same outward movement of helix I. There is a slight decrease in distance between the residue pair, followed by the same lag time of 3 μs, and finally an increase in distance as the extracellular end of helix I bends away from the helical bundle. D) Conformation of CCR2 before (teal) and after (purple) ligand dissociation.

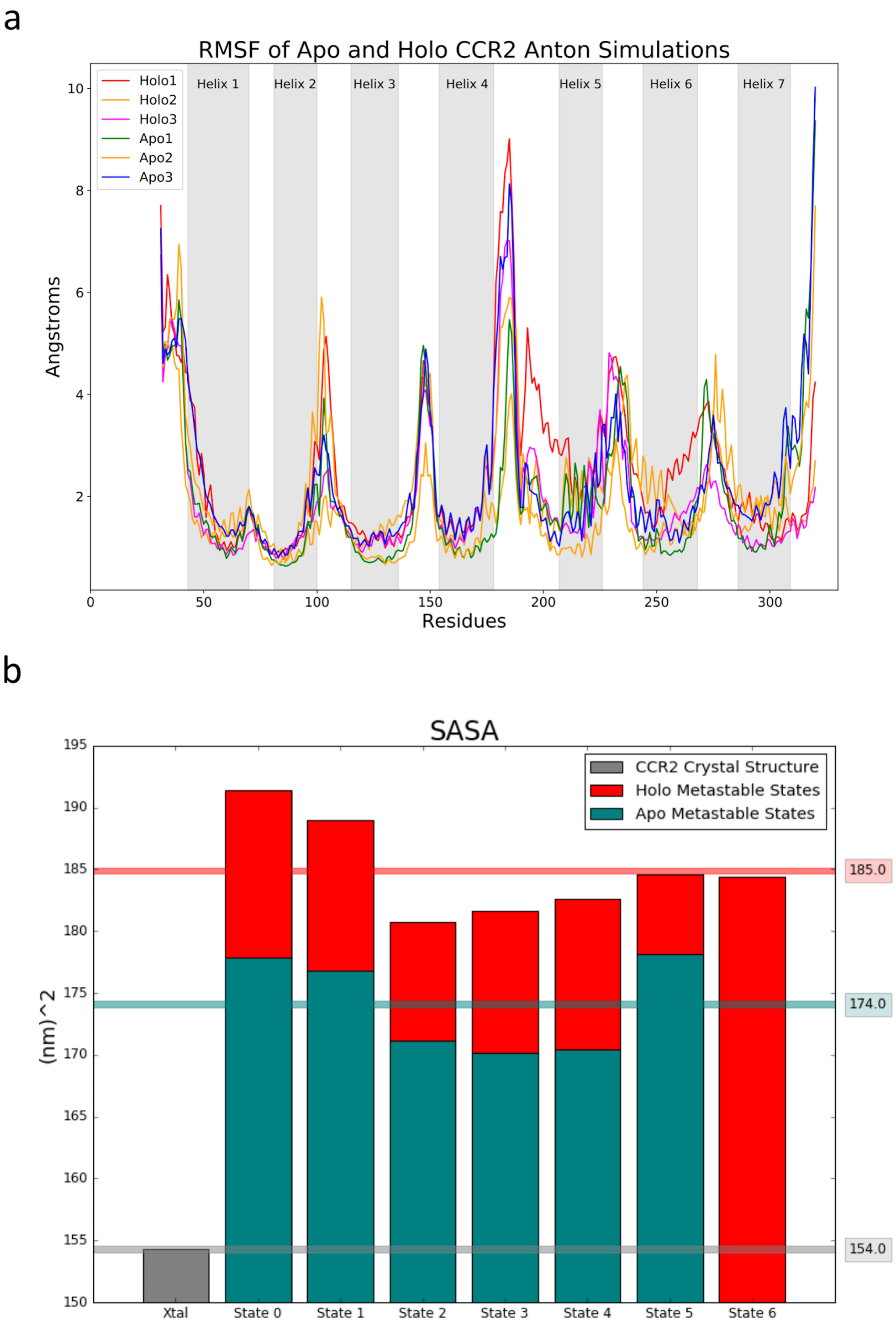


Figure 2.17: A) RMSF of each residue for individual simulations of apo and holo CCR2. B) SASA for each metastable macrostate of apo and holo, compared to the CCR2 crystal structure.

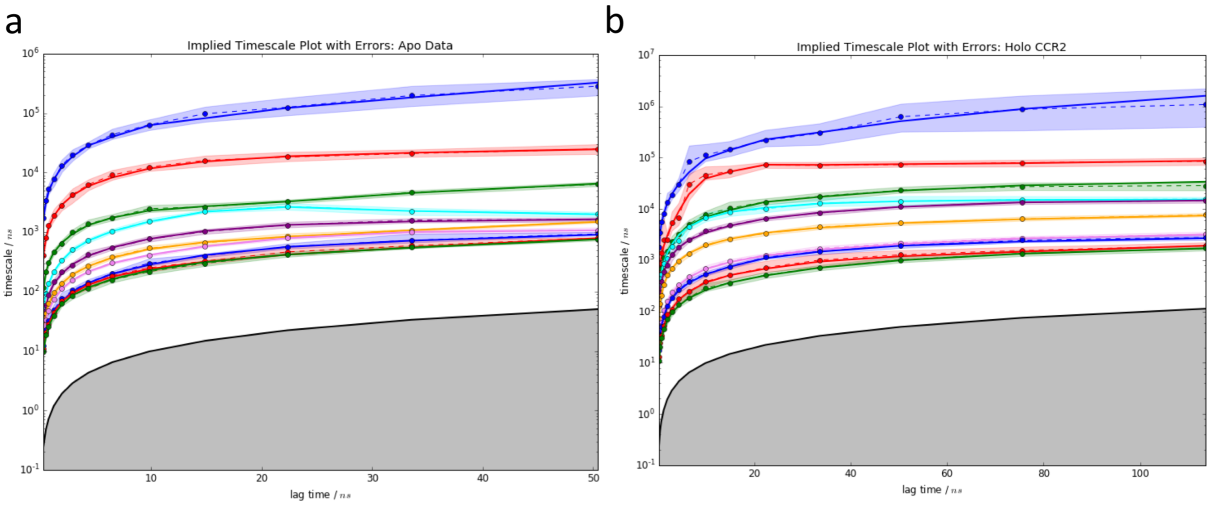


Figure 2.18: Implied timescale plots for A) apo and B) holo CCR2.

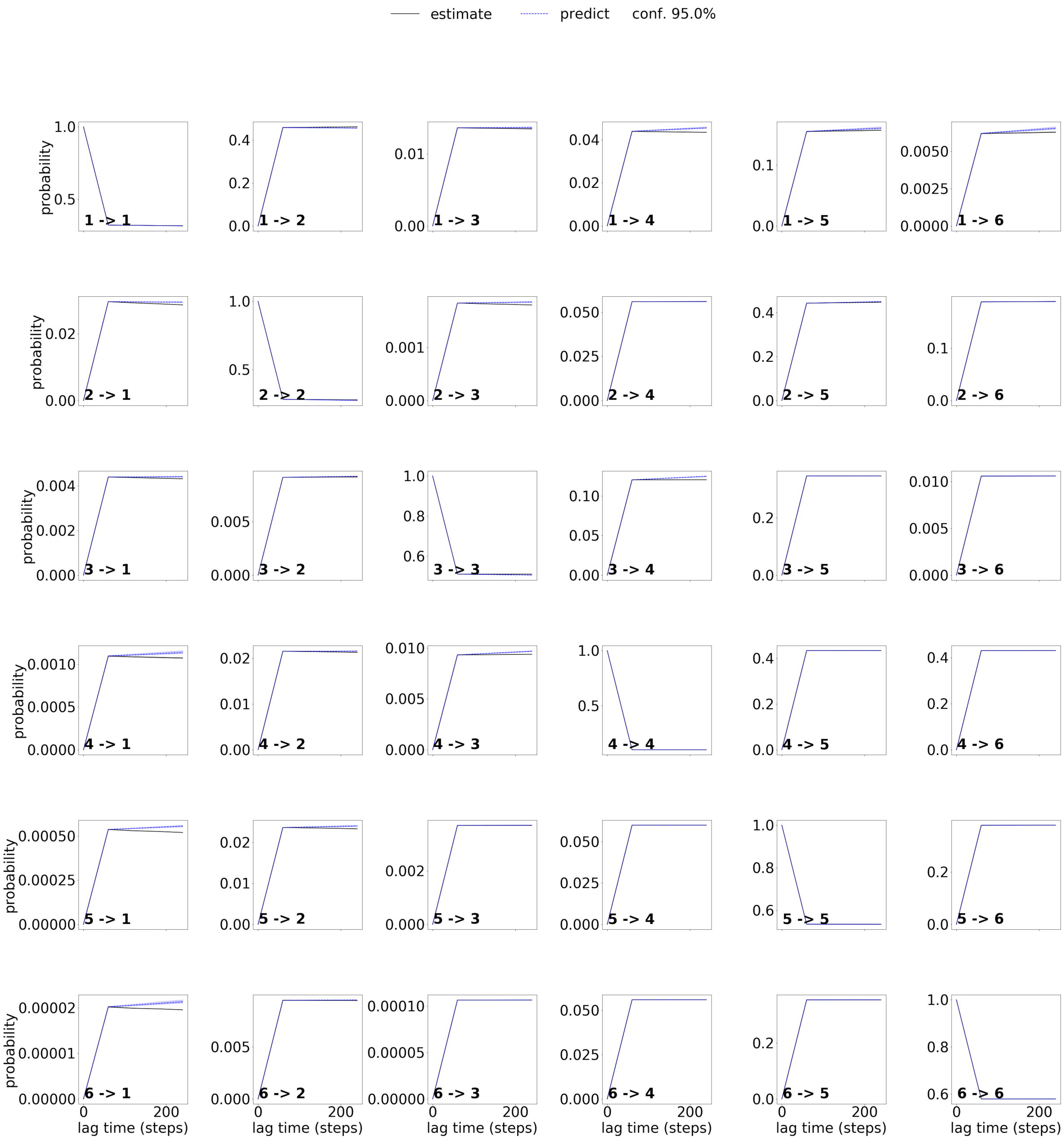


Figure 2.19: CK Test for apo CCR2.

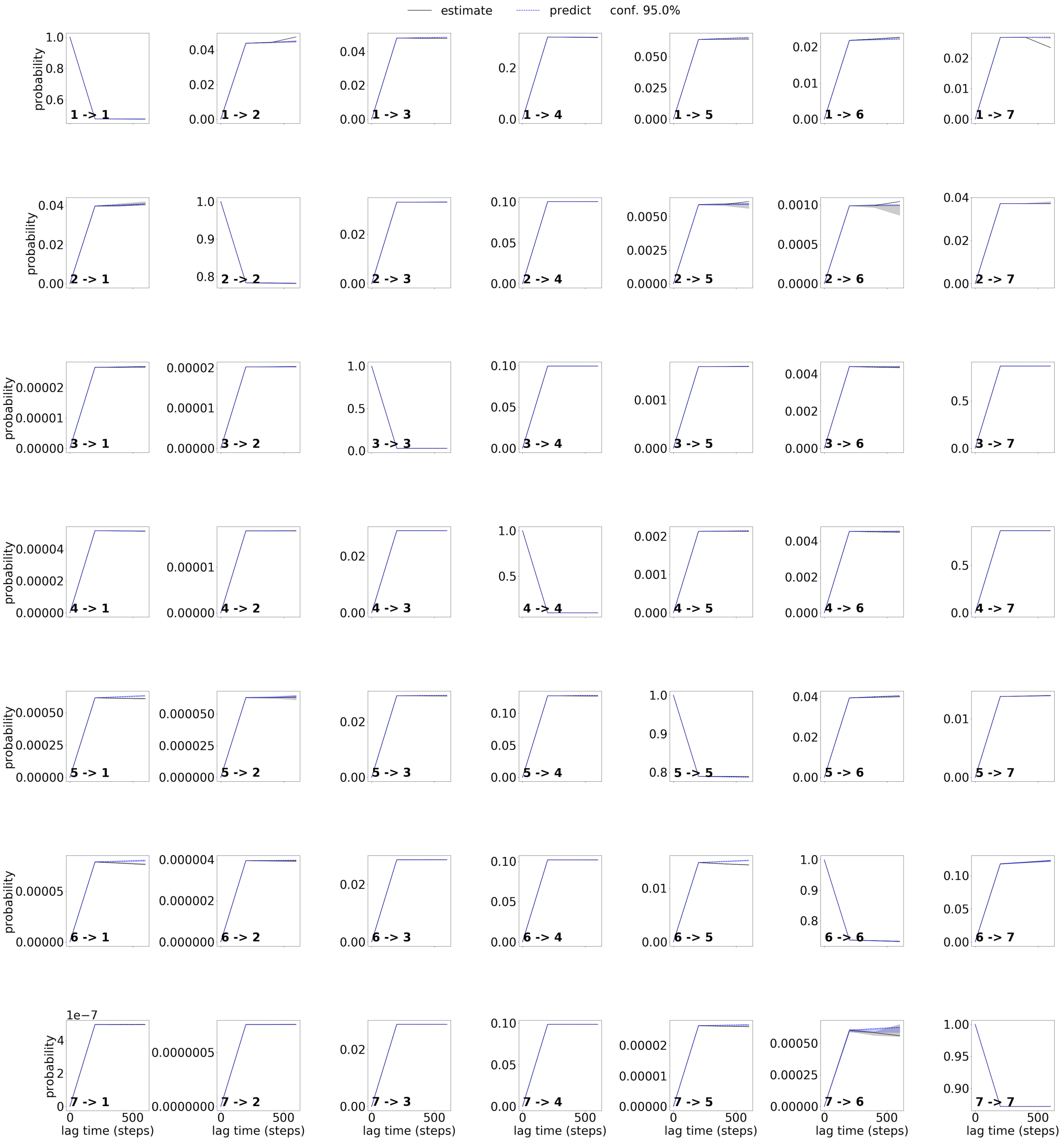


Figure 2.20: CK Test for holo CCR2

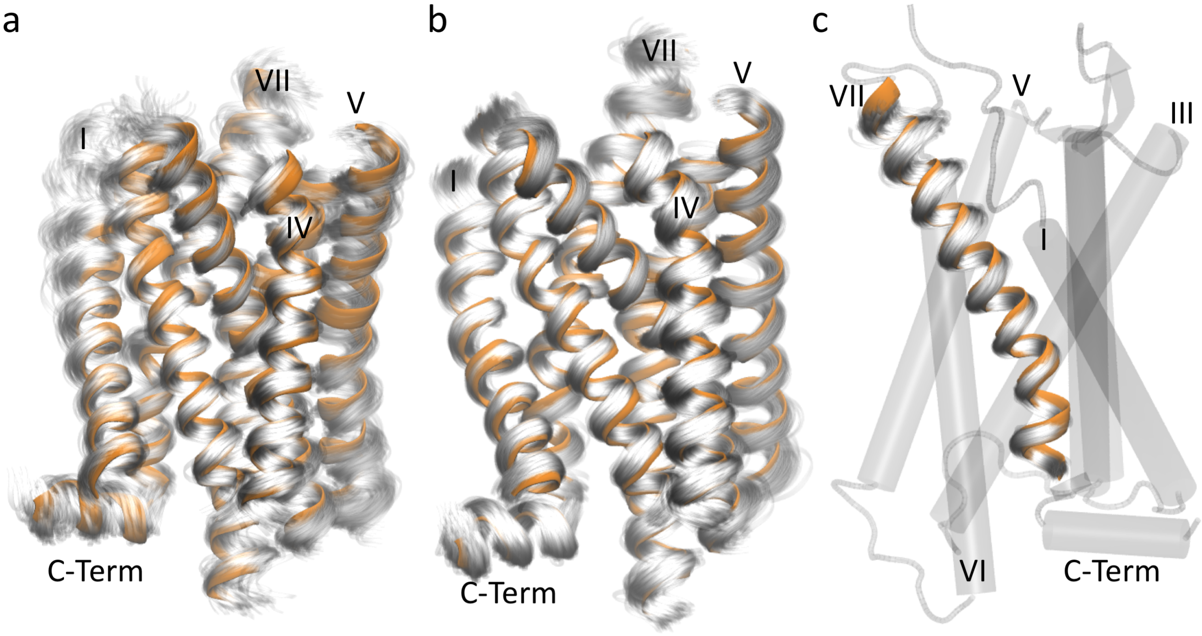


Figure 2.21: The centroid of apo macrostate A, in orange, is compared to A) the 214 other microstate centroids in white (average RMSD of alpha helices from representative structure: 2.01 Å, standard deviation 0.50 Å), and B) the 1,024 frames from its microstate in white (average RMSD of alpha helices from centroid: 0.872 Å, standard deviation 0.139 Å). C) Helix VII of the same microstate (average RMSD from centroid: 1.035 Å, standard deviation 0.259 Å.)

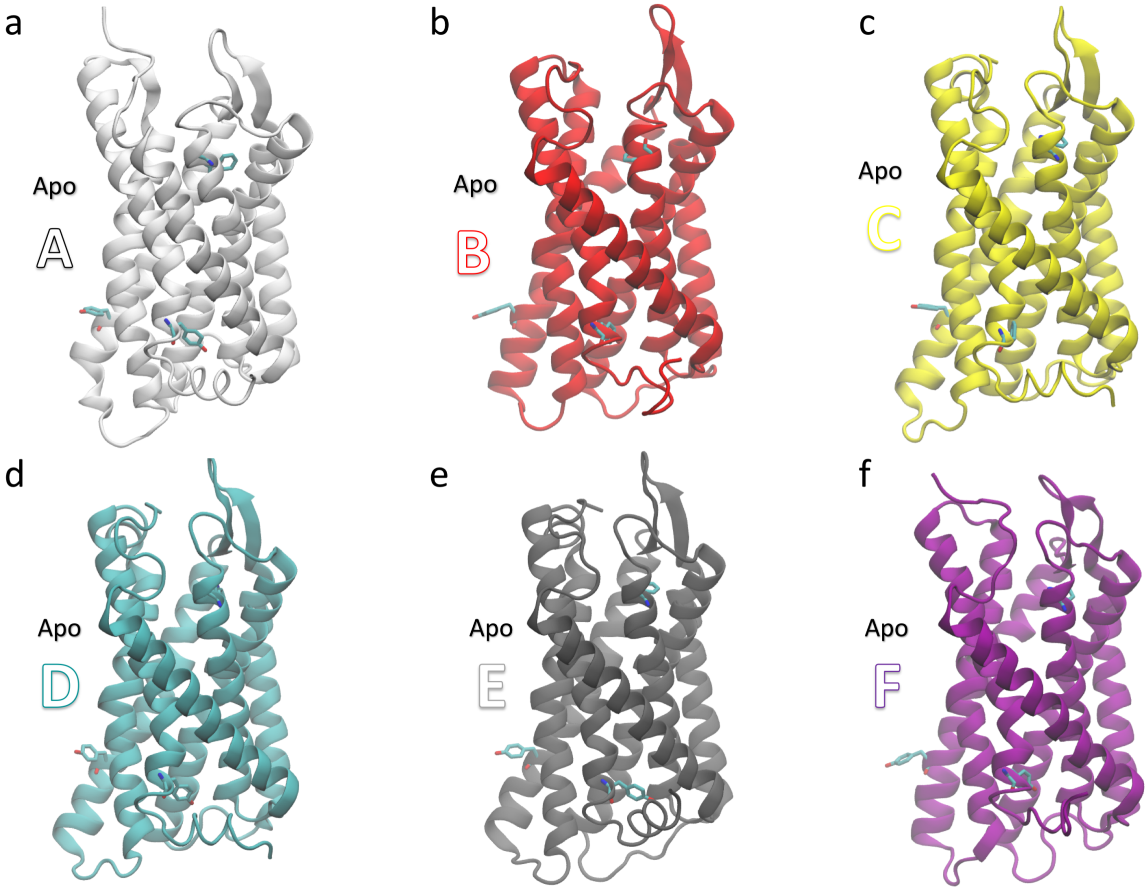


Figure 2.22: The apo macrostates.

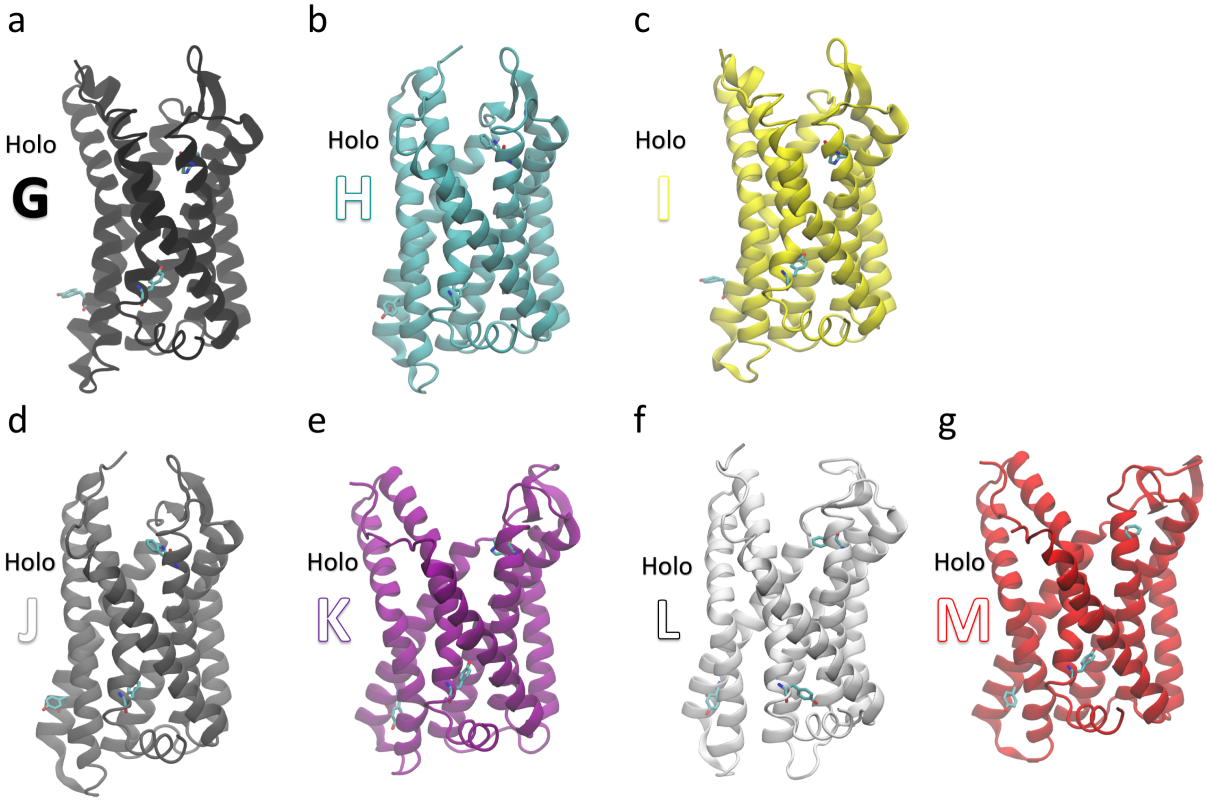


Figure 2.23: The holo macrostates.

References

(1) Ben-Baruch, A. *Cancer and Metastasis Reviews* **2006**, *25*, 357–371.

(2) O'Connor, T.; Borsig, L.; Heikenwalder, M. *Endocrine, Metabolic and Immune Disorders - Drug Targets* **2015**, *15*, 105–118.

(3) Solomon, M.; Balasa, B.; Sarvetnick, N. *Autoimmunity* **2010**, *43*, 156–163.

(4) Scholten, D. J.; Canals, M.; Mussang, D.; Roumen, L.; Smit, M.; Wijtmans, M.; de Graaf, C.; Vischer, H.; Leurs, R. *Br. J. Pharmacol* **2012**, *165*, 1617–1643.

(5) Lim, S.; Yuzhalin, A.; Gordon-Weeks, A.; Muschel, R. *Oncotarget* **2016**, *7*, 28697–710.

(6) Solari, R.; Pease, J. E.; Begg, M. *Eur. J. Pharmacol.* **2015**, *746*, 363–367.

(7) Horuk, R. *Nature reviews drug discovery* **2009**, *8*, 23–33.

(8) Shahlaei, M.; Fassihi, A.; Papaleo, E.; Pourfarzam, M. *Chemical biology & drug design* **2013**, *82*, 534–545.

(9) Chavan, S.; Pawar, S.; Singh, R.; Sobhia, M. E. *Molecular Diversity* **2012**, *16*, 401–413.

(10) Kothandan, G.; Gadhe, C. G.; Cho, S. J. *PloS one* **2012**, *7*, e32864.

(11) Zheng, P.; Zeng, B.; Zhou, C.; Liu, M.; Fang, Z.; Xu, X.; Zeng, L.; Chen, J.; Fan, S.; Du, X.; Zhang, X.; Yang, D.; Yang, Y.; Meng, H.; Li, W.; Melgiri, N. D.; Licinio, J.; Wei, H.; Xie, P. *Molecular Psychiatry* **2016**, DOI: 10.1038/mp.2016.44.

(12) Latorraca, N. R.; Venkatakrishnan, A. J.; Dror, R. O. *Chemical Reviews* **2017**, *117*, 139–155.

(13) Venkatakrishnan, A. J.; Deupi, X.; Lebon, G.; Tate, C. G.; Schertler, G. F.; Babu, M. M. *Nature* **2013**, *494*, 185–194.

(14) Zhang, Q.; Zhou, M.; Zhao, L.; Jiang, H.; Yang, H. *Biochemistry* **2018**, DOI: 10.1021/acs.biochem.8b00146.

(15) Katritch, V.; Cherezov, V.; Stevens, R. *Annual review of pharmacology and toxicology* **2013**, *53*, 531–556.

(16) Manglik, A.; Kim, T. H.; Masureel, M.; Altenbach, C.; Yang, Z. Y.; Hilger, D.; Lerch, M. T.; Kobilka, T. S.; Thian, F. S.; Hubbell, W. L.; Prosser, R. S.; Kobilka, B. K. *Cell* **2015**, *162*, 1431–1431.

(17) Katritch, V.; Cherezov, V.; Stevens, R. C. *Trends Pharmacol. Sci.* **2012**, *33*, 17–27.

(18) Malik, R. U.; Ritt, M.; DeVree, B. T.; Neubig, R. R.; Sunahara, R. K.; Sivaramakrishnan, S. *J. Biol. Chem.* **2013**, *288*, 17167–17178.

(19) Yao, X. J.; Velez Ruiz, G.; Whorton, M. R.; Rasmussen, S. G. F.; DeVree, B. T.; Deupi, X.; Sunahara, R. K.; Kobilka, B. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 9501–9506.

(20) Nygaard, R.; Zou, Y. Z.; Dror, R. O.; Mildorf, T. J.; Arlow, D. H.; Manglik, A.; Pan, A. C.; Liu, C. W.; Fung, J. J.; Bokoch, M. P.; Thian, F. S.; Kobilka, T. S.; Shaw, D. E.; Mueller, L.; Prosser, R. S.; Kobilka, B. K. *Cell* **2013**, *152*, 532–542.

(21) Bockenhauer, S.; Furstenberg, A.; Yao, X. J.; Kobilka, B. K.; Moerner, W. E. *J. Phys. Chem. B.* **2011**, *115*, 13328–13338.

(22) Shaw, D. E.; Grossman, J. P.; Bank, J. A.; Batson, B.; Butts, J. A.; Chao, J. C.; Deneroff, M. M.; Dror, R. O.; Even, A.; Fenton, C. H.; Forte, A.; Gagliardo, J.; Gill, G.; Greskamp, B.; Ho, C. R.; Ierardi, D. J.; Iserovich, L.; Kuskin, J. S.; Larson, R. H.; Layman, T.; Lee, L. S.; Lerer, A. K.; Li, C.; Killebrew, D.; Mackenzie, K. M.; Mok, S. Y. H.; Moraes, M. A.; Mueller, R.; Nociolo, L. J.; Peticolas, J. L.; Quan, T.; Ramot, D.; Salmon, J. K.; Scarpazza, D. P.; Ben Schafer, U.; Siddique, N.; Snyder, C. W.; Spengler, J.; Tang, P. T. P.; Theobald, M.; Toma, H.; Towles, B.; Vitale, B.; Wang, S. C.; Young, C. **2014**, DOI: 10.1109/SC.2014.9.

(23) Taylor, B. C.; Lee, C. T.; Amaro, R. E. Data from "Structural basis for ligand modulation of the CCR2 conformational landscape.", 2018.

(24) Bowman, G. R.; Voelz, V. a.; Pande, V. S. *Current Opinion in Structural Biology* **2011**, *21*, 4–11.

(25) Swope, W. C.; Pitera, J. W. *J. Phys. Chem. B* **2004**, *108*, 6571–6581.

(26) Singhal, N.; Snow, C. D.; Pande, V. S. *Journal of Chemical Physics* **2004**, DOI: 10.1063/1.1738647.

(27) Malmstrom, R. D.; Lee, C. T.; Van Wart, A. T.; Amaro, R. E. *Journal of Chemical Theory and Computation* **2014**, DOI: 10.1021/ct5002363.

(28) Amaro, R. E.; Mulholland, A. J. *Nature Reviews Chemistry* **2018**, *2*, 0148.

(29) Amaro, R. E.; Baudry, J.; Chodera, J.; Demir, Ö.; McCammon, A.; Miao, Y.; Smith, J. C. *Biophysical Journal* **2018**, *114*, 2271–2278.

(30) Bowman, G. R.; Pande, V. S.; Noe, F., *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*; Springer Netherlands: 2014; Vol. 797.

(31) Noe, F.; Horenko, I.; Schütte, C.; Smith, J. C. *The Journal of Chemical Physics* **2007**, *126*, 155102.

(32) Prinz, J. H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schtte, C.; No??, F. *Journal of Chemical Physics* **2011**, *134*, DOI: 10.1063/1.3565032.

(33) Schutte, C.; Fischer, A.; Huisinga, W.; Deuflhard, P. *Journal of Computational Physics* **1999**, *151*, 146–168.

(34) Noe, F.; Nüske, F. *Multiscale Modeling & Simulation* **2013**, *11*, 635–655.

(35) Schwantes, C. R.; Pande, V. S. *J. Chem. Theory Comput.* **2013**, *9*, 2000–2009.

(36) Perez-Hernandez, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noe, F. *J. Chem. Phys.* **2013**, *139*, 015102.

(37) Lebon, G.; Warne, T.; Edwards, P. C.; Bennett, K.; Langmead, C. J.; Leslie, A. G. W.; Tate, C. G. *Nature* **2011**, *474*, 521–525.

(38) Burg, J. S.; Ingram, J. R.; Venkatakrishnan, A. J.; Jude, K. M.; Dukkipati, A.; Feinberg, E. N.; Angelini, A.; Waghray, D.; Dror, R. O.; Ploegh, H. L.; Garcia, K. C. *Science* **2015**, *347*, 1113–1117.

(39) Tan, Q.; Zhu, Y.; Li, J.; Chen, Z.; Han, G. W.; Kufareva, I.; Li, T.; Ma, L.; Fenalti, G.; Li, J.; Zhang, W.; Xie, X.; Yang, H.; Jiang, H.; Cherezov, V.; Liu, H.; Stevens, R. C.; Zhao, Q.; Wu, B. *Science* **2013**, *341*, 1387–1390.

(40) Oswald, C.; Rappas, M.; Kean, J.; Doré, A. S.; Errey, J. C.; Bennett, K.; Deflorian, F.; Christopher, J. A.; Jazayeri, A.; Mason, J. S.; Congreve, M.; Cooke, R. M.; Marshall, F. H. *Nature* **2016**, *540*, 462–465.

(41) Qin, L.; Kufareva, I.; Holden, L. G.; Wang, C.; Zheng, Y.; Zhao, C.; Fenalti, G.; Wu, H.; Han, G. W.; Cherezov, V.; Abagyan, R.; Stevens, R. C.; Handel, T. M. *Science* **2015**, *347*, 1117–1122.

(42) Caliman, A.; Swift, S.; Wang, Y.; Miao, Y.; McCammon, J. *Protein Science: A Publication of the Protein Society* **2015**, *24*, 1004–1012.

(43) Jastrzebska, B.; Palczewski, K.; Golczak, M. *The Journal of biological chemistry* **2011**, *286*, 18930–7.

(44) Angel, T. E.; Chance, M. R.; Palczewski, K. *Proceedings of the National Academy of Sciences of the United States of America* **2009**, *106*, 8555–8560.

(45) Choe, H.-W.; Kim, Y. J.; Park, J. H.; Morizumi, T.; Pai, E. F.; KrauSS, N.; Hofmann, K. P.; Scheerer, P.; Ernst, O. P. *Nature* **2011**, *471*, 651–655.

(46) Huang, W.; Manglik, A.; Venkatakrishnan, A. J.; Laeremans, T.; Feinberg, E. N.; Sanborn, A. L.; Kato, H. E.; Livingston, K. E.; Thorsen, T. S.; Kling, R. C.; Granier, S.; Gmeiner, P.; Husbands, S. M.; Traynor, J. R.; Weis, W. I.; Steyaert, J.; Dror, R. O.; Kobilka, B. K. *Nature* **2015**, *524*, 315–321.

(47) Yuan, S.; Filipek, S.; Palczewski, K.; Vogel, H. *Nature communications* **2014**, *5*, 4733.

(48) Rinne, A.; Birk, A.; Bünemann, M. *Proceedings of the National Academy of Sciences of the United States of America* **2013**, *110*, 1536–41.

(49) Katritch, V.; Fenalti, G.; Abola, E. E.; Roth, B. L.; Cherezov, V.; Stevens, R. C. *Trends in biochemical sciences* **2014**, *39*, 233–44.

(50) Yuan, S.; Vogel, H.; Filipek, S. *Angewandte Chemie International Edition* **2013**, *52*, 10112–10115.

(51) Vickery, O. N.; Carvalheda, C. A.; Zaidi, S. A.; Pisliakov, A. V.; Katritch, V.; Zachariae, U. *Structure* **2018**, *26*, 171–180.

(52) Miao, Y.; Caliman, A.; McCammon, J. *Biophysical Journal* **2015**, *108*, 1796–1806.

(53) Berkhout, T. A.; Blaney, F. E.; Bridges, A. M.; Cooper, D. G.; Forbes, I. T.; Gribble, A. D.; Groot, P. H. E.; Hardy, A.; Ife, R. J.; Kaur, R.; Moores, K. E.; Shillito, H.; Willetts, J.; Witherington, J. *Journal of medicinal chemistry* **2003**, *46*, 4070–86.

(54) Hall, S. E.; Mao, A.; Nicolaidou, V.; Finelli, M.; Wise, E. L.; Nedjai, B.; Kanjanapangka, J.; Harirchian, P.; Chen, D.; Selchau, V.; Ribeiro, S.; Schyler, S.; Pease, J. E.; Horuk, R.; Vaidehi, N. *Molecular pharmacology* **2009**, *75*, 1325–36.

(55) Cherney, R.; Nelson, D.; Lo, Y.; Yang, G.; Scherle, P.; Jezak, H.; Solomon, K.; Carter, P.; Decicco, C. *Bioorg Med Chem Lett.* **2008**, *18*, 5063–5.

(56) Kozakov, D.; Grove, L.; Hall, D.; Bohnuud, T.; Mottarella, S.; Luo, L.; Xia, B.; Beglov, D.; Vajda, S. *Nature Protocols* **2015**, *10*, 733–755.

(57) Bowman, G. R.; Bolin, E. R.; Hart, K. M.; Maguire, B. C.; Marqusee, S. *Proceedings of the National Academy of Sciences of the United States of America* **2015**, *112*, 2734–2739.

(58) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J. H.; Noé, F. *Journal of Chemical Theory and Computation* **2015**, *11*, 5525–5542.

(59) Dasse, O. A.; Evans, J. L.; Zhai, H.-X.; Zou, D.; Kintigh, J. T.; Chan, F.; Hamilton, K.; Hill, E.; Eckman, J. B.; Higgins, P. J.; Volosov, A.; Collart, P.; Nicolas, J.-M.; Kondru, R.; Schwartz, C. *Letters in Drug Design and Discovery* **2007**, *4*, 263–271.

(60) Carter, P. H.; Brown, G. D.; Cherney, R. J.; Batt, D. G.; Chen, J.; Clark, C. M.; Cvijic, M. E.; Duncia, J. V.; Ko, S. S.; Mandlekar, S.; Mo, R.; Nelson, D. J.; Pang, J.; Rose, A. V.; Santella, J. B.; Tebben, A. J.; Traeger, S. C.; Xu, S.; Zhao, Q.; Barrish, J. C. *ACS Med Chem Lett.* **2015**, *6*, 439–444.

(61) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. *Journal of Computational Chemistry* **2008**, *29*, 1859–1865.

(62) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D. *Journal of Computational Chemistry* **2010**, *31*, 671–690.

(63) Huang, J.; Mackerell, A. D. *Journal of Computational Chemistry* **2013**, *34*, 2135–2145.

(64) Humphrey W. Dalke, A.; Schulten, K. *J. Molec. Graphics* **1996**, *14*, 33–38.

(65) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. *Journal of Chemical Theory and Computation* **2011**, *7*, 3412–3419.

(66) Pande, V. S.; Beauchamp, K.; Bowman, G. R. *Methods* **2010**, *52*, 99–105.

(67) Prinz, J.-H.; Keller, B.; Noe, F. *Phys. Chem. Chem. Phys.* **2011**, *13*, 16912–16927.

(68) Senne, M.; Trendelkamp-Schroer, B.; Mey, A. S. J. S.; Schütte, C.; Noe, F. *Journal of Chemical Theory and Computation* **2012**, *8*, 2223–2238.

(69) Cronkite-Ratcliff, B.; Pande, V. *Bioinformatics* **2013**, *29*, 950–952.

(70) Nüske, F.; Keller, B. G.; Pérez-Hernández, G.; Mey, A. S. J. S.; Noe, F. *Journal of Chemical Theory and Computation* **2014**, *10*, 1739–1752.

(71) Noe, F.; Wu, H.; Prinz, J.-H.; Plattner, N. *The Journal of Chemical Physics* **2013**, *139*, 184114.

Chapter 3

Best Practices for Analysing Microbiomes

3.1 Abstract

Complex microbial communities shape the dynamics of various environments, ranging from the mammalian gastrointestinal tract to the soil. Advances in DNA sequencing technologies and data analysis have provided drastic improvements in microbiome analyses, for example, in taxonomic resolution, false discovery rate control and other properties over earlier methods. In this Review, we discuss the best practices for performing a microbiome study, including experimental design, choice of molecular analysis technology, methods for data analysis and the integration of multiple -omics data sets. We focus on recent findings that suggest that operational taxonomic unit-based analyses should be replaced for new methods that are based on exact sequence variants, methods for integrating metagenomic and metabolomic data and issues surrounding compositional data analysis, where advances have been particularly rapid. We note that although some of these approaches are new, it is important to keep sight of the classic issues that arise during experimental design and relate to research reproducibility. We describe how keeping these issues in mind allow researchers to obtain more insight from their microbiome data sets.

3.2 Introduction

Advances in DNA sequencing technologies have transformed our capacity to investigate the composition and dynamics of complex microbial communities that inhabit diverse environments from mammalian gastrointestinal tracts to deep ocean sediments. These developments have led to vast increases in the number of microbiome studies being performed in many fields of science, from clinical research to biotechnology. With this transformation, researchers are often left holding massive amounts of data and confronted with a bewildering array of computational tools and methods for analyzing their data. Conducting a robust experiment is not trivial in microbiome research, and as with any study, experimental methods, environmental factors and analysis methods can impact results. Standards for data collection and analysis are still emerging in the field, yet many compelling results can be achieved with current practices.

Microbiome analysis methods and standards are rapidly advancing. In particular, recommendations concerning differential abundance testing, using exact sequence variants rather than operational taxonomic units (OTUs) and performing correlation analysis have evolved quickly in the past two years. We can expect a similar pace of development in several other areas, including metagenomic taxonomy and functional assignment; integration of data sets from multiple sequencing runs; and further improvement in machine learning, compositional data analysis and multi-omics analyses. However, many of the most fundamental issues that concern microbiome studies arise from statistical and experimental design issues. The most important challenge for the field is to integrate new approaches that are unique to microbiome studies, while remembering standard practices that are broadly applicable to all scientific studies.

Although it is impossible to be fully comprehensive in one article, this Review aims to provide straightforward guidelines for designing, executing and analyzing a microbiome experiment, with particular focus on human, model organism and environmental microbiomes. We direct the reader to more specialized reviews on specific topics where these exist.

3.3 Experimental design.

Designing an experiment that generates meaningful data is an important first step in your analysis. Typical scientific questions, such as case-control and longitudinal interventions or studies can all be studied in the context of the microbiome. Researchers can identify potential differences in microbial community structure and composition, genetics, or functional variation either between separate communities or over time. Notably, the general approach to microbiome analysis is applicable regardless of sample origin (Box 1). However, specific details of the analysis may depend on the sample origin; for example, 16S ribosomal RNA (rRNA) amplicon regions have variable success among different sample types in recapitulating results from metagenomic sequencing data [1].

The other primary considerations when assessing different sample types are experimental design and sample collection. We have observed many confounding issues during human microbiome studies and therefore we emphasize the importance of experimental design when performing these studies, though often many of the same considerations apply to animal models and environmental samples (Box 2).

Meticulous experimental design is crucial for obtaining accurate and meaningful results from microbiome studies. Many confounding factors, if not controlled, can obscure patterns in microbiome data (Figure 3.1). Careful curation of metadata, appropriate controls including extraction and reagent blanks, and thoughtful study designs that isolate and interrogate variables of interest are all essential. First, the scope of the experiment must be defined, and an appropriate experimental design selected for the question of interest. For example, cross-sectional studies are useful for finding differences in microbial communities between different human populations, such as healthy individuals and those with diseases, or individuals living in different geographic regions. However, due to the large variation in the microbiome between individuals and the profound influence of lifestyle [2, 3], diet [4], medication [5, 6] and physiology, differences between populations may arise from factors other than the disease of interest. For example, initial reports of changes in the microbiome in diabetic individuals were confounded by effects of the drug metformin [5].

Longitudinal studies, especially prospective longitudinal studies that collect baseline samples before disease onset, can help resolve these issues, although they are more expensive. For ease in downstream statistical analyses, longitudinal studies should plan the timing of sample collection carefully: for human studies, this may mean collecting samples at identical time points for each subject. Interestingly, community instability rather than the specific taxa present at a single time point can be a strong predictor of disease activity [7]. For example, individuals with inflammatory bowel disease (IBD) exhibit greater

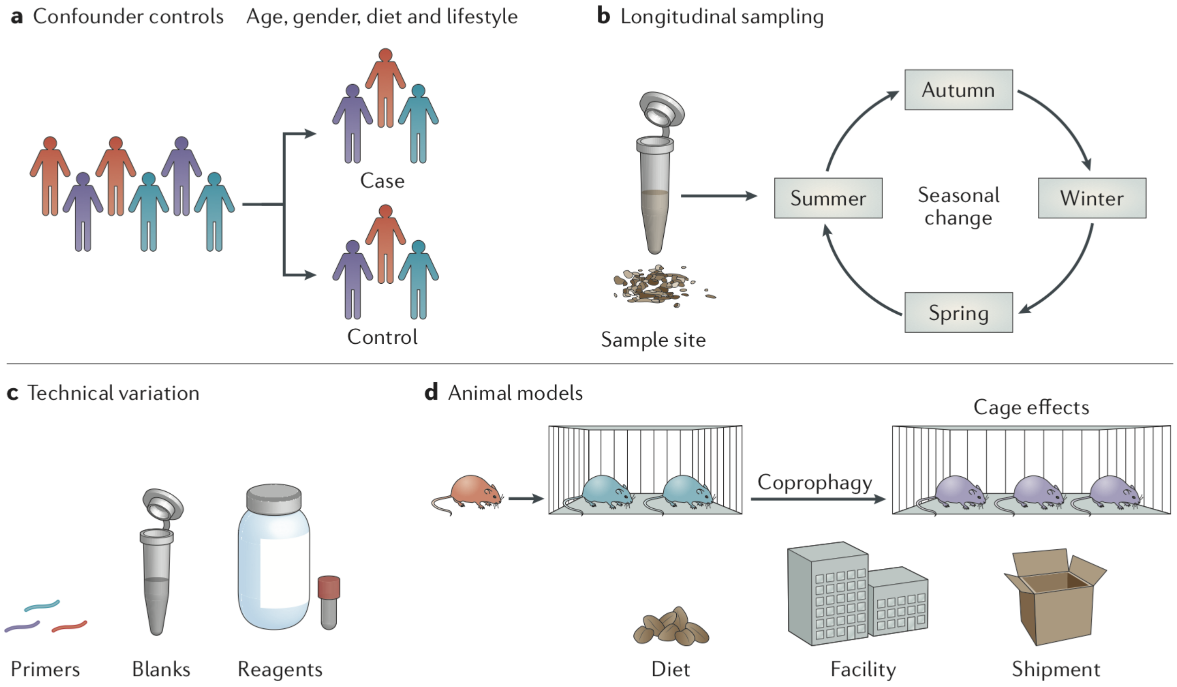


Figure 3.1: Experimental design considerations for microbiome experiments. Conducting a robust microbiome experiment warrants careful attention to numerous factors. Stratification by potential confounders (for example, age, gender, diet, lifestyle factors and medications) can help resolve differences in microbiota between groups of interest which might otherwise be masked by a confounder-effect [5]. Longitudinal studies are especially powerful as they both control for confounding factors and allow for the assessment of community stability [7]. Similar considerations apply to animal studies, though the additional impact of coprophagy must be addressed in experimental design. For all studies, standardizing technical factors and sample processing is essential to control for variation introduced by kit reagents, primers, sample storage, and other factors. The collection and curation of metadata about all aspects of each sample, from clinical variables to sample processing, is crucial for data interpretation; without metadata, it is difficult to draw meaningful conclusions from sequencing data.

microbiome fluctuations than control cohorts [7]. Interventional studies, including double blind randomized control studies, are especially useful for identifying specific effects of a course of treatment on the microbiome and disease state. Designing a study with an analysis plan and specific experimental questions to interrogate can help determine the sample size. For example, to test the effects of a new broad-spectrum antibiotic on the mouse gut microbiota, more samples may be required to look at specific taxa shifts compared to assessing how alpha diversity (a quantitative measure of community diversity) changes with antibiotic treatment, as baseline microbiota composition varies between mice. The antibiotic may be expected to decrease alpha diversity in all mice, but it could perturb their microbial community composition in different ways. For any study design, appropriate methods to assess statistical power should be employed in order to discern technical variability and real biological results [8]. However, statistical power and effect size analysis remains a challenge in microbiome research [9]. Some methods that are currently used for power and effect size analysis are based on PERMANOVA [8], Dirichlet Multinomial [10] or random forest analysis [11]. As these methods are further developed to integrate metagenomics, metatranscriptomics, metaproteomics and metabolomics data sets, study design and selection of appropriate sample size will also improve. For specific experimental design considerations, we recommend reviewing the design of other successful studies with similar sample types and desired outcomes. We expand on important considerations for microbiome experimental design below.

3.4 Box 1. Good working practices.

It is crucial for microbiome analyses to be reproducible. Similar microbiome studies can often have conflicting results, and without proper documentation of sample collection, data processing, and analysis methods, it is difficult to re-examine the data and reconcile these differences. As the field evolves, it will be necessary to revisit early experiments and potentially re-analyze the data with updated tools. Reproducibility is paramount for this process to be possible and efficient. When collecting samples, details of the collection process should be recorded in the experimental metadata to ensure that as much variability as possible is accounted for. Additionally, the Genome Standards Consortium minimum information standards (MIxS) for marker genes (MIMARKS) and metagenomes (MIMS) [12] should be adhered to.

These unified standards enable comparisons across data sets. During bioinformatics processing, researchers should track all of the commands that they ran and all software versions that they used, and deposit their raw data and metadata in public repositories. We recommend using tools such as Jupyter Notebooks (http://jupyter.org) or R Markdown (https://rmarkdown.rstudio.com/) to facilitate this, and then storing the notebooks in a revision control management system such as GitHub (https://github.com). Some software packages, such as QIIME 2 [13] (https://qiime2.org) and Galaxy (https://usegalaxy.org/) automatically track this information for researchers through an integrated data provenance tracking system. Qiita (http://qiita.microbio.me) and EBI (http://www.ebi.ac.uk/) are powerful meta-analysis and data archiving tools, respectively, and when combined allow a researcher to analyze their microbiome data in the context of tens of thousands of other samples, which enables the data to be re-used by future researchers.

3.5 Box 2. Considerations for different microbiomes.

Although microbiome data analysis methods are widely applicable to many sample types and environments, experimental design, and method selection require careful consideration for different sample types. First, one must consider the composition of the sample and feasibility of use for different methods. For samples that are heavily contaminated with non-microbial DNA, such as tissue, shotgun metagenomic sequencing may not be feasible without non-microbial DNA depletion. Depending on the experimental question, samples heavily contaminated with relic DNA from dead microorganisms, such as soil, may require physical removal of relic DNA by propidium monoazide [14] or other methods prior to DNA extraction. The amount of sample to collect is also determined by sample type. Whereas a high biomass fecal sample may only require a swab, samples with low microbial density may necessitate larger volumes and potentially concentration for sufficient DNA extraction. For example, ocean microbiome samples are usually large volumes of water run though a filter to trap and concentrate the target organisms prior to DNA extraction [15]. Though in all cases appropriate controls should be included, low biomass environments, such as blood, spinal fluid or laboratory clean rooms, particularly necessitate controls that have gone through the entire sampling process to fully characterize contaminants. DNA contaminants can be found in numerous reagents, including swabs, DNA extraction kits and PCR reagents [16]. Furthermore, the method of sample preservation is both dictated by analysis method and sample type. For example, metatranscriptomics requires an RNAse inhibitor and metabolomics requires sample preservation that does not interfere with metabolite extraction or data collection. In addition to sampling considerations, study design and metadata collection also require careful tailoring to sample type and environment. For example, animal studies require an evaluation of co-housing cage effects and should stratify experimental groups into multiple cages. Fresh samples should be collected and the mouse of origin should be recorded in the metadata. Environmental samples require collection of metadata related to environmental conditions, such as pH, salinity, elevation, and depth for soil samples. The manner of collection is highly dependent on sample type and cannot be detailed for all possible samples in this Review. We recommend consulting well-validated protocols related to the sample type of interest. In any case, methods of collection, preservation, and storage should remain consistent across all samples in a study to avoid introducing confounding variation. Sample composition can be effected by outgrowth of certain microorganisms during storage at room temperature [17].

3.5.1 Defining controls and exclusion criteria.

Defining clear inclusion and exclusion criteria limits confounding covariates. For instance, variability in recovery time from antibiotics among individuals [18] suggests that individuals that were treated with antibiotics in the preceding 6 months should be excluded from most microbiome studies. Similarly, recovery of the skin microbiome after hand washing takes ˜2 hours[19]. In case-control experimental designs, controls must be appropriately selected and matched. Age and sex are common control criteria, despite the relatively weak effect of sex on most human microbiomes across body sites [20, 21], while other variables such as medication and diet are often more important confounders to control for. The relative effect sizes of these microbiome variables are still emerging[9]. Collection of comprehensive clinical data collection is crucial for identifying confounders that cannot be controlled. This topic has been extensively reviewed in Ref. [22]. Environmental studies must also account for similar confounders, as plot-to-plot variation is a widely recognized confounding phenomenon in the ecological literature that should be addressed with nested statistical tests [23].

3.5.2 Animal models.

The predominant animal models for studying the microbiome are rodents, such as mice. Other models with varying microbial complexity such as bobtail squid, insects or zebrafish are often useful for studying specific interactions between host and microorganisms (for example, how the microbiome and the host genetics influence each other) [24]. Nevertheless, rodents are often preferred because they are well-characterized and have many physiological similarities to humans. Rodent microbiome studies require particularly careful design. As rodents are coprophagic, cagemate fecal microbiomes become more homogenous over time, so experiments must be replicated across multiple cages to control for cage effects [25]. Parental effects also necessitate randomizing littermates between cages and allowing for normalization. Single-housing stresses mice [26], and is thus often technically or ethically infeasible. Even genetically identical rodents may differ in their microbiomes due to environmental factors including diet, litter, vendor, shipment and facility [27, 28]. Additionally, early life microbial exposures greatly impact the established microbiota and can influence immune system development [29]. Similar considerations apply to other co-housed model organisms, for example, zebrafish [30].

3.5.3 Technical variation.

Technical variability among experimental methods ranging from DNA extraction to sequencing is high [31, 32]. The same reagent kits must be used for all samples in a study [16], and multiple baseline samples should be collected to assess intrinsic variability among time points in longitudinal studies. Using blanks during sampling, DNA extraction, PCR and sequencing is essential for detecting contamination. Reads that are derived from microorganisms introduced as contaminants or that grow during shipping can sometimes be reduced during analysis [17], though samples should be at -80 řC when possible [33]. For field studies or other situations where freezing is not possible, ambient storage methods, such as 95% ethanol or commercial products such as RNAlater or the OMNIgene Gut kit can be used [34]. Mock communities (reference samples with a known composition) are useful for standardizing analyses [35], as is including the same standard specimens in each DNA sequencing run [36]. In general, reconciling microbiome data that were generated using different methods remains an unsolved challenge.

Depending on the scope of their experiment (which includes the overall experimental design, sample types and source, sequencing method, and other factors that are discussed below), researchers can aim to gain a broad, community-level overview of their samples, a detailed genomic-level understanding, or even characterization of the functional variation in microbial communities.

3.6 Sequencing targets and methods.

Different methods for surveying microbial communities, including marker gene, metagenome, and metatransciptome sequencing, can produce varying results. All widely-used methods have strengths and weaknesses, so the question, hypothesis, sample type and analysis goals should inform the choice of method (Table 1). Here we discuss the trade-offs between cost, robustness, resolution and difficulty for marker gene, metagenome and metatranscriptome sequencing. We outline the best workflow for each method in Figure 3.2. To attain a high-level, but low resolution overview, the preferred method is marker gene sequencing. Metagenomic sequencing provides more detail by analyzing the total DNA in a sample, allowing strain-level resolution and detection of genes that can provide information on molecular functions. We also discuss metatranscriptomic sequencing of total RNA, which is used to characterize gene expression in the microbial community.

Table 3.1: Pros and cons of genomic analyses for evaluating microbial communities

Method	Pros	Cons
Marker gene analysis	- Quick, simple and inexpensive sample preparation and analysis [13, 37] - Correlates well with genomic content [38–42] - Amenable to low biomass and highly host-contaminated samples - Large existing public data sets for comparison [22, 37, 47]	- No live/dead/active discrimination - Subject to amplification biases [43] - Choice of primers, and variable region magnifies biases [44–46] - Requires a priori knowledge of microbial community [48] - Resolution typically limited to genus level at best - Appropriate negative controls required - Functional information is limited [40, 41]
Whole metagenome analysis	- Can directly infer the relative abundance of microbial functional genes, microbial taxonomic and phylogenetic identity to species/strains level is attainable known organisms [49] - Does not assume knowledge of microbial community (that is, captures phage, viruses, plasmids, microbial eukaryotes, etc.) - No PCR-related biases - Can estimate in situ growth rates for target organisms with sequenced genomes [50] - Can allow assembly of population-averaged microbial genomes [51, 52] - Can be mined for novel gene families	- Relatively expensive, laborious, and complex sample preparation and analysis - Contamination from host-derived DNA and organelles may obscure microbial signatures - Viruses and plasmids are not typically well annotated by default pipelines - Deep sequencing depths are typically required relative to other methods - No live/dead/active discrimination - Population-averaged microbial genomes tend to be inaccurate due to assembly artifacts
Metatranscriptome analysis	- Can estimate which microorganisms in a community are actively transcribing when paired with marker gene analysis - Inherently discriminates between active live organisms versus dormant or dead microorganisms and extracellular DNA - Captures dynamic intra-individual variation51 - Directly evaluates microbial activity, including response to intervention/event exposure [57]	- Most expensive, laborious, and complex sample preparation and analysis [53] - Host mRNA contamination and rRNA must be removed [54–56] - Requires careful sample collection and storage - Data is biased toward organisms with high transcription rate - Requires paired DNA sequencing to decouple transcription rates from bacterial abundance changes

3.6.1 Marker gene analysis.

Marker gene sequencing uses primers that target a specific region of a gene of interest in order to determine microbial phylogenies of a sample. This region typically contains a highly variable region that can be used for detailed identification, flanked by highly conserved regions that can serve as binding sites for PCR primers. Marker gene amplification and sequencing (such as 16S rRNA for bacteria and archaea and internal transcribed spacer (ITS) for fungi) is a well-tested, fast and cost-effective method for obtaining a low resolution view of a microbial community. This approach works well for host DNA contaminated samples, such as tissue and low-biomass samples. However, because DNA sequences vary in these primer-amplified regions, primers do not have equal affinity for all possible DNA sequences, and consequently induce bias during PCR amplification. Other sources of inherent bias in marker gene sequencing include variable region selection, amplicon size [44], and the number of PCR cycles [43]. Low-biomass samples are particularly susceptible to bias introduced by over amplification as the PCR cycle number increases, contaminating microorganisms are increasingly over-represented [16]. Optimizing primer selection can help mitigate bias, but this requires a priori knowledge of microbial community composition to assess taxonomic resolution and coverage of the target community [48]. However, even well-optimized primers are often limited to genus level taxonomic resolution. Marker gene sequencing generally correlates well with genomic content [38–42] and is applicable to the broadest range of sample types and study designs.

3.6.2 Whole metagenome analysis.

Metagenomics is the method of sequencing all microbial genomes within a sample. Metagenomic sequencing yields more detailed genomic information and taxonomic resolution than marker gene sequencing alone, but it is relatively expensive to prepare, sequence and analyze the samples. This method captures all DNA present in the sample, including viral and eukaryotic DNA. Given adequate sequencing depth (the number of sequencing reads per sample), taxonomic resolution to species or strain level [49] and the assembly of whole microbial genomes from short DNA sequence reads is possible [51]. However, de novo annotation of functional genes is not possible in such settings. Metagenomic sequencing profiles

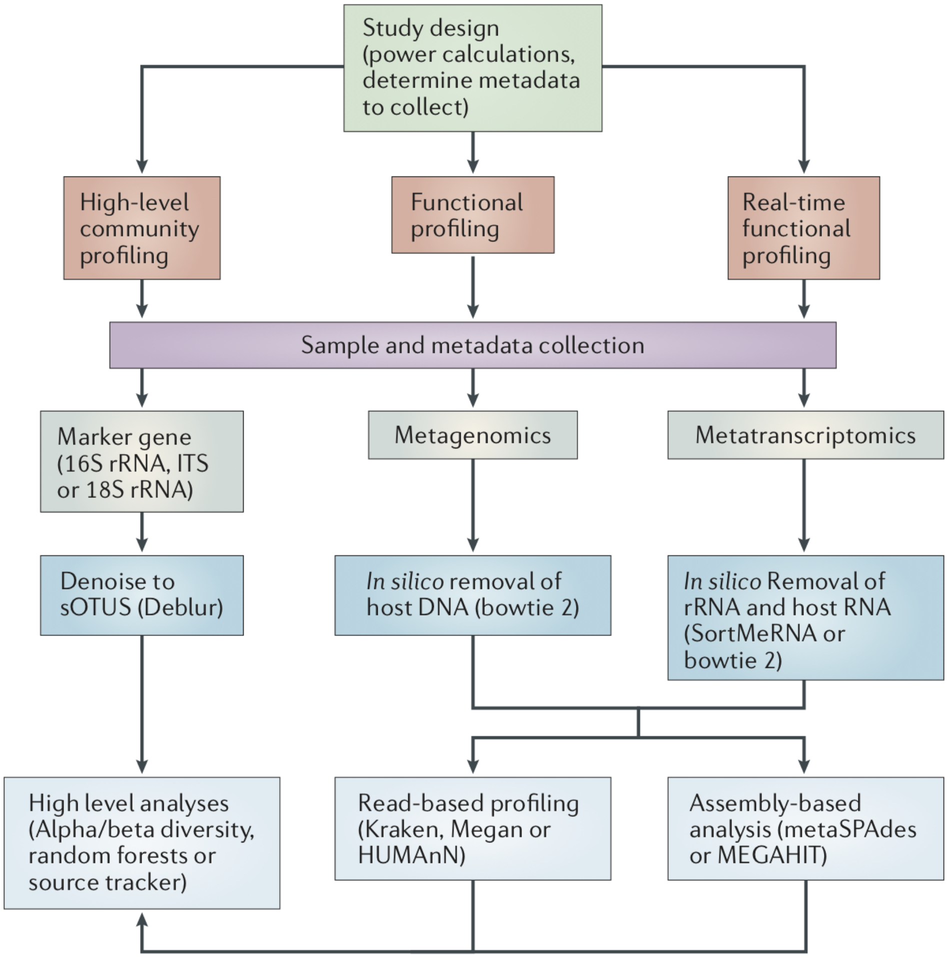

Study design (power calculations, determine metadata to collect)
High-level community profiling
Functional profiling
Real-time functional profiling
Sample and metadata collection
Marker gene (16S rRNA, ITS or 18S rRNA)
Metagenomics
Metatranscriptomics
Denoise to sOTUS (Deblur)
In silico removal of host DNA (bowtie 2)
In silico Removal of rRNA and host RNA (SortMeRNA or bowtie 2)
High level analyses (Alpha/beta diversity, random forests or source tracker)
Read-based profiling (Kraken, Megan or HUMAnN)
Assembly-based analysis (metaSPAdes or MEGAHIT)

Figure 3.2: Best workflow for 16S ribosomal RNA, metagenomic, and metatranscriptomic sequencing. After careful design and sample collection, microbiome data is generated from 16S ribosomal RNA (rRNA), metagenomic or metatranscriptomic sequencing. After performing 16S rRNA sequencing, we recommend using Deblur[58] to resolve sequence data into single-sequence variants called sub-operational taxonomic units (sOTUs). Although DADA2 and Deblur achieve the similar results, Deblur is an order of magnitude faster than DADA2, is parallelizable, and shows greater stability (that is, it obtains the same sOTUs across different samples) [58]. Metagenomics and metatranscriptomics first require pre-processing to remove either host DNA or rRNA and host RNA. The resultant sequencing data can be analyzed by either read-based profiling using state-of-the-art tools such as Kraken [59], Megan [60], or HUMAnN [61], or by assembly-based analyses, with tools such as metaSPAdes [62] and MEGAHIT [63]. For each of these three methods, higher level analyses (for example, alpha and beta diversity, taxonomic profiling and machine learning) are subsequently used to find overall patterns in microbiome variation. Random Forests regression has been effective in many applications, ranging from dating time since death of a corpse [64] to providing an index for microbiome maturation [65]. SourceTracker [66], a Bayesian estimator of the sources that make up each unknown community, is useful for classifying microbial samples according to environment of origin [67].

the functional capacity of an entire community at the gene level [61], moving well beyond the limits of marker gene analysis. However, biases that are introduced by library construction, assembly and reference databases for annotation are less understood than biases that exist in well-characterized marker gene approaches. As the metagenomics field matures, these annotation steps will continue to be improved and validated. For a comprehensive review on metagenomics, we direct the reader to Ref. [68].

3.6.3 Metatranscriptome analysis.

Metatranscriptomics uses RNA sequencing to profile transcription in microbiomes, providing information on gene expression and the active functional output of the microbiome. Metatranscriptomics differs from both marker gene and metagenomic sequencing that sequence DNA in a sample, regardless of cell viability or activity. Although there are methods for depleting relic DNA from dead cells [14], sequencing microbial RNA provides better insight into the functional activity of a microbial community, though it is biased towards organisms with higher rates of transcription. It is worth noting that propidium monoazide (PMA) depletion of relic DNA is an alternative method to identify live microorganisms [69]. Host RNA contamination, particularly the highly abundant rRNAs, is also an important consideration and methods to exclude rRNAs from samples should be considered [54]. RNA must be carefully preserved to avoid degradation in all cases, though certain sample types may warrant specialized protocols for RNA purification. For example, soil samples require removal of enzyme-inhibiting humic substances [70, 71]. Despite these technical difficulties, metatranscriptomic data can offer unique insight; transcriptomes vary more within individuals than metagenomes [72], and metatranscriptomics can reveal microbial community response to perturbations, such as xenobiotic exposure [57]. For a comprehensive review on metatranscriptomics analysis of the microbiome, we direct the reader to Ref. [73].

3.7 Analyses.

Ideally, each microbiome study would analyze samples with all three of the methods discussed above. In most cases, however, there is not enough sample material or enough project funding for performing all three analyses, and in some cases, the samples might not be amenable to one of the sequencing methods. It is therefore paramount that the researcher chooses the method of sequencing that is most effective for answering their specific questions. If there are no budget constraints, we recommend performing metagenomics rather than marker gene sequencing. However, it is common practice to perform marker gene sequencing to gain a low resolution understanding of the microbial community composition. Next, depending on the focus of the study, the researcher can move on to metagenomic and metatranscriptomic sequencing, though this may require a second study for appropriate sample collection and processing.

3.7.1 Marker gene analyses.

As noted above, marker gene approaches are sensitive to technical factors such as primer choice [45], so well-validated protocols such as those used with the diverse sample set in the Earth Microbiome Project should be used [37]. The first step in analyzing marker gene amplicon data is to remove sequencing errors: despite very low sequencing error rates (for example, in Illumina sequencing 0.1% per nucleotide [74]), most of the apparent sequence diversity arises from sequencing errors [75, 76]. Until recently, this problem was addressed by clustering similar sequences into OTUs [13, 77]. Clustering sequences into OTUs, termed OTU picking, consolidates similar sequences (usually with a 97% similarity threshold) into single features, merging sequence variants including those introduced by sequence error into a single OTU that can be used in subsequent analysis. However, this method misses subtle and real biological sequence variation, such as single nucleotide polymorphisms (SNPs) that would be consolidated into single OTUs [78]. Oligotyping[79] improves upon traditional OTU picking by including position-specific information from 16S rRNA sequencing to identify subtle nucleotide variation and by discriminating between closely related but distinct taxa. Algorithms such as Deblur [58] and DADA2 [80] use error profiles to resolve sequence data into exact-sequence features (the marker gene sequence) called sub-OTUs (sOTUs). The resulting output from these methods is a table of DNA sequences and counts of these different sequences per sample rather than OTU groups. We recommend that these methods replace OTU-based approaches for all applications, except when it is necessary to combine sequence data that were generated using different technologies (that is, Illumina sequencing and 454 pyrosequencing) or with different primer sets, when mapping to a common reference database of full-length sequences is often still needed [81].

One key analysis step is to assign taxonomic names to microbial sequences in the data. Taxonomy is typically assigned by machine learning approaches such as the RDP classifier [82], which uses Naive Bayes models that are trained on oligonucleotide frequencies at the genus level to achieve 80% accuracy in genus-level assignments. Popular microbiome analysis packages such as QIIME [13] and Mothur [77] provide support for taxonomic classification. In principle, exact matching to reference databases (three of the most characterized and frequently used are Greengenes, RDP, and Silva) should provide better specificity in taxonomic assignment, but the sensitivity of this approach is poor given the large number of unknown taxa. Furthermore, de novo phylogenetic trees that are constructed from short marker gene sequences are typically poorly resolved, so insertion of marker gene sequences into a characterized reference tree that is based on full-length sequences [83] is desirable, given the importance of phylogenetic metrics [84]. Unclassified microorganisms should be checked for organelle sequences, and for many studies, chloroplast and mitochondria sequences should be excluded before proceeding with analysis (although for intestinal samples, these sequences can be useful for identifying consumed foods and thus should not be disregarded completely).

Predictive functional profiling [39–42] is a technique for linking marker gene studies with available microbial genomes to make predictions about metagenomic content and thus the putative biological functions of a microbial community. This analysis generally requires a reference-based OTU table. Methods based on evolutionary models (for example, PICRUSt[40]) provide confidence intervals on these predictions of gene content, which will tend to be wider in regions of the tree distant from reference genome sequences, and narrower where many reference genomes are available. Thus, the availability of sufficient closely related reference genomes is a main factor that influences the accuracy of these results. Another limitation for predictive functional profiling is that some families of bacteria possess a very similar 16S rRNA variable region, despite being phenotypically and genotypically divergent.

Most statistical analyses that are applied to microbiome data that is generated from marker gene sequencing can also be applied to other types of -omics analyses, and are described below in the Higher-level analyses section.

3.7.2 Metagenome and metatranscriptome analyses.

Surveying the complete nucleic acid profile of a sample yields rich information that can be used to investigate a broad range of taxonomic, functional, and evolutionary aspects of microbial communities even contaminants can provide important details [85]. As with marker gene-based surveys, the analytical methods must be carefully chosen to consider the sample origin and the specific hypotheses under investigation. Here, we discuss the best approaches to perform these analyses. Read-based profiling takes the unassembled DNA or mRNA sequence reads and compares them against reference databases to assign taxonomy or annotate genes. With the ever-increasing size of modern query datasets and databases, methods are continually being refined to improve the speed of read-based profiling. Many tools utilize k-mers, assigning taxonomy to short DNA fragments of length k, such as Kraken [59] or employ the Burrows-Wheeler transform which compresses the database by merging similar sequences (for example, Bowtie2 [86] and Centrifuge [87]). For a more comprehensive guide to tool selection, we direct the reader to Ref. [88]. Marker gene methods (such as MetaPhlAn2 [89] and TIPP [90]) use specific genomic regions for taxonomy assignment, focusing on universal, single-copy elements. Beyond taxonomy assignment, others tools such as HUMAnN2 [61] can also be used for annotating genes and metabolic pathways. Some tools, including MEGAN [60], incorporate both of these functionalities, and can be a preferred method when both annotations are desired. Because each read is considered independently, read-based methods scale efficiently to large, complex data sets, such as soil microbiome data sets. It is important to note that as taxonomic or functional assignment depends on homology between the single read and a reference, database choice is crucial. For well-characterized environments like the human gut, curated genome databases such as RefSeq [91] and protein family databases like Pfam [92] or UniRef [93] increase the accuracy of results and decrease computational costs. For samples from poorly characterized environments, the use of large databases such as NCBI nr and nt and IMG/MG [94] should be considered because the databases are larger, despite the increased computational complexity and decreased assignment specificity. Specialized databases must be used to annotate specific taxonomic or functional categories, such as PHASTER [95] for bacteriophages, Resfams [96] for antibiotic resistance genes and FOAM for environmental samples [97]. Additionally, numerous metagenomic data catalogues are available for many sample types, including Tara for ocean samples [15], the BGI catalogue for mouse gut samples[98] and MetaHit for human gut samples [99]. Another method for analyzing metagenome and metatranscriptome sequencing reads is to assemble the short reads into longer sequences (contigs). These contigs can be further sorted or binned by similarity to assemble partial to full genomes of microorganisms. This allows data exploration beyond taxa and gene annotation, enabling the prediction of multi-gene biosynthetic pathways or even metabolic reconstructions with tools such as antiSMASH [100]. However, assembly-based analyses are not universally applicable; higher biodiversity, the presence of many related strains in samples or low coverage yields fragmented assemblies and can obscure taxa from downstream analyses. For example, soil samples are often difficult to assemble due to the high microbial diversity and uneven distribution [101]. For samples that avoid these complications, metagenome assemblies provide valuable bespoke reference databases for read-based and assembly-based metatranscriptome analyses [102, 103], thus recovering the microbial dark matter that is absent in curated databases [104]. Recommended tools for assembly-based analyses include metaSPAdes [62] and MEGAHIT [63]. A comprehensive discussion of these and other tools can be found in Ref. [105]. To assemble partial to full genomes of individual microorganisms, contigs are sorted (binned) into separate putative genomes with tools such as MaxBin2 [106] and CONCOCT [107], which evaluate nucleotide composition and abundance patterns across samples to perform sorting (binning). To evaluate the quality of these binned and assembled genomes, single-copy gene profiling tools such as CheckM [108] that use common single-copy genes to estimate genome completeness and contamination can be used. Additionally, visualization tools like VizBin [109] display clustering of metagenomic sequences without alignment to a reference database, allowing researchers to visually inspect the sequence clustering of related organisms and assist with evaluating bin quality. Employing integrated workflow tools to automate data processing such as Anvio [110], ATLAS [111], or MetAMOS [112], is highly recommended because assembly-based methods are complex. In order to compare samples with varying sequencing read counts, various methods of normalization can be employed. Common methods of normalization include: read counts per million (counts are scaled by the total number of reads), transcripts per kilobase million (counts scaled by number of reads and length of reads), and converting the data to relative abundance. Additionally, there are various tools for performing normalization including edgeR [113] and DESeq2 [114]. New tools for both read-based and assembly-based approaches are under rapid development. When possible, specific analytical decisions should be made based on performance on well-studied or synthetic datasets (such as the Critical Assessment of Metagenomic Information [115]) that are most similar to the microbial community of interest.

3.8 Higher-level analyses.

Processing microbiome data generates a matrix that relates feature abundance (taxa or genes) to samples. This output is deceptively simple; microbiome data is highly dimensional, often representing thousands of different taxa, and sparse with many zeros present in the matrix, requiring careful statistical treatment to extract meaningful results. Overall patterns in microbiome variation are typically assessed by alpha and beta diversity. Alpha diversity quantifies feature diversity within individual samples and can be compared across sample groups. For example, when comparing a sample from an individual with a disease to a healthy control, the researcher can use alpha diversity to compare the mean species diversity between the two samples. Measures of species richness (for example, the number of observed species, or Chao1 abundance estimator, which estimates true species diversity) and phylogenetic measures (Faiths phylogenetic diversity) are sensitive to the number of sequences per sample, whereas measures that combine richness and evenness (Shannon index) are much less so. However, it should be noted that these methods have been evaluated exclusively for 16S rRNA data, and may not apply to other microbiome data types. Beta diversity compares feature dissimilarity between each pair of samples, generating a distance matrix of beta diversity distances between all pairs of samples. Metric selection can influence the results obtained [84, 116] and should be chosen with biological data interpretation in mind. Quantitative metrics (Bray-Curtis, Canberra and weighted UniFrac) use feature abundance data in calculations whereas qualitative metrics (binary-Jaccard and unweighted UniFrac) only consider the presence or absence of features. Phylogenetic measures such as UniFrac typically provide interpretable biological patterns [117], though these metrics require a phylogenetic tree and thus cannot be used for direct comparison with omics data that lack trees. Software for performing alpha and beta diversity calculations includes QIIME [13], Mothur [77], and the R package Vegan [118]. The non-parametric permutation tests PERMANOVA and ANOSIM are used for assessing significant beta diversity clustering between groups, but PERMANOVA may perform better on datasets with varying dispersions within groups [119]. Calculation of meaningful alpha and beta diversity measures requires the researcher to control for the sampling effort (that is, the number of sequences per sample obtained), as this can differ by orders of magnitude. The current best solution for UniFrac is rarefaction [120], though for the special case of pairwise differential abundance testing, the full sample set should be used [121]. For visualizing beta diversity data, ordination techniques, such as principal coordinates analysis (PCoA) or principal component analysis (PCA), are commonly used. These methods reduce large and complex distance matrices into a visually manageable two dimensional or three dimensional representations of sample distances. Samples can then be colored by various metadata categories to visualize clustering in an unsupervised manner. EMPeror offers an interactive framework for manipulating PCoA plots [122]. Another common analysis approach is to look at differentially abundant microorganisms or functional elements (for example, genes and pathways) in the comparison groups of interest (that is, treatment versus control). Identifying microbial taxa that explain differences between communities is particularly challenging because microbiome data sets are high-dimensional (that is, they include thousands of taxa), sparse and compositional. Compositionality is the crux of the problem [123]; when the proportion of one microorganism increases, the proportions of others must decrease for the proportions to sum to 1. For example, suppose a patient is administered a drug that increases the growth rate in only a single microbial genus, while not affecting the growth of others. Although the other microorganisms are not impacted by the drug, they would have decreased in relative abundance due to the outgrowth of the single microbial genus. This poses challenges for many classical methods, such as parametric statistical tests (for example, Students t-test and ANOVA), and measures of correlation including Spearmans rank correlation, often leading to completely unacceptable false discovery rates above 90% [120, 124, 125]. Recently, compositionally-aware methods have addressed this problem of compositionality and relative abundance. One approach is to force strong biological assumptions on the statistical test: for example, Lovells proportionality metric detects only positive correlations [126]. Other tools that are widely applicable and have been optimized for microbiome data, such as SparCC [127] and SPEIC-EASI [128], assume that few species are correlated, so most correlation coefficients are zero. BAnOCC [129] is another tool for addressing the compositionality problem that makes no assumptions about the data. We recommend another approach that does not assume few species are correlated, which is to test for differences between microbial communities using the isometric log ratio transform (ilr) [sic]. The isometric log ratio transform approach controls for false positives due to proportionality by testing for the changes in log ratios between microbial abundances, commonly referred to as balances. Balances can be constructed using prior knowledge such as evolutionary history [117, 130, 131] or microbial niche differentiation in response to environmental factors such as pH [132]. After the ilr transform is applied, standard statistical tools such as multivariate response, linear regression and classification can effectively test for differences on the balances or log ratios between microorganisms rather than the raw microbial abundances, controlling for compositionally. Other recent methods use absolute quantification to address compositionality by complementing sequencing with microbial cell counts in each sample [133, 134].

Machine learning is emerging as an especially useful technique for determining how microbiome data can be used to separate samples based on current state (usually determined by metadata categories, such as healthy state versus diseased state) [135, 136] or, excitingly, to predict future state [137, 138]. For instance, it is possible to model the severity and susceptibility of gingivitis based on an individuals oral microbiota [137]. Random Forests regression, a machine learning technique, has been effective in many applications, ranging from dating time since death of a corpse [64] to providing a model for determining microbiome maturation in child development [65]. SourceTracker [66], a Bayesian estimator of the microbial sources that make up an unknown community, is useful for classifying microbial samples according to environment of origin [67]. Importantly, machine learning analyses need a substantial sample size and should always be coupled with cross-validation, independent test sets, or other experimental and biological confirmation to ensure robust findings.

3.9 Integrating other omics data.

Knowing the composition of a microbial community is no longer a sufficient research goal; we want to know the function of the community. Integrating other data types including marker gene sequencing, metagenomics, metatranscriptomics, metaproteomics, metabolomics and other techniques for a given study is crucial for a comprehensive understanding of the composition and function of microbial communities. For example, changes in the metabolite profile of a microbial community reflect changes in its biosynthetic activity, mRNA and protein expression, and protein activity [139]. Multi-omics analysis integrates chemical and biological knowledge to provide a more complete picture of a biological system and is an active area of research with largely untested methods (Figure 3.3).

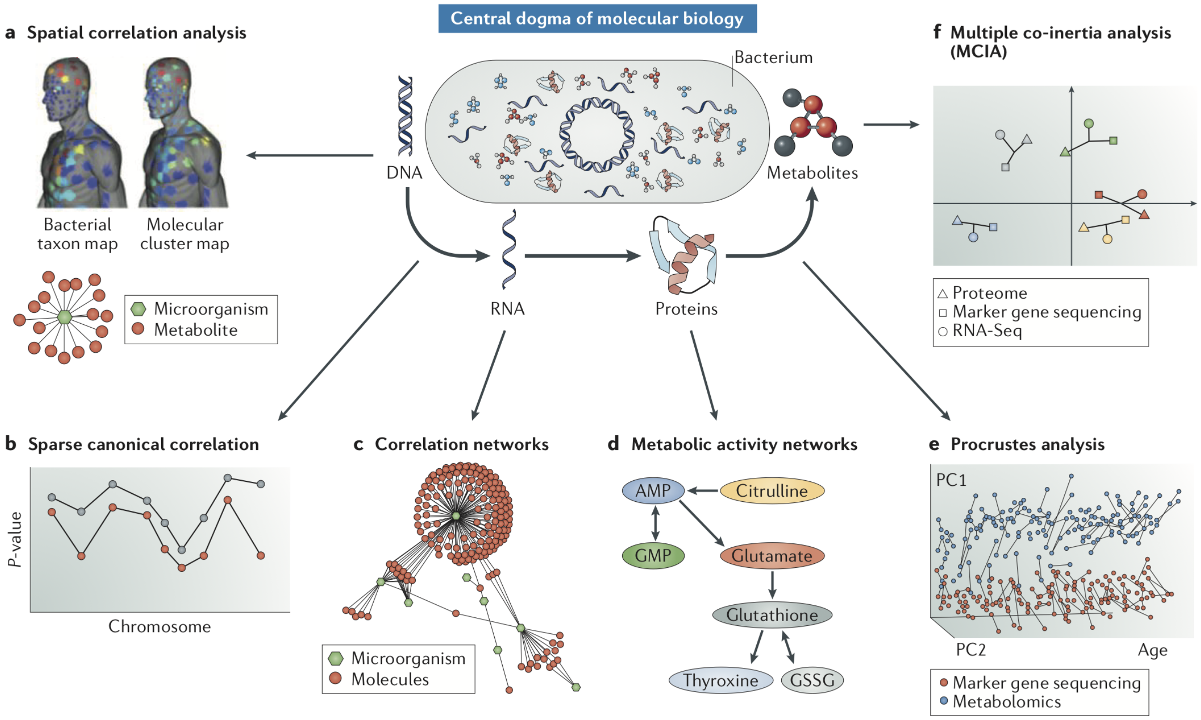


Figure 3.3: Integrating -omics data with microbiome data. The central dogma of molecular biology of progression from genes to downstream metabolic products is reflected by the compendia of corresponding -omes co-occurring within the cell. Linking the knowledge from different -omics studies constitutes the multi-omics analysis. Panels around the cell represent some integration examples of various -omics data with marker gene sequencing: a) Three dimensional visualization of mapped molecular and microbial (or any other) features aids our understanding of spatial correlation thereof. b) Sparse canonical correlation analysis [140] identifying linear combinations of the two sets of variables that are highly correlated with each other. c) Correlation network analysis shows clustering of a particular microorganism with metabolites that are potentially produced and/or processed by it. d) Metabolic activity networks help to predict microbial community structure and function by mathematical modelling of the molecular mechanisms of particular organism(s). e) Procrustes analysis enables the direct comparison of different -omics data sets with the same internal structure on a single PCoA plot to reveal trends in the data. f) Multiple co-inertia analysis (MCIA) enables multidimensional comparisons through graphical representation, so that the similarity of different -omics data can be more easily understood.

Integrating multi-omics data types is inherently difficult. For example, gene expression and metabolism operate on different timescales [141], and microorganisms produce many metabolites, often only in response to molecular signals from other species [142]. Also, the sparse nature of metagenomic and metabolomic data (where the data matrices are composed mostly of zeros) is much greater than of metaproteomic data and this may pose technical problems for some methods. Although the integration of different -omics data sets is a work in progress, tools that integrate these datasets are becoming increasingly available. For example, XCMS Online integrates metabolomic data with metabolic pathways, as well as transcriptomic and proteomic data [143]. Traditional correlation methods such as Pearson and Spearman could enable pairwise correlation between features across -omics data sets. However, these are prone to false positives due to the sparsity and high-dimensionality of microbiome and metabolome datasets. Procrustes analysis [144] uses dimensionally-reduced data to test if patterns (distances) between samples in one dataset is observed in the other, essentially correlating ordination spaces rather than individual features (tested using Mantel [145] or PROcrustes randomization TEST). Other methods integrate -omics datasets by not only taking into account the relationships between samples, but also associating samples to particular metadata categories of interest (such as examining healthy versus diseased or control versus treatment groups). These methods include co-inertia analysis, which uses dimensionality reduction to associate sample patterns in two data sets and relevant metadata [146], and partial least-squares [147], related methods such as canonical correlation analysis [140], or robust sparse canonical correlation analysis, which is a variation of the method to deal with sparse -omics data [148]. Advanced integrative analysis tools include molecular networking with GNPS [149] to identify metabolites and pathway annotations [150], and general systems biology tools, exemplified by XCMS Online [143]. Increasingly, multi-omics studies are investigating temporal patterns in addition to spatial patterns. Spatial mapping [151], that can now be performed with the tool ili[151], adds a powerful dimension to multi-omics studies through visual representations that are readily amenable to human interpretation.

Integration with other -omics data can be performed using various statistical methodologies [152]. However, these techniques have been shown to perform suboptimally on microbiome data sets [125]. Furthermore, simply finding correlations in various -omics data by itself is only the first step. Establishing causation and correlation across data sets is the next challenge. Box 3 gives an example of the integration of metabolome and microbiome data sets and corresponding approaches to move beyond correlation and determine causation. Correction for multiple comparisons is crucial in multi-omic analyses; data sets can contain thousands of different microorganisms and metabolites, so significant correlations are expected by random chance. Measures to correct significance testing for multiple comparisons include the

False Discovery Rate (for example, Benjamini-Hochberg correction) or, for more conservative corrections, the Family-Wise Error (for example, Bonferroni correction). Using these methods to penalize multiple comparisons in conjunction with statistical models that incorporate sparsity and compositionality [125], false discovery rates in large multi-omic comparisons can be reduced.

Despite these challenges, the future potential for -omics data integration is promising. In particular, there are numerous examples where metagenome, metatranscriptome and metabolome data have been successfully integrated, illuminating gene regulation in microbiomes [38] and correlating the presence of microorganisms with metabolites [153]. Such studies have provided insights beyond the capacity of single omics, such as gut bacterial metabolism of xenobiotics [57] and how antibiotic-induced microbiome depletion creates a favorable metabolomic environment for Clostridium difficile [154]. Comparatively, the integration of metaproteomics data with microbiome data is a relatively newer field of investigation, though there are many recent examples of successful integration ranging from identifying biomarkers of Crohns disease [155] to examining microbial protein production in layers of permafrost [156]. Additionally, tool development for metaproteomics annotations and analysis is ongoing [157, 158]. Overall, integrating -omics data can provide a more holistic and mechanistic understanding of microbiomes from DNA identification to functional production of metabolites and proteins and ideally lead to more actionable scientific insights.

3.10 Box 3. Metabolomics and the microbiome.

Microbially produced metabolites influence host physiology, can shape microbial community dynamics and are involved in both health and disease. These metabolites can have both beneficial (for example, short-chain fatty acids (SCFAs) [159]) and detrimental effects on the host (for example, the genotoxin colibactin [160]). However, identifying a metabolite as sourced from the microbiome is particularly challenging. Even more challenging is identifying which microorganism or collection of microorganisms produced or modified a particular metabolite. Here are several strategies to address this problem:

1. Compare metabolites from natural samples to those from cultured isolates of microbiome-isolated

microorganisms. One useful approach is matching tandem mass spectrometry data from cultured isolates to clinical or environmental samples, showing that a particular metabolite signature can be sourced from the cultured microorganism [161].

2. Map metabolites detected in a microbiome sample to paired genome or metagenomic data. Some metabolites are unique to particular microbial taxa. Detection of these metabolites in a natural sample can enable determination of their likely source by mining paired genomic data for genes known to produce that metabolite. For example, 2,3-butanedione, a unique fermentation product, is a microbial metabolite produced by Streptococcus spp. Detection of this metabolite in clinical samples along with the biosynthetic genes, facilitates mapping of reads to the biochemical pathway back to the genome of the organism of origin [153].

3. Build co-occurrence networks of microorganisms and metabolites. Co-occurrence or correlation methods associate microorgaisms with metabolite features. This is an active area of research, but available algorithms that have been optimized for detecting correlations between microorganisms in sparse microbiome data include SparCC [127], CCLasso [162], and others [125, 163]. However, this approach warrants caution because of the high false discovery rates across the large multivariate datasets.

4. Germ free versus specific pathogen free murine models. These comparisons identify metabolites from the microbiome as metabolites detected in colonized mice but not in uncolonized mice are likely produced by microorganisms. Gnotobiotic mice (mono-colonized or with defined communities) help identify specific microorganisms that produce metabolites of interest [164].

3.11 Conclusions

In this Review, we have discussed how all stages of conducting a microbiome study, from designing the experiment to collecting and storing the samples, to obtaining insight from graphical displays of the sequence data, can substantially impact the results and their biological interpretation. As the effects of many of these technical steps are large compared to the real biological variability to be explained, standardization is necessary in order to compare and combine separate studies, and the first efforts to do this and to provide recommendations and best practices, such as the International Human Microbiome Standards and the Microbiome Quality Control Project (MBQC), are already under way. Including bioinformatics pipelines and controls into these standardization efforts, and in particular using cloud-enabled reproducible computing resources that run open-source code on publicly available data to reproduce scientific claims of publications, is a rapidly emerging area that will bring consistency and comparability to the microbiome field. An important part of such efforts will be spike-in standards (which have already been so important to standardizing microarrays), and standardized biologically realistic samples that can be used to quantify systems-level accuracy in microbiome assays.

This article has focused primarily on DNA-level analyses at the whole-community level, but as expression-level profiling and single-cell profiling techniques continue to advance, many similar considerations will apply to those types of data also. Avoiding the mistakes that have been repeated frequently in other expensive assays, such as inadequate sample size and validation, and employing best practices for standards, sample handling, compositional data analysis and other frequent pitfalls, will accelerate progress in these areas. Using standardized and well-characterized sample sets, such as those developed in MBQC and in the Earth Microbiome Project, can greatly shorten the time needed to understand the value and unique insights provided by a new technique.

As the field trends towards ever-larger data sets, understanding subtle confounding factors long known to epidemiologists and taking more care with longitudinal study designs will become increasingly important. The value of interventional studies over observational studies is considerable, especially when human, animal model and in vitro data can be correlated across scales and systems. Increased standardization of techniques and dissemination of methods with low noise and bias will greatly increase the ability of the microbiome field to deliver on the promise of translatability from lab-scale studies to the clinic, field or natural environment.

3.12 Acknowledgments

Chapter 3, in full, is a modified reprint of the material as it appears in “R. Knight[†], A. Vrbanac[†], B. C. Taylor[†], A. Aksenov, C. Callewaert, J. Debelius, A. Gonzalez, T. Kosciolek, L. McCall, D. McDonald, A. V. Melnik, J. T. Morton, J. Navas, R. A. Quinn, J. G. Sanders, A. D. Swafford, L. R. Thompson, A. Tripathi, Z. Z. Xu, J. R. Zaneveld, Q. Zhu, J. G. Caporaso, and P.C. Dorrestein. *Best Practices for Analysing Microbiomes*. Nature Reviews Microbiology (2018). DOI: 10.1038/s41579-018-0029-9.”. The dissertation author was a primary coinvestigator and author of this work. With the three co-first authors, the dissertation author conceptualized the scope of the manuscript, researched the data for the manuscript, designed and created the figures, and wrote the manuscript.

This review is informed by our work funded by the National Institutes of Health, National Science Foundation, Alfred P. Sloan Foundation, John Templeton Foundation and W. M. Keck Foundation, as well as that of hundreds of collaborators on the Human Microbiome Project, American Gut Project and Earth Microbiome Project.

References

(1) Meisel, J. S.; Hannigan, G. D.; Tyldsley, A. S.; SanMiguel, A. J.; Hodkinson, B. P.; Zheng, Q.; Grice, E. A. *Journal of Investigative Dermatology* **2016**, DOI: 10.1016/j.jid.2016.01.016.

(2) Falony, G.; Joossens, M.; Vieira-Silva, S.; Wang, J.; Darzi, Y.; Faust, K.; Kurilshikov, A.; Bonder, M. J.; Valles-Colomer, M.; Vandeputte, D.; Tito, R. Y.; Chaffron, S.; Rymenans, L.; Verspecht, C.; Sutter, L. D.; Lima-Mendez, G.; D'hoe, K.; Jonckheere, K.; Homola, D.; Garcia, R.; Tigchelaar, E. F.; Eeckhaudt, L.; Fu, J.; Henckaerts, L.; Zhernakova, A.; Wijmenga, C.; Raes, J. *Science* **2016**, DOI: 10.1126/science.aad3503.

(3) Noguera-Julian, M.; Rocafort, M.; Guillén, Y.; Rivera, J.; Casadellà, M.; Nowak, P.; Hildebrand, F.; Zeller, G.; Parera, M.; Bellido, R.; Rodríguez, C.; Carrillo, J.; Mothe, B.; Coll, J.; Bravo, I.; Estany, C.; Herrero, C.; Saz, J.; Sirera, G.; Torrela, A.; Navarro, J.; Crespo, M.; Brander, C.; Negredo, E.; Blanco, J.; Guarner, F.; Calle, M. L.; Bork, P.; Sönnerborg, A.; Clotet, B.; Paredes, R. *EBioMedicine* **2016**, DOI: 10.1016/j.ebiom.2016.01.032.

(4) Wu, G. D.; Chen, J.; Hoffmann, C.; Bittinger, K.; Chen, Y. Y.; Keilbaugh, S. A.; Bewtra, M.; Knights, D.; Walters, W. A.; Knight, R.; Sinha, R.; Gilroy, E.; Gupta, K.; Baldassano, R.; Nessel, L.; Li, H.; Bushman, F. D.; Lewis, J. D. *Science* **2011**, DOI: 10.1126/science.1208344.

(5) Forslund, K.; Hildebrand, F.; Nielsen, T.; Falony, G.; Le Chatelier, E.; Sunagawa, S.; Prifti, E.; Vieira-Silva, S.; Gudmundsdottir, V.; Krogh Pedersen, H.; Arumugam, M.; Kristiansen, K.; Yvonne Voigt, A.; Vestergaard, H.; Hercog, R.; Igor Costea, P.; Roat Kultima, J.; Li, J.; Jørgensen, T.; Levenez, F.; Dore, J.; Bjørn Nielsen, H.; Brunak, S.; Raes, J.; Hansen, T.; Wang, J.; Dusko Ehrlich, S.; Bork, P.; Pedersen, O. *Nature* **2015**, DOI: 10.1038/nature15766.

(6) Jackson, M. A.; Goodrich, J. K.; Maxan, M. E.; Freedberg, D. E.; Abrams, J. A.; Poole, A. C.; Sutter, J. L.; Welter, D.; Ley, R. E.; Bell, J. T.; Spector, T. D.; Steves, C. J. *Gut* **2016**, DOI: 10.1136/gutjnl-2015-310861.

(7) Halfvarson, J.; Brislawn, C. J.; Lamendella, R.; Vázquez-Baeza, Y.; Walters, W. A.; Bramer, L. M.; D'Amato, M.; Bonfiglio, F.; McDonald, D.; Gonzalez, A.; McClure, E. E.; Dunklebarger, M. F.; Knight, R.; Jansson, J. K. *Nature Microbiology* **2017**, DOI: 10.1038/nmicrobiol.2017.4.

(8) Kelly, C. J.; Zheng, L.; Campbell, E. L.; Saeedi, B.; Scholz, C. C.; Bayless, A. J.; Wilson, K. E.; Glover, L. E.; Kominsky, D. J.; Magnuson, A.; Weir, T. L.; Ehrentraut, S. F.; Pickel, C.; Kuhn, K. A.; Lanis, J. M.; Nguyen, V.; Taylor, C. T.; Colgan, S. P. *Cell Host and Microbe* **2015**, DOI: 10.1016/j.chom.2015.03.005.

(9) Debelius, J.; Song, S. J.; Vazquez-Baeza, Y.; Xu, Z. Z.; Gonzalez, A.; Knight, R. *Genome Biology* **2016**, DOI: 10.1186/s13059-016-1086-x.

(10) La Rosa, P. S.; Brooks, J. P.; Deych, E.; Boone, E. L.; Edwards, D. J.; Wang, Q.; Sodergren, E.; Weinstock, G.; Shannon, W. D. *PLoS ONE* **2012**, DOI: 10.1371/journal.pone.0052078.

(11) Knights, D.; Costello, E. K.; Knight, R. *FEMS Microbiology Reviews* **2011**, DOI: 10.1111/j.1574-6976.2010.00251.x.

(12) Yilmaz, P.; Kottmann, R.; Field, D.; Knight, R.; Cole, J. R.; Amaral-Zettler, L.; Gilbert, J. A.; Karsch-Mizrachi, I.; Johnston, A.; Cochrane, G.; Vaughan, R.; Hunter, C.; Park, J.; Morrison, N.; Rocca-Serra, P.; Sterk, P.; Arumugam, M.; Bailey, M.; Baumgartner, L.; Birren, B. W.; Blaser, M. J.; Bonazzi, V.; Booth, T.; Bork, P.; Bushman, F. D.; Buttigieg, P. L.; Chain, P. S.; Charlson, E.; Costello, E. K.; Huot-Creasy, H.; Dawyndt, P.; Desantis, T.; Fierer, N.; Fuhrman, J. A.; Gallery, R. E.; Gevers, D.; Gibbs, R. A.; Gil, I. S.; Gonzalez, A.; Gordon, J. I.; Guralnick, R.; Hankeln, W.; Highlander, S.; Hugenholtz, P.; Jansson, J.; Kau, A. L.; Kelley, S. T.; Kennedy, J.; Knights, D.; Koren, O.; Kuczynski, J.; Kyrpides, N.; Larsen, R.; Lauber, C. L.; Legg, T.; Ley, R. E.; Lozupone, C. A.; Ludwig, W.; Lyons, D.; Maguire, E.; Methé, B. A.; Meyer, F.; Muegge, B.; Nakielny, S.; Nelson, K. E.; Nemergut, D.; Neufeld, J. D.; Newbold, L. K.; Oliver, A. E.; Pace, N. R.; Palanisamy, G.; Peplies, J.; Petrosino, J.; Proctor, L.; Pruesse, E.; Quast, C.; Raes, J.; Ratnasingham, S.; Ravel, J.; Relman, D. A.; Assunta-Sansone, S.; Schloss, P. D.; Schriml, L.; Sinha, R.; Smith, M. I.; Sodergren, E.; Spor, A.; Stombaugh, J.; Tiedje, J. M.; Ward, D. V.; Weinstock, G. M.; Wendel, D.; White, O.; Whiteley, A.; Wilke, A.; Wortman, J. R.; Yatsunenko, T.; Glöckner, F. O. *Nature Biotechnology* **2011**, DOI: 10.1038/nbt.1823.

(13) Caporaso, J. G.; Kuczynski, J.; Stombaugh, J.; Bittinger, K.; Bushman, F. D.; Costello, E. K.; Fierer, N.; Pa, A. G.; Goodrich, J. K.; Gordon, J. I.; Huttley, G. A.; Kelley, S. T.; Knights, D.; Koenig, J. E.; Ley, R. E.; Lozupone, C. A.; McDonald, D.; Muegge, B. D.; Pirrung, M.; Reeder, J.; Sevinsky, J. R.; Turnbaugh, P. J.; Walters, W. A.; Widmann, J.; Yatsunenko, T.; Zaneveld, J.; Knight, R. *Nature Methods* **2010**, DOI: 10.1038/nmeth.f.303.

(14) Carini, P.; Marsden, P. J.; Leff, J. W.; Morgan, E. E.; Strickland, M. S.; Fierer, N. *Nature Microbiology* **2016**, DOI: 10.1038/nmicrobiol.2016.242.

(15) Sunagawa, S.; Coelho, L. P.; Chaffron, S.; Kultima, J. R.; Labadie, K.; Salazar, G.; Djahanschiri, B.; Zeller, G.; Mende, D. R.; Alberti, A.; Cornejo-Castillo, F. M.; Costea, P. I.; Cruaud, C.; D'Ovidio, F.; Engelen, S.; Ferrera, I.; Gasol, J. M.; Guidi, L.; Hildebrand, F.; Kokoszka, F.; Lepoivre, C.; Lima-Mendez, G.; Poulain, J.; Poulos, B. T.; Royo-Llonch, M.; Sarmento, H.; Vieira-Silva, S.; Dimier, C.; Picheral, M.; Searson, S.; Kandels-Lewis, S.; Boss, E.; Follows, M.; Karp-Boss, L.; Krzic, U.; Reynaud, E. G.; Sardet, C.; Sieracki, M.; Velayoudon, D.; Bowler, C.; De Vargas, C.; Gorsky, G.; Grimsley, N.; Hingamp, P.; Iudicone, D.; Jaillon, O.; Not, F.; Ogata, H.; Pesant, S.; Speich, S.; Stemmann, L.; Sullivan, M. B.; Weissenbach, J.; Wincker, P.; Karsenti, E.; Raes, J.; Acinas, S. G.; Bork, P. *Science* **2015**, DOI: 10.1126/science.1261359.

(16) Salter, S. J.; Cox, M. J.; Turek, E. M.; Calus, S. T.; Cookson, W. O.; Moffatt, M. F.; Turner, P.; Parkhill, J.; Loman, N. J.; Walker, A. W. *BMC Biology* **2014**, DOI: 10.1186/s12915-014-0087-z.

(17) Amir, A.; McDonald, D.; Navas-Molina, J. A.; Debelius, J.; Morton, J. T.; Hyde, E.; Robbins-Pianka, A.; Knight, R. *mSystems* **2017**, DOI: 10.1128/msystems.00199-16.

(18) Dethlefsen, L.; Relman, D. A. *Proceedings of the National Academy of Sciences of the United States of America* **2011**, DOI: 10.1073/pnas.1000087107.

(19) Fierer, N.; Hamady, M.; Lauber, C. L.; Knight, R. *Proceedings of the National Academy of Sciences of the United States of America* **2008**, DOI: 10.1073/pnas.0807920105.

(20) Costello, E. K.; Lauber, C. L.; Hamady, M.; Fierer, N.; Gordon, J. I.; Knight, R. *Science* **2009**, DOI: 10.1126/science.1177486.

(21) Huttenhower, C. et al. *Nature* **2012**, DOI: 10.1038/nature11234.

(22) McDonald, D.; Birmingham, A.; Knight, R. *Microbiome* **2015**, DOI: 10.1186/s40168-015-0117-2.

(23) Ramette, A. *FEMS Microbiology Ecology* **2007**, DOI: 10.1111/j.1574-6941.2007.00375.x.

(24) Kostic, A. D.; Howitt, M. R.; Garrett, W. S. *Genes and Development* **2013**, DOI: 10.1101/gad.212522.112.

(25) Ridaura, V. K.; Faith, J. J.; Rey, F. E.; Cheng, J.; Duncan, A. E.; Kau, A. L.; Griffin, N. W.; Lombard, V.; Henrissat, B.; Bain, J. R.; Muehlbauer, M. J.; Ilkayeva, O.; Semenkovich, C. F.; Funai, K.; Hayashi, D. K.; Lyle, B. J.; Martini, M. C.; Ursell, L. K.; Clemente, J. C.; Van Treuren, W.; Walters, W. A.; Knight, R.; Newgard, C. B.; Heath, A. C.; Gordon, J. I. *Science* **2013**, DOI: 10.1126/science.1241214.

(26) Reber, S. O.; Siebler, P. H.; Donner, N. C.; Morton, J. T.; Smith, D. G.; Kopelman, J. M.; Lowe, K. R.; Wheeler, K. J.; Fox, J. H.; Hassell, J. E.; Greenwood, B. N.; Jansch, C.; Lechner, A.; Schmidt, D.; Uschold-Schmidt, N.; Füchsl, A. M.; Langgartner, D.; Walker, F. R.; Hale, M. W.; Perez, G. L.; Van Treuren, W.; González, A.; Halweg-Edwards, A. L.; Fleshner, M.; Raison, C. L.; Rook, G. A.; Peddada, S. D.; Knight, R.; Lowry, C. A. *Proceedings of the National Academy of Sciences of the United States of America* **2016**, DOI: 10.1073/pnas.1600324113.

(27) Ley, R. E.; Bäckhed, F.; Turnbaugh, P.; Lozupone, C. A.; Knight, R. D.; Gordon, J. I. *Proceedings of the National Academy of Sciences of the United States of America* **2005**, DOI: 10.1073/pnas.0504978102.

(28) Friswell, M. K.; Gika, H.; Stratford, I. J.; Theodoridis, G.; Telfer, B.; Wilson, I. D.; McBain, A. J. *PLoS ONE* **2010**, DOI: 10.1371/journal.pone.0008584.

(29) Snijders, A. M.; Langley, S. A.; Kim, Y. M.; Brislawn, C. J.; Noecker, C.; Zink, E. M.; Fansler, S. J.; Casey, C. P.; Miller, D. R.; Huang, Y.; Karpen, G. H.; Celniker, S. E.; Brown, J. B.; Borenstein, E.; Jansson, J. K.; Metz, T. O.; Mao, J. H. *Nature Microbiology* **2016**, DOI: 10.1038/nmicrobiol.2016.221.

(30) Stagaman, K.; Burns, A. R.; Guillemin, K.; Bohannan, B. J. *ISME Journal* **2017**, DOI: 10.1038/ismej.2017.28.

(31) Sinha, R.; Abu-Ali, G.; Vogtmann, E.; Fodor, A. A.; Ren, B.; Amir, A.; Schwager, E.; Crabtree, J.; Ma, S.; Abnet, C. C.; Knight, R.; White, O.; Huttenhower, C. *Nature Biotechnology* **2017**, DOI: 10.1038/nbt.3981.

(32) Costea, P. I.; Zeller, G.; Sunagawa, S.; Pelletier, E.; Alberti, A.; Levenez, F.; Tramontano, M.; Driessen, M.; Hercog, R.; Jung, F. E.; Kultima, J. R.; Hayward, M. R.; Coelho, L. P.; Allen-Vercoe, E.; Bertrand, L.; Blaut, M.; Brown, J. R.; Carton, T.; Cools-Portier, S.; Daigneault, M.; Derrien, M.; Druesne, A.; De Vos, W. M.; Finlay, B. B.; Flint, H. J.; Guarner, F.; Hattori, M.; Heilig, H.; Luna, R. A.; Van Hylckama Vlieg, J.; Junick, J.; Klymiuk, I.; Langella, P.; Le Chatelier, E.; Mai, V.; Manichanh, C.; Martin, J. C.; Mery, C.; Morita, H.; O'toole, P. W.; Orvain, C.; Patil, K. R.; Penders, J.; Persson, S.; Pons, N.; Popova, M.; Salonen, A.; Saulnier, D.; Scott, K. P.; Singh, B.; Slezak, K.; Veiga, P.; Versalovic, J.; Zhao, L.; Zoetendal, E. G.; Ehrlich, S. D.; Dore, J.; Bork, P. *Nature Biotechnology* **2017**, DOI: 10.1038/nbt.3960.

(33) Fouhy, F.; Deane, J.; Rea, M. C.; O'Sullivan, Ó.; Ross, R. P.; O'Callaghan, G.; Plant, B. J.; Stanton, C. *PLoS ONE* **2015**, DOI: 10.1371/journal.pone.0119355.

(34) Song, S. J.; Amir, A.; Metcalf, J. L.; Amato, K. R.; Xu, Z. Z.; Humphrey, G.; Knight, R. *mSystems* **2016**, DOI: 10.1128/msystems.00021-16.

(35) Ward, D. V.; Gevers, D.; Giannoukos, G.; Earl, A. M.; Methé, B. A.; Sodergren, E.; Feldgarden, M.; Ciulla, D. M.; Tabbaa, D.; Arze, C.; Appelbaum, E.; Aird, L.; Anderson, S.; Ayvaz, T.; Belter, E.; Bihan, M.; Bloom, T.; Crabtree, J.; Courtney, L.; Carmichael, L.; Dooling, D.; Erlich, R. L.; Farmer, C.; Fulton, L.; Fulton, R.; Gao, H.; Gill, J. A.; Haas, B. J.; Hemphill, L.; Hall, O.; Hamilton, S. G.; Hepburn, T. A.; Lennon, N. J.; Joshi, V.; Kells, C.; Kovar, C. L.; Kalra, D.; Li, K.; Lewis, L.; Leonard, S.; Muzny, D. M.; Mardis, E.; Mihindukulasuriya, K.; Magrini, V.; O'Laughlin, M.; Pohl, C.; Qin, X.; Ross, K.; Ross, M. C.; Rogers, Y. H. A.; Singh, N.; Shang, Y.; Wilczek-Boney, K.; Wortman, J. R.; Worley, K. C.; Youmans, B. P.; Yooseph, S.; Zhou, Y.; Schloss, P. D.; Wilson, R.; Gibbs, R. A.; Nelson, K. E.; Weinstock, G.; DeSantis, T. Z.; Petrosino, J. F.; Highlander, S. K.; Birren, B. W. *PLoS ONE* **2012**, DOI: 10.1371/journal.pone.0039315.

(36) Chase, J.; Fouquier, J.; Zare, M.; Sonderegger, D. L.; Knight, R.; Kelley, S. T.; Siegel, J.; Caporaso, J. G. *mSystems* **2016**, DOI: 10.1128/msystems.00022-16.

(37) Thompson, L. R. et al. *Nature* **2017**, DOI: 10.1038/nature24621.

(38) Zaneveld, J. R.; Lozupone, C.; Gordon, J. I.; Knight, R. *Nucleic Acids Research* **2010**, DOI: 10.1093/nar/gkq066.

(39) Okuda, S.; Tsuchiya, Y.; Kiriyama, C.; Itoh, M.; Morisaki, H. *Nature Communications* **2012**, DOI: 10.1038/ncomms2203.

(40) Langille, M. G.; Zaneveld, J.; Caporaso, J. G.; McDonald, D.; Knights, D.; Reyes, J. A.; Clemente, J. C.; Burkepile, D. E.; Vega Thurber, R. L.; Knight, R.; Beiko, R. G.; Huttenhower, C. *Nature Biotechnology* **2013**, DOI: 10.1038/nbt.2676.

(41) ASShauer, K. P.; Wemheuer, B.; Daniel, R.; Meinicke, P. *Bioinformatics* **2015**, DOI: 10.1093/bioinformatics/btv287.

(42) Jun, S. R.; Robeson, M. S.; Hauser, L. J.; Schadt, C. W.; Gorin, A. A. *BMC Research Notes* **2015**, DOI: 10.1186/s13104-015-1462-8.

(43) Bonnet, R.; Suau, A.; Doré, J.; Gibson, G. R.; Collins, M. D. *International Journal of Systematic and Evolutionary Microbiology* **2002**, DOI: 10.1099/ijs.0.01755-0.

(44) Walker, A. W.; Martin, J. C.; Scott, P.; Parkhill, J.; Flint, H. J.; Scott, K. P. *Microbiome* **2015**, DOI: 10.1186/s40168-015-0087-4.

(45) Soergel, D. A.; Dey, N.; Knight, R.; Brenner, S. E. *ISME Journal* **2012**, DOI: 10.1038/ismej.2011.208.

(46) Liu, Z.; Lozupone, C.; Hamady, M.; Bushman, F. D.; Knight, R. *Nucleic Acids Research* **2007**, DOI: 10.1093/nar/gkm541.

(47) Consortium, T. I. H. (R. N. *Cell Host and Microbe* **2014**, DOI: 10.1016/j.chom.2014.08.014.

(48) Walters, W. A.; Caporaso, J. G.; Lauber, C. L.; Berg-Lyons, D.; Fierer, N.; Knight, R. *Bioinformatics* **2011**, DOI: 10.1093/bioinformatics/btr087.

(49) Scholz, M.; Ward, D. V.; Pasolli, E.; Tolio, T.; Zolfo, M.; Asnicar, F.; Truong, D. T.; Tett, A.; Morrow, A. L.; Segata, N. *Nature Methods* **2016**, DOI: 10.1038/nmeth.3802.

(50) Korem, T.; Zeevi, D.; Suez, J.; Weinberger, A.; Avnit-Sagi, T.; Pompan-Lotan, M.; Matot, E.; Jona, G.; Harmelin, A.; Cohen, N.; Sirota-Madi, A.; Thaiss, C. A.; Pevsner-Fischer, M.; Sorek, R.; Xavier, R. J.; Elinav, E.; Segal, E. *Science* **2015**, DOI: 10.1126/science.aac4812.

(51) Mukherjee, S.; Seshadri, R.; Varghese, N. J.; Eloe-Fadrosh, E. A.; Meier-Kolthoff, J. P.; Göker, M.; Coates, R. C.; Hadjithomas, M.; Pavlopoulos, G. A.; Paez-Espino, D.; Yoshikuni, Y.; Visel, A.; Whitman, W. B.; Garrity, G. M.; Eisen, J. A.; Hugenholtz, P.; Pati, A.; Ivanova, N. N.; Woyke, T.; Klenk, H. P.; Kyrpides, N. C. *Nature Biotechnology* **2017**, DOI: 10.1038/nbt.3886.

(52) Sangwan, N.; Xia, F.; Gilbert, J. A. *Microbiome* **2016**, DOI: 10.1186/s40168-016-0154-5.

(53) Bikel, S.; Valdez-Lara, A.; Cornejo-Granados, F.; Rico, K.; Canizales-Quinteros, S.; Soberón, X.; Del Pozo-Yauner, L.; Ochoa-Leyva, A. *Computational and Structural Biotechnology Journal* **2015**, DOI: 10.1016/j.csbj.2015.06.001.

(54) Giannoukos, G.; Ciulla, D. M.; Huang, K.; Haas, B. J.; Izard, J.; Levin, J. Z.; Livny, J.; Earl, A. M.; Gevers, D.; Ward, D. V.; Nusbaum, C.; Birren, B. W.; Gnirke, A. *Genome Biology* **2012**, DOI: 10.1186/gb-2012-13-3-r23.

(55) Sultan, M.; Amstislavskiy, V.; Risch, T.; Schuette, M.; Dökel, S.; Ralser, M.; Balzereit, D.; Lehrach, H.; Yaspo, M. L. *BMC Genomics* **2014**, DOI: 10.1186/1471-2164-15-675.

(56) Peano, C.; Pietrelli, A.; Consolandi, C.; Rossi, E.; Petiti, L.; Tagliabue, L.; De Bellis, G.; Landini, P. *Microbial Informatics and Experimentation* **2013**, DOI: 10.1186/2042-5783-3-1.

(57) Maurice, C. F.; Haiser, H. J.; Turnbaugh, P. J. *Cell* **2013**, DOI: 10.1016/j.cell.2012.10.052.

(58) 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.; Knight, R. *mSystems* **2017**, DOI: 10.1128/msystems.00191-16.

(59) Wood, D. E.; Salzberg, S. L. *Genome Biology* **2014**, DOI: 10.1186/gb-2014-15-3-r46.

(60) Huson, D. H.; Beier, S.; Flade, I.; Górska, A.; El-Hadidi, M.; Mitra, S.; Ruscheweyh, H. J.; Tappu, R. *PLoS Computational Biology* **2016**, DOI: 10.1371/journal.pcbi.1004957.

(61) Abubucker, S.; Segata, N.; Goll, J.; Schubert, A. M.; Izard, J.; Cantarel, B. L.; Rodriguez-Mueller, B.; Zucker, J.; Thiagarajan, M.; Henrissat, B.; White, O.; Kelley, S. T.; Methé, B.; Schloss, P. D.; Gevers, D.; Mitreva, M.; Huttenhower, C. *PLoS Computational Biology* **2012**, DOI: 10.1371/journal.pcbi.1002358.

(62) Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A. A.; Dvorkin, M.; Kulikov, A. S.; Lesin, V. M.; Nikolenko, S. I.; Pham, S.; Prjibelski, A. D.; Pyshkin, A. V.; Sirotkin, A. V.; Vyahhi, N.; Tesler, G.; Alekseyev, M. A.; Pevzner, P. A. *Journal of Computational Biology* **2012**, DOI: 10.1089/cmb.2012.0021.

(63) Li, D.; Liu, C. M.; Luo, R.; Sadakane, K.; Lam, T. W. *Bioinformatics* **2015**, DOI: 10.1093/bioinformatics/btv033.

(64) Metcalf, J. L.; Xu, Z. Z.; Weiss, S.; Lax, S.; Van Treuren, W.; Hyde, E. R.; Song, S. J.; Amir, A.; Larsen, P.; Sangwan, N.; Haarmann, D.; Humphrey, G. C.; Ackermann, G.; Thompson, L. R.; Lauber, C.; Bibat, A.; Nicholas, C.; Gebert, M. J.; Petrosino, J. F.; Reed, S. C.; Gilbert, J. A.; Lynne, A. M.; Bucheli, S. R.; Carter, D. O.; Knight, R. *Science* **2016**, DOI: 10.1126/science.aad2646.

(65) Subramanian, S.; Huq, S.; Yatsunenko, T.; Haque, R.; Mahfuz, M.; Alam, M. A.; Benezra, A.; Destefano, J.; Meier, M. F.; Muegge, B. D.; Barratt, M. J.; VanArendonk, L. G.; Zhang, Q.; Province, M. A.; Petri, W. A.; Ahmed, T.; Gordon, J. I. *Nature* **2014**, DOI: 10.1038/nature13421.

(66) Knights, D.; Kuczynski, J.; Charlson, E. S.; Zaneveld, J.; Mozer, M. C.; Collman, R. G.; Bushman, F. D.; Knight, R.; Kelley, S. T. *Nature Methods* **2011**, DOI: 10.1038/nmeth.1650.

(67) Lax, S.; Smith, D. P.; Hampton-Marcell, J.; Owens, S. M.; Handley, K. M.; Scott, N. M.; Gibbons, S. M.; Larsen, P.; Shogan, B. D.; Weiss, S.; Metcalf, J. L.; Ursell, L. K.; Vázquez-Baeza, Y.; Van Treuren, W.; Hasan, N. A.; Gibson, M. K.; Colwell, R.; Dantas, G.; Knight, R.; Gilbert, J. A. *Science* **2014**, DOI: 10.1126/science.1254529.

(68) Quince, C.; Walker, A. W.; Simpson, J. T.; Loman, N. J.; Segata, N. *Nature Biotechnology* **2017**, DOI: 10.1038/nbt.3935.

(69) Emerson, J. B.; Adams, R. I.; Román, C. M.; Brooks, B.; Coil, D. A.; Dahlhausen, K.; Ganz, H. H.; Hartmann, E. M.; Hsu, T.; Justice, N. B.; Paulino-Lima, I. G.; Luongo, J. C.; Lymperopoulou, D. S.; Gomez-Silvan, C.; Rothschild-Mancinelli, B.; Balk, M.; Huttenhower, C.; Nocker, A.; Vaishampayan, P.; Rothschild, L. J. *Microbiome* **2017**, DOI: 10.1186/s40168-017-0285-3.

(70) Wang, Y.; Hayatsu, M.; Fujii, T. *Microbes and Environments* **2012**, DOI: 10.1264/jsme2.ME11304.

(71) Tveit, A. T.; Urich, T.; Svenning, M. M. *Applied and Environmental Microbiology* **2014**, DOI: 10.1128/AEM.01030-14.

(72) Franzosa, E. A.; Morgan, X. C.; Segata, N.; Waldron, L.; Reyes, J.; Earl, A. M.; Giannoukos, G.; Boylan, M. R.; Ciulla, D.; Gevers, D.; Izard, J.; Garrett, W. S.; Chan, A. T.; Huttenhower, C. *Proceedings of the National Academy of Sciences of the United States of America* **2014**, DOI: 10.1073/pnas.1319284111.

(73) Bashiardes, S.; Zilberman-Schapira, G.; Elinav, E. *Bioinformatics and Biology Insights* **2016**, DOI: 10.4137/BBI.S34610.

(74) Glenn, T. C. *Molecular Ecology Resources* **2011**, DOI: 10.1111/j.1755-0998.2011.03024.x.

(75) Kunin, V.; Engelbrektson, A.; Ochman, H.; Hugenholtz, P. *Environmental Microbiology* **2010**, DOI: 10.1111/j.1462-2920.2009.02051.x.

(76) Reeder, J.; Knight, R. *Nature Methods* **2009**, DOI: 10.1038/nmeth0909-636.

(77) Schloss, P. D.; Westcott, S. L.; Ryabin, T.; Hall, J. R.; Hartmann, M.; Hollister, E. B.; Lesniewski, R. A.; Oakley, B. B.; Parks, D. H.; Robinson, C. J.; Sahl, J. W.; Stres, B.; Thallinger, G. G.; Van Horn, D. J.; Weber, C. F. *Applied and Environmental Microbiology* **2009**, DOI: 10.1128/AEM.01541-09.

(78) Callahan, B. J.; McMurdie, P. J.; Holmes, S. P. *ISME Journal* **2017**, DOI: 10.1038/ismej.2017.119.

(79) Eren, A. M.; Maignien, L.; Sul, W. J.; Murphy, L. G.; Grim, S. L.; Morrison, H. G.; Sogin, M. L. *Methods in Ecology and Evolution* **2013**, DOI: 10.1111/2041-210X.12114.

(80) Callahan, B. J.; McMurdie, P. J.; Rosen, M. J.; Han, A. W.; Johnson, A. J. A.; Holmes, S. P. *Nature Methods* **2016**, DOI: 10.1038/nmeth.3869.

(81) Lozupone, C. A.; Li, M.; Campbell, T. B.; Flores, S. C.; Linderman, D.; Gebert, M. J.; Knight, R.; Fontenot, A. P.; Palmer, B. E. *Cell Host and Microbe* **2013**, DOI: 10.1016/j.chom.2013.08.006.

(82) Wang, Q.; Garrity, G. M.; Tiedje, J. M.; Cole, J. R. *Applied and Environmental Microbiology* **2007**, DOI: 10.1128/AEM.00062-07.

(83) McDonald, D.; Price, M. N.; Goodrich, J.; Nawrocki, E. P.; Desantis, T. Z.; Probst, A.; Andersen, G. L.; Knight, R.; Hugenholtz, P. *ISME Journal* **2012**, DOI: 10.1038/ismej.2011.139.

(84) Kuczynski, J.; Liu, Z.; Lozupone, C.; McDonald, D.; Fierer, N.; Knight, R. *Nature Methods* **2010**, DOI: 10.1038/nmeth.1499.

(85) Olm, M. R.; Butterfield, C. N.; Copeland, A.; Boles, T. C.; Thomas, B. C.; Banfield, J. F. *mBio* **2017**, DOI: 10.1128/mBio.01969-16.

(86) Langmead, B.; Salzberg, S. L. *Nature Methods* **2012**, DOI: 10.1038/nmeth.1923.

(87) Kim, D.; Song, L.; Breitwieser, F. P.; Salzberg, S. L. *Genome Research* **2016**, DOI: 10.1101/gr.210641.116.

(88) McIntyre, A. B.; Ounit, R.; Afshinnekoo, E.; Prill, R. J.; Hénaff, E.; Alexander, N.; Minot, S. S.; Danko, D.; Foox, J.; Ahsanuddin, S.; Tighe, S.; Hasan, N. A.; Subramanian, P.; Moffat, K.; Levy, S.; Lonardi, S.; Greenfield, N.; Colwell, R. R.; Rosen, G. L.; Mason, C. E. *Genome Biology* **2017**, DOI: 10.1186/s13059-017-1299-7.

(89) Truong, D. T.; Franzosa, E. A.; Tickle, T. L.; Scholz, M.; Weingart, G.; Pasolli, E.; Tett, A.; Huttenhower, C.; Segata, N. *Nature Methods* **2015**, DOI: 10.1038/nmeth.3589.

(90) Nguyen, N. P.; Mirarab, S.; Liu, B.; Pop, M.; Warnow, T. *Bioinformatics* **2014**, DOI: 10.1093/bioinformatics/btu721.

(91) O'Leary, N. A.; Wright, M. W.; Brister, J. R.; Ciufo, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D.; Astashyn, A.; Badretdin, A.; Bao, Y.; Blinkova, O.; Brover, V.; Chetvernin, V.; Choi, J.; Cox, E.; Ermolaeva, O.; Farrell, C. M.; Goldfarb, T.; Gupta, T.; Haft, D.; Hatcher, E.; Hlavina, W.; Joardar, V. S.; Kodali, V. K.; Li, W.; Maglott, D.; Masterson, P.; McGarvey, K. M.; Murphy, M. R.; O'Neill, K.; Pujar, S.; Rangwala, S. H.; Rausch, D.; Riddick, L. D.; Schoch, C.; Shkeda, A.; Storz, S. S.; Sun, H.; Thibaud-Nissen, F.; Tolstoy, I.; Tully, R. E.; Vatsan, A. R.; Wallin, C.; Webb, D.; Wu, W.; Landrum, M. J.; Kimchi, A.; Tatusova, T.; DiCuccio, M.; Kitts, P.; Murphy, T. D.; Pruitt, K. D. *Nucleic Acids Research* **2016**, DOI: 10.1093/nar/gkv1189.

(92) Finn, R. D.; Coggill, P.; Eberhardt, R. Y.; Eddy, S. R.; Mistry, J.; Mitchell, A. L.; Potter, S. C.; Punta, M.; Qureshi, M.; Sangrador-Vegas, A.; Salazar, G. A.; Tate, J.; Bateman, A. *Nucleic Acids Research* **2016**, DOI: 10.1093/nar/gkv1344.

(93) Suzek, B. E.; Wang, Y.; Huang, H.; McGarvey, P. B.; Wu, C. H. *Bioinformatics* **2015**, DOI: 10.1093/bioinformatics/btu739.

(94) Markowitz, V. M.; Chen, I. M. A.; Palaniappan, K.; Chu, K.; Szeto, E.; Grechkin, Y.; Ratner, A.; Jacob, B.; Huang, J.; Williams, P.; Huntemann, M.; Anderson, I.; Mavromatis, K.; Ivanova, N. N.; Kyrpides, N. C. *Nucleic Acids Research* **2012**, DOI: 10.1093/nar/gkr1044.

(95) Arndt, D.; Grant, J. R.; Marcu, A.; Sajed, T.; Pon, A.; Liang, Y.; Wishart, D. S. *Nucleic acids research* **2016**, DOI: 10.1093/nar/gkw387.

(96) Gibson, M. K.; Forsberg, K. J.; Dantas, G. *ISME Journal* **2015**, DOI: 10.1038/ismej.2014.106.

(97) Prestat, E.; David, M. M.; Hultman, J.; Ta, N.; Lamendella, R.; Dvornik, J.; Mackelprang, R.; Myrold, D. D.; Jumpponen, A.; Tringe, S. G.; Holman, E.; Mavromatis, K.; Jansson, J. K. *Nucleic Acids Research* **2014**, DOI: 10.1093/nar/gku702.

(98) Xiao, L.; Feng, Q.; Liang, S.; Sonne, S. B.; Xia, Z.; Qiu, X.; Li, X.; Long, H.; Zhang, J.; Zhang, D.; Liu, C.; Fang, Z.; Chou, J.; Glanville, J.; Hao, Q.; Kotowska, D.; Colding, C.; Licht, T. R.; Wu, D.; Yu, J.; Sung, J. J. Y.; Liang, Q.; Li, J.; Jia, H.; Lan, Z.; Tremaroli, V.; Dworzynski, P.; Nielsen, H. B.; Bäckhed, F.; Doré, J.; Le Chatelier, E.; Ehrlich, S. D.; Lin, J. C.; Arumugam, M.; Wang, J.; Madsen, L.; Kristiansen, K. *Nature Biotechnology* **2015**, DOI: 10.1038/nbt.3353.

(99) Qin, J.; Li, R.; Raes, J.; Arumugam, M.; Burgdorf, K. S.; Manichanh, C.; Nielsen, T.; Pons, N.; Levenez, F.; Yamada, T.; Mende, D. R.; Li, J.; Xu, J.; Li, S.; Li, D.; Cao, J.; Wang, B.; Liang, H.; Zheng, H.; Xie, Y.; Tap, J.; Lepage, P.; Bertalan, M.; Batto, J. M.; Hansen, T.; Le Paslier, D.; Linneberg, A.; Nielsen, H. B.; Pelletier, E.; Renault, P.; Sicheritz-Ponten, T.; Turner, K.; Zhu, H.; Yu, C.; Li, S.; Jian, M.; Zhou, Y.; Li, Y.; Zhang, X.; Li, S.; Qin, N.; Yang, H.; Wang, J.; Brunak, S.; Doré, J.; Guarner, F.; Kristiansen, K.; Pedersen, O.; Parkhill, J.; Weissenbach, J.; Bork, P.; Ehrlich, S. D.; Wang, J.; Antolin, M.; Artiguenave, F.; Blottiere, H.; Borruel, N.; Bruls, T.; Casellas, F.; Chervaux, C.; Cultrone, A.; Delorme, C.; Denariaz, G.; Dervyn, R.; Forte, M.; Friss, C.; Van De Guchte, M.; Guedon, E.; Haimet, F.; Jamet, A.; Juste, C.; Kaci, G.; Kleerebezem, M.; Knol, J.; Kristensen, M.; Layec, S.; Le Roux, K.; Leclerc, M.; Maguin, E.; Melo Minardi, R.; Oozeer, R.; Rescigno, M.; Sanchez, N.; Tims, S.; Torrejon, T.; Varela, E.; De Vos, W.; Winogradsky, Y.; Zoetendal, E. *Nature* **2010**, DOI: 10.1038/nature08821.

(100) Medema, M. H.; Blin, K.; Cimermancic, P.; De Jager, V.; Zakrzewski, P.; Fischbach, M. A.; Weber, T.; Takano, E.; Breitling, R. *Nucleic Acids Research* **2011**, DOI: 10.1093/nar/gkr466.

(101) Howe, A. C.; Jansson, J. K.; Malfatti, S. A.; Tringe, S. G.; Tiedje, J. M.; Brown, C. T. *Proceedings of the National Academy of Sciences of the United States of America* **2014**, DOI: 10.1073/pnas.1402564111.

(102) Ye, Y.; Tang, H. *Bioinformatics* **2016**, DOI: 10.1093/bioinformatics/btv510.

(103) Narayanasamy, S.; Jarosz, Y.; Muller, E. E.; Heintz-Buschart, A.; Herold, M.; Kaysen, A.; Laczny, C. C.; Pinel, N.; May, P.; Wilmes, P. *Genome Biology* **2016**, DOI: 10.1186/s13059-016-1116-8.

(104) Hug, L. A.; Baker, B. J.; Anantharaman, K.; Brown, C. T.; Probst, A. J.; Castelle, C. J.; Butterfield, C. N.; Hernsdorf, A. W.; Amano, Y.; Ise, K.; Suzuki, Y.; Dudek, N.; Relman, D. A.; Finstad, K. M.; Amundson, R.; Thomas, B. C.; Banfield, J. F. *Nature Microbiology* **2016**, DOI: 10.1038/nmicrobiol.2016.48.

(105) Vollmers, J.; Wiegand, S.; Kaster, A. K. *PLoS ONE* **2017**, DOI: 10.1371/journal.pone.0169662.

(106) Wu, Y. W.; Simmons, B. A.; Singer, S. W. *Bioinformatics* **2016**, DOI: 10.1093/bioinformatics/btv638.

(107) Alneberg, J.; Bjarnason, B. S.; De Bruijn, I.; Schirmer, M.; Quick, J.; Ijaz, U. Z.; Lahti, L.; Loman, N. J.; Andersson, A. F.; Quince, C. *Nature Methods* **2014**, DOI: 10.1038/nmeth.3103.

(108) Parks, D. H.; Imelfort, M.; Skennerton, C. T.; Hugenholtz, P.; Tyson, G. W. *Genome Research* **2015**, DOI: 10.1101/gr.186072.114.

(109) Laczny, C. C.; Sternal, T.; Plugaru, V.; Gawron, P.; Atashpendar, A.; Margossian, H. H.; Coronado, S.; der Maaten, L. V.; Vlassis, N.; Wilmes, P. *Microbiome* **2015**, DOI: 10.1186/s40168-014-0066-1.

(110) Eren, A. M.; Esen, O. C.; Quince, C.; Vineis, J. H.; Morrison, H. G.; Sogin, M. L.; Delmont, T. O. *PeerJ* **2015**, DOI: 10.7717/peerj.1319.

(111) Allen White III, R.; Brown, J.; Colby, S.; Overall, C. C.; Lee, J.-Y.; Zucker, J.; Glaesemann, K. R. G.; Jansson, C.; Jansson, J. K. *PeerJ* **2017**, DOI: 10.7287/peerj.preprints.2843v1.

(112) Treangen, T. J.; Koren, S.; Sommer, D. D.; Liu, B.; Astrovskaya, I.; Ondov, B.; Darling, A. E.; Phillippy, A. M.; Pop, M. *Genome Biology* **2013**, DOI: 10.1186/gb-2013-14-1-r2.

(113) Robinson, M. D.; McCarthy, D. J.; Smyth, G. K. *Bioinformatics* **2009**, DOI: 10 . 1093 / bioinformatics/btp616.

(114) Anders, S.; Huber, W. *Genome Biology* **2010**, DOI: 10.1186/gb-2010-11-10-r106.

(115) Sczyrba, A.; Hofmann, P.; Belmann, P.; Koslicki, D.; Janssen, S.; Dröge, J.; Gregor, I.; Majda, S.; Fiedler, J.; Dahms, E.; Bremges, A.; Fritz, A.; Garrido-Oter, R.; Jørgensen, T. S.; Shapiro, N.; Blood, P. D.; Gurevich, A.; Bai, Y.; Turaev, D.; Demaere, M. Z.; Chikhi, R.; Nagarajan, N.; Quince, C.; Meyer, F.; Balvoiut, M.; Hansen, L. H.; Sørensen, S. J.; Chia, B. K.; Denis, B.; Froula, J. L.; Wang, Z.; Egan, R.; Don Kang, D.; Cook, J. J.; Deltel, C.; Beckstette, M.; Lemaitre, C.; Peterlongo, P.; Rizk, G.; Lavenier, D.; Wu, Y. W.; Singer, S. W.; Jain, C.; Strous, M.; Klingenberg, H.; Meinicke, P.; Barton, M. D.; Lingner, T.; Lin, H. H.; Liao, Y. C.; Silva, G. G. Z.; Cuevas, D. A.; Edwards, R. A.; Saha, S.; Piro, V. C.; Renard, B. Y.; Pop, M.; Klenk, H. P.; Göker, M.; Kyrpides, N. C.; Woyke, T.; Vorholt, J. A.; Schulze-Lefert, P.; Rubin, E. M.; Darling, A. E.; Rattei, T.; McHardy, A. C. *Nature Methods* **2017**, DOI: 10.1038/nmeth.4458.

(116) Barwell, L. J.; Isaac, N. J.; Kunin, W. E. *Journal of Animal Ecology* **2015**, DOI: 10.1111/1365-2656.12362.

(117) Hamady, M.; Lozupone, C.; Knight, R. *ISME Journal* **2010**, DOI: 10.1038/ismej.2009.97.

(118) Dixon, P. *Journal of Vegetation Science* **2003**, DOI: 10.1111/j.1654-1103.2003.tb02228.x.

(119) Anderson, M. J.; Walsh, D. C. *Ecological Monographs* **2013**, DOI: 10.1890/12-2010.1.

(120) Weiss, S.; Xu, Z. Z.; Peddada, S.; Amir, A.; Bittinger, K.; Gonzalez, A.; Lozupone, C.; Zaneveld, J. R.; Vázquez-Baeza, Y.; Birmingham, A.; Hyde, E. R.; Knight, R. *Microbiome* **2017**, DOI: 10.1186/s40168-017-0237-y.

(121) McMurdie, P. J.; Holmes, S. *PLoS Computational Biology* **2014**, DOI: 10.1371/journal.pcbi.1003531.

(122) Vázquez-Baeza, Y.; Pirrung, M.; Gonzalez, A.; Knight, R. *GigaScience* **2013**, DOI: 10.1186/2047-217X-2-16.

(123) Jones, M. C.; Aitchison, J. *Journal of the Royal Statistical Society. Series A (General)* **1987**, DOI: 10.2307/2982045.

(124) Mandal, S.; Van Treuren, W.; White, R. A.; Eggesbø, M.; Knight, R.; Peddada, S. D. *Microbial Ecology in Health & Disease* **2015**, DOI: 10.3402/mehd.v26.27663.

(125) Weiss, S.; Van Treuren, W.; Lozupone, C.; Faust, K.; Friedman, J.; Deng, Y.; Xia, L. C.; Xu, Z. Z.; Ursell, L.; Alm, E. J.; Birmingham, A.; Cram, J. A.; Fuhrman, J. A.; Raes, J.; Sun, F.; Zhou, J.; Knight, R. *ISME Journal* **2016**, DOI: 10.1038/ismej.2015.235.

(126) Lovell, D.; Pawlowsky-Glahn, V.; Egozcue, J. J.; Marguerat, S.; Bähler, J. *PLoS Computational Biology* **2015**, DOI: 10.1371/journal.pcbi.1004075.

(127) Friedman, J.; Alm, E. J. *PLoS Computational Biology* **2012**, DOI: 10.1371/journal.pcbi.1002687.

(128) Kurtz, Z. D.; Müller, C. L.; Miraldi, E. R.; Littman, D. R.; Blaser, M. J.; Bonneau, R. A. *PLoS Computational Biology* **2015**, DOI: 10.1371/journal.pcbi.1004226.

(129) Schwager, E.; Mallick, H.; Ventz, S.; Huttenhower, C. *PLoS Computational Biology* **2017**, DOI: 10.1371/journal.pcbi.1005852.

(130) Washburne, A. D.; Silverman, J. D.; Leff, J. W.; Bennett, D. J.; Darcy, J. L.; Mukherjee, S.; Fierer, N.; David, L. A. *PeerJ* **2017**, DOI: 10.7717/peerj.2969.

(131) Silverman, J. D.; Washburne, A. D.; Mukherjee, S.; David, L. A. *eLife* **2017**, DOI: 10.7554/eLife.21887.

(132) Morton, J. T.; Sanders, J.; Quinn, R. A.; McDonald, D.; Gonzalez, A.; Vázquez-Baeza, Y.; Navas-Molina, J. A.; Song, S. J.; Metcalf, J. L.; Hyde, E. R.; Lladser, M.; Dorrestein, P. C.; Knight, R. *mSystems* **2017**, DOI: 10.1128/msystems.00162-16.

(133) Vandeputte, D.; Kathagen, G.; D'Hoe, K.; Vieira-Silva, S.; Valles-Colomer, M.; Sabino, J.; Wang, J.; Tito, R. Y.; De Commer, L.; Darzi, Y.; Vermeire, S.; Falony, G.; Raes, J. *Nature* **2017**, DOI: 10.1038/nature24460.

(134) Kleyer, H.; Tecon, R.; Or, D. *Frontiers in Microbiology* **2017**, DOI: 10.3389/fmicb.2017.02017.

(135) Knights, D.; Parfrey, L. W.; Zaneveld, J.; Lozupone, C.; Knight, R. Human-associated microbial signatures: Examining their predictive value., 2011.

(136) Yazdani, M.; Taylor, B. C.; Debelius, J. W.; Li, W.; Knight, R.; Smarr, L. *2016 IEEE International Conference on Big Data* **2016**, DOI: 10.1109/BigData.2016.7840731.

(137) Huang, S.; Li, R.; Zeng, X.; He, T.; Zhao, H.; Chang, A.; Bo, C.; Chen, J.; Yang, F.; Knight, R.; Liu, J.; Davis, C.; Xu, J. *ISME Journal* **2014**, DOI: 10.1038/ismej.2014.32.

(138) Teng, F.; Yang, F.; Huang, S.; Bo, C.; Xu, Z. Z.; Amir, A.; Knight, R.; Ling, J.; Xu, J. *Cell Host and Microbe* **2015**, DOI: 10.1016/j.chom.2015.08.005.

(139) Roume, H.; El Muller, E.; Cordes, T.; Renaut, J.; Hiller, K.; Wilmes, P. *ISME Journal* **2013**, DOI: 10.1038/ismej.2012.72.

(140) Witten, D. M.; Tibshirani, R. J. *Statistical Applications in Genetics and Molecular Biology* **2009**, DOI: 10.2202/1544-6115.1470.

(141) Nicholson, J. K.; Lindon, J. C. *Nature* **2008**, DOI: 10.1038/4551054a.

(142) Wang, R.; Seyedsayamdost, M. R. *Nature Reviews Chemistry* **2017**, DOI: 10.1038/s41570-017-0021.

(143) Huan, T.; Forsberg, E.; Rinehart, D.; Johnson, C.; Ivanisevic, J.; Benton, H.; Fang, M.; Aisporna, A. *Nature Publishing Group* **2017**, DOI: 10.1038/nmeth.4260.

(144) Hurley, J.; Cattell, R. *Behavioral Science* **2007**, DOI: 10.1002/bs.3830070216.

(145) Mantel, N. *Cancer Research* **1967**.

(146) DOLÉDEC, S.; CHESSEL, D. *Freshwater Biology* **1994**, DOI: 10.1111/j.1365-2427.1994.tb01741.x.

(147) Boulesteix, A. L.; Strimmer, K. *Briefings in Bioinformatics* **2007**, DOI: 10.1093/bib/bbl016.

(148) Wilms, I.; Croux, C. *BMC Systems Biology* **2016**, DOI: 10.1186/s12918-016-0317-9.

(149) Wang, M. et al. *Nature Biotechnology* **2016**, DOI: 10.1038/nbt.3597.

(150) Dhanasekaran, A. R.; Pearson, J. L.; Ganesan, B.; Weimer, B. C. *BMC Bioinformatics* **2015**, DOI: 10.1186/s12859-015-0462-y.

(151) Protsyuk, I.; Melnik, A. V.; Nothias, L. F.; Rappez, L.; Phapale, P.; Aksenov, A. A.; Bouslimani, A.; Ryazanov, S.; Dorrestein, P. C.; Alexandrov, T. *Nature Protocols* **2018**, DOI: 10.1038/nprot.2017.122.

(152) McHardy, I. H.; Goudarzi, M.; Tong, M.; Ruegger, P. M.; Schwager, E.; Weger, J. R.; Graeber, T. G.; Sonnenburg, J. L.; Horvath, S.; Huttenhower, C.; McGovern, D. P.; Fornace, A. J.; Borneman, J.; Braun, J. *Microbiome* **2013**, DOI: 10.1186/2049-2618-1-17.

(153) Whiteson, K. L.; Meinardi, S.; Lim, Y. W.; Schmieder, R.; Maughan, H.; Quinn, R.; Blake, D. R.; Conrad, D.; Rohwer, F. *ISME Journal* **2014**, DOI: 10.1038/ismej.2013.229.

(154) Theriot, C. M.; Koenigsknecht, M. J.; Carlson, P. E.; Hatton, G. E.; Nelson, A. M.; Li, B.; Huffnagle, G. B.; Li, J. Z.; Young, V. B. *Nature Communications* **2014**, DOI: 10.1038/ncomms4114.

(155) Erickson, A. R.; Cantarel, B. L.; Lamendella, R.; Darzi, Y.; Mongodin, E. F.; Pan, C.; Shah, M.; Halfvarson, J.; Tysk, C.; Henrissat, B.; Raes, J.; Verberkmoes, N. C.; Fraser, C. M.; Hettich, R. L.; Jansson, J. K. *PLoS ONE* **2012**, DOI: 10.1371/journal.pone.0049138.

(156) Hultman, J.; Waldrop, M. P.; Mackelprang, R.; David, M. M.; McFarland, J.; Blazewicz, S. J.; Harden, J.; Turetsky, M. R.; McGuire, A. D.; Shah, M. B.; VerBerkmoes, N. C.; Lee, L. H.; Mavrommatis, K.; Jansson, J. K. *Nature* **2015**, DOI: 10.1038/nature14238.

(157) Jagtap, P. D.; Blakely, A.; Murray, K.; Stewart, S.; Kooren, J.; Johnson, J. E.; Rhodus, N. L.; Rudney, J.; Griffin, T. J. *Proteomics* **2015**, DOI: 10.1002/pmic.201500074.

(158) Cheng, K.; Ning, Z.; Zhang, X.; Li, L.; Liao, B.; Mayne, J.; Stintzi, A.; Figeys, D. *Microbiome* **2017**, DOI: 10.1186/s40168-017-0375-2.

(159) Ríos-Covián, D.; Ruas-Madiedo, P.; Margolles, A.; Gueimonde, M.; De los Reyes-Gavilán, C. G.; Salazar, N. *Frontiers in Microbiology* **2016**, DOI: 10.3389/fmicb.2016.00185.

(160) Balskus, E. P. *Natural Product Reports* **2015**, DOI: 10.1039/c5np00091b.

(161) Quinn, R. A.; Phelan, V. V.; Whiteson, K. L.; Garg, N.; Bailey, B. A.; Lim, Y. W.; Conrad, D. J.; Dorrestein, P. C.; Rohwer, F. L. *ISME Journal* **2016**, DOI: 10.1038/ismej.2015.207.

(162) Fang, H.; Huang, C.; Zhao, H.; Deng, M. *Bioinformatics* **2015**, DOI: 10.1093/bioinformatics/btv349.

(163) Lê Cao, K. A.; González, I.; Déjean, S. *Bioinformatics* **2009**, DOI: 10.1093/bioinformatics/btp515.

(164) Wikoff, W. R.; Anfora, A. T.; Liu, J.; Schultz, P. G.; Lesley, S. A.; Peters, E. C.; Siuzdak, G. *Proceedings of the National Academy of Sciences of the United States of America* **2009**, DOI: 10.1073/pnas.0812874106.

Chapter 4

Using Machine Learning to Identify Major Shifts in Human Gut Microbiome Protein Family Abundance in Disease

4.1 Abstract

Inflammatory Bowel Disease (IBD) is an autoimmune condition that is observed to be associated with major alterations in the gut microbiome taxonomic composition. Here we classify major changes in microbiome protein family abundances between healthy subjects and IBD patients. We use machine learning to analyze results obtained previously from computing relative abundance of ∼10,000 KEGG orthologous protein families in the gut microbiome of a set of healthy individuals and IBD patients. We develop a machine learning pipeline, involving the Kolomogorv-Smirnov test, to identify the 100 most statistically significant entries in the KEGG database. Then we use these 100 as a training set for a Random Forest classifier to determine ∼5% the KEGGs which are best at separating disease and healthy states. Lastly, we developed a Natural Language Processing classifier of the KEGG description files to predict KEGG relative over- or under- abundance. As we expand our analysis from 10,000 KEGG protein families to one million proteins identified in the gut microbiome, scalable methods for quickly identifying such

anomalies between health and disease states will be increasingly valuable for biological interpretation of sequence data.

4.2 Introduction

The exponential decline in the cost of next-generation sequencing technology and innovations in bioinformatics approaches have enabled discovery of the detailed microbial ecology of the human body that were heretofore largely unexplored. In whole genome sequencing metagenomics, the genomic DNA present in a sample is first randomly sheared and then sequenced. The output of the Illumina sequencer are “reads" which have $\sim$100 contiguous DNA bases per read. Here we use samples that have been deeply sequenced (e.g., 100-200 million reads per sample).

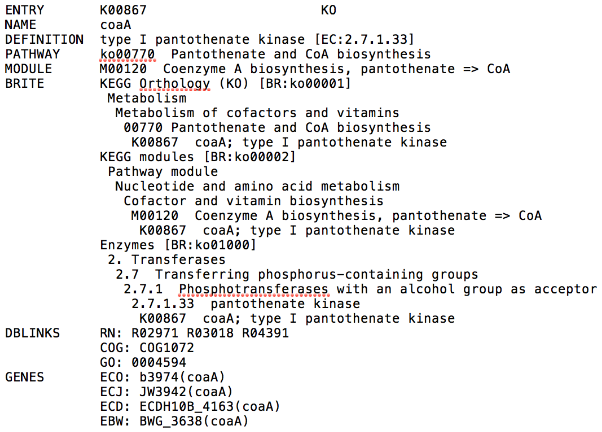
There are 80 autoimmune diseases recognized by the National Institute of Health [1], one of which is Inflammatory Bowel Disease (IBD), which is closely tied to dysbiosis of the gut microbiome [2]. Here we examine deep metagenomic sequencing of a set of healthy subjects and IBD patients to determine how microbial function changes in the health and disease state.

Despite the extensive research on the compositional changes of microbiota in IBD, the precise manner in which changes in the microbial community contributes to the disease state is only beginning to be unraveled. The microbiome DNA contains about 100 times as many genes as its human host DNA, carrying out important functions for the host, such as modulating immune development, amino acid biosynthesis, and energy harvest from food [3].

In previous work [4, 5], the bacterial species compositions were shown to be highly variable across healthy subjects, but the relative abundance of different metabolic pathways were extremely consistent between individuals and over time (see Figure 3 in [5] and Figure 2 in [4]). However, one study [5] included otherwise healthy obese individuals, and the other [4] included a cohort of individuals rigorously defined as healthy at every body site. Testing whether this pattern of constancy of metagenome-encoded functional gene frequency holds true for more acutely diseased populations is therefore of considerable interest.

This type of biological information is contained in the Kyoto Encyclopedia of Genes and Genomes

(KEGG), which is used to elucidate microbial function. KEGG is a collection of databases that contain information about genomes, biological pathways, drugs, chemicals, diseases, and protein family functions [6, 7]. In the KEGG database, each entry has a specific K number and describes an orthologous protein family with a particular biological function. Each KEGG also has a text entry such as the one shown in Figure 4.1 (accessible through the BioServices Python package [8]). Understanding the functional profiles of IBD microbiomes, including their differences in health and disease states, instead of just the taxonomic structure of microbial communities, will help inform drug development and other treatment options for patients.



```
ENTRY       K00867                      KO
NAME        coaA
DEFINITION  type I pantothenate kinase [EC:2.7.1.33]
PATHWAY     ko00770  Pantothenate and CoA biosynthesis
MODULE      M00120  Coenzyme A biosynthesis, pantothenate => CoA
BRITE       KEGG Orthology (KO) [BR:ko00001]
             Metabolism
              Metabolism of cofactors and vitamins
               00770 Pantothenate and CoA biosynthesis
                K00867  coaA; type I pantothenate kinase
            KEGG modules [BR:ko00002]
             Pathway module
              Nucleotide and amino acid metabolism
               Cofactor and vitamin biosynthesis
                M00120  Coenzyme A biosynthesis, pantothenate => CoA
                 K00867  coaA; type I pantothenate kinase
            Enzymes [BR:ko01000]
             2. Transferases
              2.7  Transferring phosphorus-containing groups
               2.7.1  Phosphotransferases with an alcohol group as acceptor
                2.7.1.33  pantothenate kinase
                 K00867  coaA; type I pantothenate kinase
DBLINKS     RN: R02971 R03018 R04391
            COG: COG1072
            GO: 0004594
GENES       ECO: b3974(coaA)
            ECJ: JW3942(coaA)
            ECD: ECDH10B_4163(coaA)
            EBW: BWG_3638(coaA)
```

Figure 4.1: An example of a KEGG description file as queried from the KEGG database for K00867.

Several studies have explored the function of the IBD microbiome using the KEGG database. For instance, Morgan et al. [9] created a broad map with 16S sequences of the gut microbiota of a large cohort of patients with IBD, and then chose a representative 11 samples to perform metagenomic sequencing and analyze with the KEGG database. This analysis revealed that moderate perturbation of microbiome composition corresponds with major perturbation of metabolic and functional pathways. Greenblum et al. [10] developed a metabolic network of KEGG enzymes to study the enzymatic variation in the gut microbiome of patients with IBD. Tong et al. [11] identified functional microbial communities using 16S rRNA sequencing, enhancing the analysis with reference sequences from Greengenes and then annotating the predicted genes with the KEGG database. Erickson et al. [12] found a number of KEGGs and KEGG pathways that were altered in Ileal Crohn's Disease.

In our previous work [13], we used deep metagenomic sequencing data, instead of predicting genes from 16S sequencing data, to compute relative abundances of the entire $\sim$10,000 entry KEGG database. Here we extend these results, using machine-learning techniques to discover the most significant over- and under- abundant KEGGs in the disease state compared to healthy subjects.

In section II (Previous Work), we discuss our data collection process and previous results. In section III (Methods), we present our proposed algorithms and workflows. We have two specific workflows. The first workflow is to identify KEGGs that are over or under abundant in disease states based on relative abundance data obtained from stool samples from healthy and disease cohorts. In the second workflow we train an NLP classifier using the KEGG description files to predict over and under abundant set of KEGGs that we identified from our first workflow. In section IV (Results) we show the results and evaluate our proposed workflows and we conduced in section (V).

4.3 Previous work

4.3.1 Cohort selection and data extraction

The three main subtypes of IBD are Ileal Crohn's Disease (ICD), Colonic Crohns Disease (CCD), and Ulcerative Colitis (UC) [14]. In our earlier research we developed a study with examples of each subtype of IBD, as well as a set of healthy subjects. The description of the set of individuals is contained in our earlier paper [13]. In summary, we downloaded 2.4 TBs of raw reads from 34 healthy individuals, 6 samples from UC and 15 from ICD selected from the NIH National Center for Biotechnology Information (NCBI) BioProjects 46321, 46881 and 43021. An additional seven samples were deeply sequenced (200 million reads per sample) by the J. Craig Venter Institute from an adult with early CCD. Table 4.1 summarizes our dataset and cohort nomenclature.

4.3.2 Feature annotation

To clarify how our previous study computed the KEGG relative abundances across our patient set, we describe the technical process we followed. First, we created a reference database of known (as of Sept 2012) gut microbe genomes consisting of 2,471 complete and 5,543 draft Bacterial and Archaeal genomes, 2,399 complete virus genomes, 26 complete Fungal genomes, and 309 HMP Eukaryote Reference

Table 4.1: Cohort sample distribution

Sample distribution for the various cohorts in our dataset.

Cohort	Abbreviation	Number of Samples
Healthy subjects	HE	34
Ulcerative colitis	UC	6
Ileal Crohn's disease	CD	15
Colonic Crohn's disease	LS	7
	Total samples:	62

Genomes, for a grand total of 10,012 genomes representing 30GB of sequences. We then used the San Diego Supercomputer Center's Gordon supercomputer to align our 6.4 billion reads from the healthy and IBD samples against the reference database we created. The database was used to calculate the relative taxonomic distribution in each sample.

In addition, high quality filtered reads were also assembled into contigs (using Velvet [15]) and Open Reading Frames (ORFs) were predicted from the contigs using Metagene [16]. Protein families were identified using the KEGG database [6, 7]. All the ORFs were aligned to KEGG sequence database using BLASTP. A curated KEGG reference database was generated by clustering all KEGG sequences at 90% sequence identity with CD-HIT [17]. If all sequences in a CD-HIT cluster belonged to the same protein orthology family (KO), the longest representative sequence was used in the reference database; otherwise all sequences were retained.

The curated database recovered more than 99% of the original hits and was 10 times faster [18]. Only the top score non-overlapping alignments from the ORFs to KEGG BLAST alignment results were used in counting the KEGG protein abundance. KEGG abundance was calculated as the number of times a KEGG protein is found in a sample, normalized against the reference protein length and predicted ORF length. The abundance of a protein family was calculated as its abundance divided by the sum of the abundance of all protein families.

The computations described above consumed 180,000 core-hours (provided by Director Michael Norman) on the Gordon supercomputer at the San Diego Supercomputer Center. About half of this time was required for the KEGG analysis.

The resulting output was a database of 10,012 KEGG entries with relative abundance for each

KEGG for each of the 62 human gut microbiome samples in Table I. This database was completed in August 2014 and is available upon request.

4.4 Methods

The dataset we examine represents a matrix of $10,012 \times 62$ or 620,744 entries. This is the matrix on which we use machine learning techniques to ascertain if there are biomedically relevant patterns in this dataset. In the near future not only will the number of samples increase by an order of magnitude, but we will be able to compute directly the relative abundance of $\sim$1 million genes, or two orders of magnitude over our current KEGG dataset. Thus, within a year, we expect our matrix to grow by three orders of magnitude in scale. This paper is our pilot to develop machine learning algorithms which will scale with the increase in data size.

First, we use Principal Component Analysis (PCA) to investigate our data set both across samples and across KEGGs. That is, we apply PCA to both the $62 \times 10,012$ matrix (a PCA across samples) and to the transpose matrix (thus a $10,012 \times 62$ matrix, or a PCA across KEGGs). This reveals insights into the structure of the data from both a samples and KEGGs perspective.

Second, we develop a KEGG relative abundance classifier that predicts over or under abundant KEGGs in the disease state compared to the healthy. Third, having classified all KEGGs, we then train a classifier using the queries from the KEGG database to predict if a KEGG is over or under abundant in disease state only using the text in the description file (example description file is shown in Figure 4.1). Note that we deploy this 2-step process since currently ground truth for which KEGGs are over or under abundant in disease state is not well understood.

4.4.1 Discrimination between healthy and IBD cohort using relative abundances

The workflow for discriminating between healthy and IBD cohorts in our samples uses the Kolmogorov-Smirnov (KS) test and Random Forests. We chose the KS test since it does not use any assumptions on the distribution of the data. We chose Random Forests since such classifiers are scale invariant, non-linear, and robust to outliers, missing values, and overfitting. Note that the overall design of

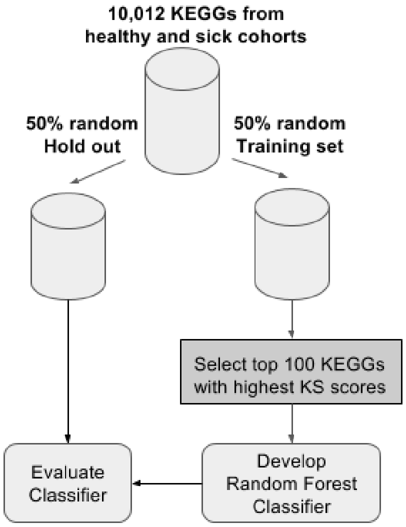


Figure 4.2: Workflow for developing a Random Forest classifier to discriminate between over and under abundant KEGGs in the diseased state. After splitting the data into the training set, we select the top 100 KEGGs that have the highest Kolomogorv-Smirnov (KS) score. We then use these 100 most significant KEGGs to train a Random Forest classifier and use to predict and evaluate on the remainder of the dataset.

the workflow is our main aim, not the specific choice of classifier (Random Forest) and statistical test (KS).

As indicated in Figure 4.2, we first randomly partition the ∼10,000 KEGGs into a 50% hold-out set and a 50% training set. We develop our classifier from the training set and apply the classifier on all KEGGs (the union of the training and hold-out sets). The primary reason for splitting the KEGGs into two sets is to simulate the scenario of scaling our workflow to new KEGGs that are introduced in the database. In other words, we simulate, in a controlled manner, the issues with developing machine learning models from databases that grow larger over time. Thus, our workflow assumes that we are training our model with a database that is smaller than the time of the application of the model.

Since there is no single ground truth on which KEGGs should be over and under abundant in disease state (as is common in many biological datasets), from the training set we use the KS test to determine the subset of the KEGGs that are the most statistically significant between the disease and healthy cohorts. From the KS test we select the 100 KEGGs with the highest KS scores, and determine whether these are over or under abundant in IBD cohorts. The over or under abundance is determined by comparing the median of the abundance for each of the 100 KEGGs between the healthy and IBD cohorts. In short, we have determined the most statistically significant over and under abundant KEGGs in disease state.

Using these 100 most significant KEGGs, we train a Random Forest classifier (defaults of [19], using 500 trees, sampling with replacement, and the square root of the number of predictors) to model if a KEGG is over or under abundant relative to the healthy state. With this classifier we can also compute a confidence score (probability that a KEGG is under abundant in the disease cohorts) as estimated by the Random Forest model to all the KEGGs in our data. This is possible as we report the KEGGs that have the highest (corresponding to under abundance of IBD cohorts relative to healthy) and lowest (corresponding to over abundance of IBD cohorts relative to health) confidence scores. Note that we take this approach of only training the Random Forest classifier on 100 KEGGs and applying to the remaining KEGGs to ensure that we do not overfit.

4.4.2 Discrimination between healthy and IBD cohorts using KEGG description files

Once we have classified all KEGGs as either over or under abundant, we can now develop a classifier that determines if a KEGG is over or under abundant in a subject based on the KEGG description file (example of a description file for one KEGG is shown in Figure 4.1). Since the KEGG description file is a text file and numerous approaches and methods are available for Natural Language Processing, we here present a baseline model to serve as a benchmark for future models. In our baseline approach, we use the “raw" KEGG description file as queried from the database [6, 7]. In this baseline approach, we extract bag-of-words unigram features weighted by Term Frequency-Inverse Document Frequency scaling (TF-IDF, as implemented in [20]), thus ignoring the word order and hierarchical structure of the description file (more sophisticated features, such as [21, 22], can be explored for future work). Our workflow for this approach is shown in Figure 4.3.

In the bag-of-words approach to NLP, in each document we count the number of occurrences of the terms in our dictionary (fixed-size vocabulary), creating a count vector that we can use as a feature vector for machine learning tasks. The idea here is that high occurrences of specific terms reflect the content or subject of the document. These raw counts of terms are referred to as “term frequencies". Since the raw counts of the terms may inadvertently weight so-called “stop words" (such as “the") that do not reflect the subject of a document, we normalize the raw counts of each term to diminish bias from commonly

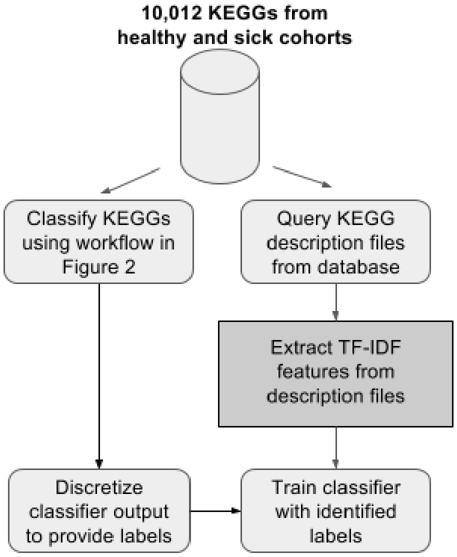

10,012 KEGGs from healthy and sick cohorts
Classify KEGGs using workflow in Figure 2
Query KEGG description files from database
Extract TF-IDF features from description files
Discretize classifier output to provide labels
Train classifier with identified labels

Figure 4.3: Workflow for developing an NLP classifier to discriminate between over and under abundant KEGGs in the diseased state using the classified labels as determined by the workflow from Figure 4.2. Each of the 10,012 KEGGs are queried in the KEGG database [6, 7] and stored as KEGG description files. We then extract Term Frequency-Inverse Document Frequency (TF-IDF) features from the description files as described in Equations 1 and 2. In parallel we classify each of the 10,012 KEGGs from our patient data using the workflow in Figure 4.2 and discretize the probabilities of the classifier into over and under abundant and “neutral" categories. These categories are then used as labels to train a classifier on the extracted TF-IDF features from the KEGG description file.

occurring words. A common normalization is to count the number of documents that contain each of the terms in our dictionary, and this count is referred to as “document frequency." The motivation with this normalization is that if a word is common then it should appear in most of the documents in the corpus that we are studying and will have a high document frequency count. We use the document frequencies to normalize the term frequencies to obtain the TF-IDF features.

In our application, we define each of the terms (also referred to as word or token) that occur in the KEGG description files as t and each of the KEGG description files as d (also referred to as a document in NLP). Our corpus of documents is the 10,012 KEGG description files and we compute the TF-IDF features for each KEGG as

$$\text{tf-idf}(t,d) = \text{tf}(t,d) \cdot \big(\text{idf}(t,d) + 1\big) \tag{4.1}$$

where $\text{tf}(t,d)$ is the number of occurrences (that is, the frequency) of term t in the KEGG description file d and $\text{idf}(t,d)$ is the normalization of this count with respect to the number of occurrences of t in all the other KEGG description files. This normalization $\text{idf}(t,d)$ is computed as:

$$\text{idf}(t,d) = \log \frac{1+N}{1+\text{df}(d,t)} \tag{4.2}$$

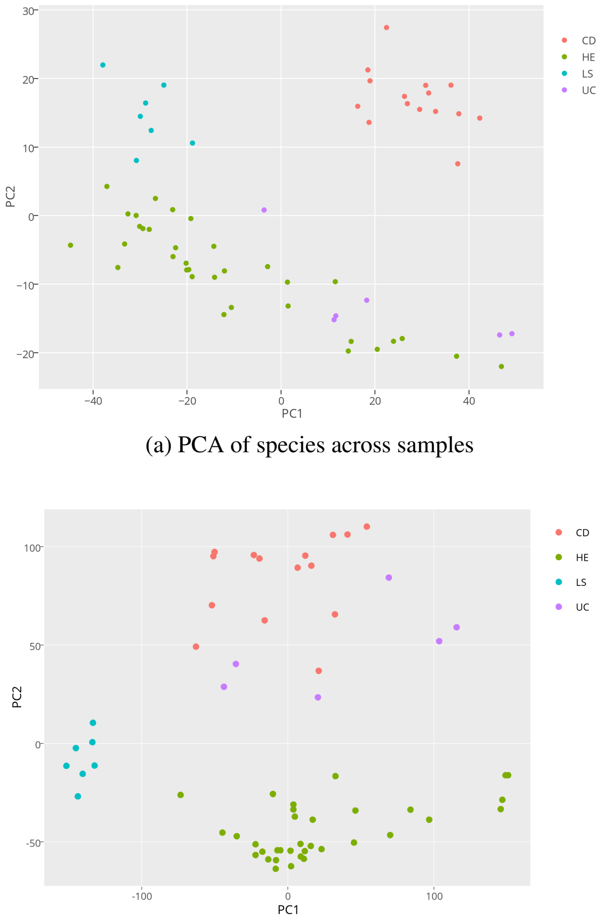
where N is the total number of KEGG description files (10,012), and $\text{df}(d,t)$ is the number of KEGG description files with the term t. Given the success of such features in NLP applications, this set of features from the description files for each KEGG serves as a baseline to develop classification models.

To develop the classification models from the bag-of-words TF-IDF features, we discretize the confidence scores of all the KEGGs in our data, as estimated by the Random Forest classifier we developed earlier, to three distinct categories: under abundant, over abundant, and neither (probability of Random Forest output greater than 0.75, less than 0.25, and between 0.25 and 0.75 respectively. These thresholds may be adjusted to allow for more False Positives/Negatives, but we do not explore adjusting these trade-offs here. Using these categories as the labels for the bag-of-words features that we computed, we proceed to train three standard classifiers (Naive Bayes, Support Vector Machine with linear kernel, and Logistic Regression) and report their average F1 score using 10-fold Cross Validation.

4.5 Results

4.5.1 Use of PCA to show KEGGs separate healthy and disease states

It has been known since 2010 (see Figure 4 in [23]) that in human gut microbiome samples the species abundances can separate healthy from IBD substates UC and ICD using PCA. Further research (see Figure 2 in [24]) showed that PCAs based on species abundance can also separate ICD into its two subtypes (CCD and ICD), as well as separating UC and healthy.



(a) PCA of species across samples

(b) PCA of KEGGs across samples

Figure 4.4: PCA of species (a) and KEGGs (b) across samples colored by the different cohorts (abbreviations and data summary shown in Table 4.1). As shown in (b), using all 10,012 KEGGs we see near perfect separation between the different cohorts. While on the other hand, in (a) using species PCA we do not see a clear separation between UC (Ulcerative colitis) and HE (healthy) groups.

Our species abundance data shows a similar separation (Figure 4a). However, we can go beyond microbial species and use PCA on the KEGG protein families by computing PCA on our data matrix with microbiome samples as rows and KEGGs as columns. This shows an even clearer separation between healthy and the three disease states (Figure 4b). Thus, it appears that there are significant differences between the KEGG relative abundances in health and each of the three IBD disease states. We next turn to using machine learning techniques to find which KEGGs are the best discriminators.

4.5.2 Classification of over and under abundant KEGGS in IBD

The Random Forest classifier that we developed according to our workflow in Figure 4.2 obtains an out-of-bag classification accuracy of 99%. Moreover, since Random Forest classifiers can give probabilistic outputs, we compute the confidence scores for how well each KEGG works as a classifier on separation of over or under abundant compared to the healthy cohort. Figure 4.5 shows the PCA of the KEGGs. We color each KEGG in the PCA scatter plot with the discretized confidence score from the Random Forest classifier [1].

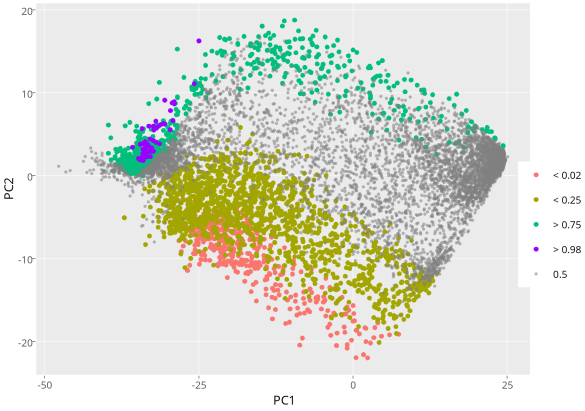


Figure 4.5: PCA of KEGGs classified with the outputs of our trained Random Forest classifier based on subject relative abundance data. The categories of over abundant, under abundant and neutral KEGGs form coherent clusters in PCA space suggesting that the classifier has not overfit and that similar KEGG distributions in our patient population are classified with a similar label.

Coloring the PCA plot with these discretized confidence scores shows that the all KEGGs are clustered according to their abundance level. That is, KEGGs that are clustered with each other in PCA space have a similar classification score as given by the Random Forest classifier. Otherwise if the classifier

[1] For the raw scores see https://plot.ly/~crude2refined/1959/pc2-vs-pc1.embed

was overfitting, then we would expect to see the color distribution of the KEGGs in the scatter plot to be distributed in a less structured and more random manner. Note also that this separation continues into the most extreme regions of confidence (over 98% and less than 2%). This set of "extreme" KEGGs corresponds to ∼500 KEGGs or 5% of all KEGGs in our dataset. These results indicate that our proposed workflow in Figure 4.2 is able to discover the best KEGG classifiers for separating over or under abundant values in IBD patients compared to healthy subjects.

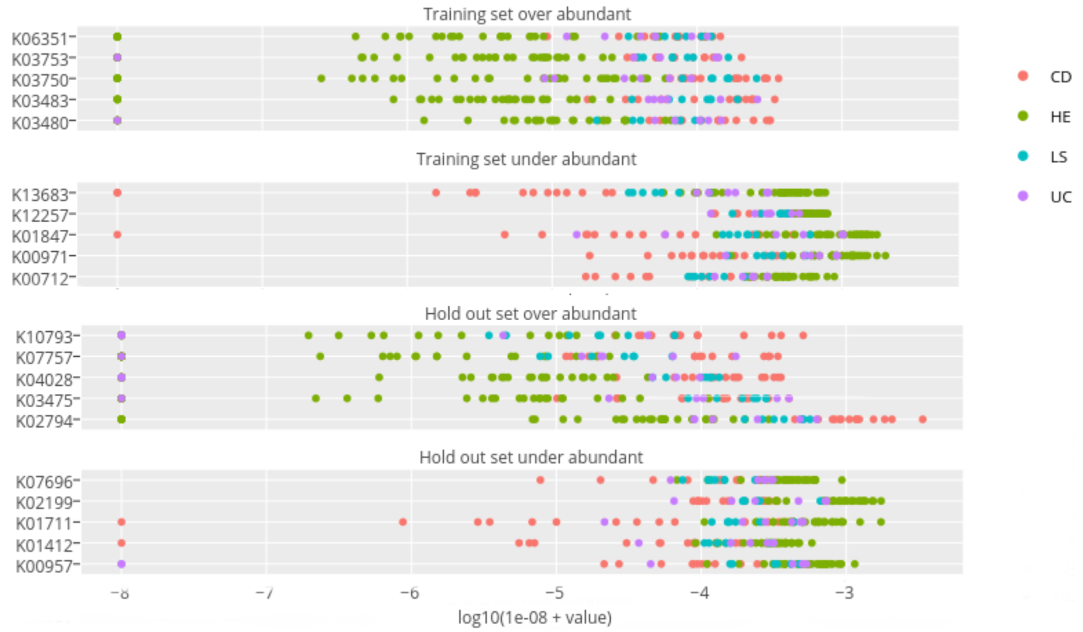


Figure 4.6: Distribution of the relative abundance of KEGGs selected by our approach that discriminate between healthy and disease states. The horizontal axis is the relative abundance values of the KEGGs on a logarithmic scale for each of the samples. See Table 4.1 for the summary of cohort samples.

Finally, we can use the classifier to find the over- and under-abundance KEGGs that most differentiate between the health and disease states. In Figure 4.6 we show a sample of the most confident KEGGs that are over and under abundant compared to the healthy cohorts for both our training and hold-out sets. This figure clearly shows the separation of the healthy and disease samples on a logarithmic scale.

It is beyond the scope of this paper to go into the biological implications of these large differences, but we note that our machine learning methodology has selected certain KEGGs that previous research has identified as important to IBD state. Notably, we find that a number of the over-abundant KEGGs identified are involved in the phospho-transferase system (PTS) (K03480, K03483, K03475, K02794, and others). The PTS is a sugar transport mechanism associated with the Firmicutes phylum, which is favored in patients with IBD [10, 25], being involved in carbohydrate uptake. Furthermore, the PTS enzyme FrvX is a known biomarker for IBD according to [26].

Another interesting over-abundant KEGG we identify is mobB (K03753), the presence of which enhances activation of nitrate reductase as discussed in Palmer et al. and Eaves et al. [27, 28]. Nitrate reduction is a critical process that produces nitric oxide, which is not synthesized by the human genome. Increased levels of nitric oxide is associated with inflammation, cancer, and IBD as several studies have shown [10, 29–31]. Under-abundant KEGGs include those that metabolize amino acids and carbohydrates (K01847, K01711, K00971, K12111, and more), which are thought to be decreased in favor of nutrient uptake in the IBD microbiome as shown in [9]. Several KEGGs involved in amino acid biosynthesis and carbohydrate metabolism are also over-abundant, reflecting the inconsistency in previous studies and the need for further analysis and more datasets.

In a future paper, we will analyze in more detail the biological significance of our hundreds of over and under abundant KEGGs that differentiate between health and IBD.

4.5.3 Development of a natural language classifier to disease association

As additional KEGGs are annotated or additional disease pathways are identified, natural language processing can help predict the association between new protein families and disease. Here, we present preliminary results on developing a baseline classifier that determines if a KEGG is over or under abundant based on the KEGG description file alone (a snippet is shown in Figure 1).

As discussed in the methods section above, we extract unigram bag-of-words TF-IDF features and train three different baseline classifiers to classify a KEGG as “over abundant," “under abundant," or “neither" based on the results of the two-stage classification above. Figure 4.5 shows the distribution of these three categories on the PCA of the KEGG relative abundances amongst the cohorts in our data. The uniformity of the distribution of these three categories in the distribution of our KEGGs suggests that the categories that we have selected for the KEGGs are sensible.

Table 4.2 shows the average F1 score for the three classifiers we considered using 10-fold cross validation. We report the F1 score since it is a more rigorous measurement of accuracy as it is the harmonic

Table 4.2: Classification accuracies using KEGG description files. Average F1 scores based on 10-fold cross validation for classifiers we trained on bag-of-words TF-IDF features from the KEGG description files to classify KEGGs as under or over abundant relative to healthy states.

Classifier	**F1-score**
Naive Bayes	0.71 ± 0.024
Support Vector Machine	0.76 ± 0.012
Logistic Regression	0.77 ± 0.007

mean of the precision and recall scores (that is, we are accounting for both type I and type II errors by not allowing class imbalance to influence our error rates). The results that we have are significantly higher than random. This suggests that our baseline TF-IDF features are able to predict if the description of the KEGGs has predictive power for discriminating between healthy and disease states. The predictive power of these features can be used in more sophisticated topic modelling approaches (such as Non-Negative Matrix Factorization [32], Latent Dirichlet Allocation [33], and word embeddings [21, 22]). Such topic models can then be used to aid biologists in comprehending the biological relationship between the numerous KEGGs in databases. We suspect that future work that takes into account the hierarchical structure and utilizes domain knowledge of the KEGG description files will significantly improve this baseline performance.

4.6 Discussion and Related Work

Microbial communities are complex networks that rely on function as well as structure. The KEGG is a well-characterized database of molecular function that has been a widely-used tool to investigate microbial function. By looking at the function of specific disease-associated microbial communities, we can better identify targets for future intervention (i.e. small molecule development to target a specific gene pathway). The motivation for using machine learning methods is to reduce the amount of time-consuming manual investigation of immense amounts of data generated from metagenomic sequencing. Using metagenomic data from a cohort of healthy and IBD-affected individuals, we developed and trained a two step classifier to identify 100 KEGG ortholog genes which are over or under abundant in IBD patients compared to healthy adults. We also demonstrated the ability of a simple natural language classifier to identify KEGGs as over or under abundant in the IBD disease state.

While there are a number of methods that could be used with KEGG protein families to discover the large changes in healthy and disease states, we turned to machine learning methods here because the next step in our project will see our matrix of samples versus function grow by a factor of 1000x, necessitating a computational approach to discovery of these patterns.

4.7 Acknowledgments

Chapter 4, in full, is a modified reprint of the material as it appears in “M. Yazdani, B. C. Taylor, L. Weizhong, J. W. Debelius, R. Knight, and L. Smarr. *Using Machine Learning to Identify Major Shifts in Human Gut Microbiome Protein Family Abundance in Disease*. IEEE Big Data (2016). DOI: 10.1109/BigData.2016.7840731.” The dissertation author was an investigator and coauthor of this work. The dissertation author contributed to research design, analysis of the Random Forest and natural language classifiers, biological interpretation of the classifier results, figure creation, and manuscript writing.

The authors would like to thank members of the UC San Diego Knight Lab and Calit2 Qualcomm Institute staff for many useful conversations. We also thank the San Diego Supercomputer Center for providing computational resources. We also acknowledge support from the Amar Foundation.

References

(1) NIH Autoimmune Diseases. Bethesda, Maryland.

(2) Walters, W. A.; Xu, Z.; Knight, R. *FEBS letters* **2014**, *588*, 4223–4233.

(3) Turnbaugh, P. J.; Ley, R. E.; Hamady, M.; Fraser-Liggett, C.; Knight, R.; Gordon, J. I. *Nature* **2007**, *449*, 804.

(4) Consortium, H. M. P. et al. *Nature* **2012**, *486*, 207–214.

(5) Turnbaugh, P. J.; Hamady, M.; Yatsunenko, T.; Cantarel, B. L.; Duncan, A.; Ley, R. E.; Sogin, M. L.; Jones, W. J.; Roe, B. A.; Affourtit, J. P., et al. *nature* **2009**, *457*, 480–484.

(6) Kanehisa, M.; Sato, Y.; Kawashima, M.; Furumichi, M.; Tanabe, M. *Nucleic acids research* **2016**, *44*, D457–D462.

(7) Kanehisa, M.; Goto, S. *Nucleic acids research* **2000**, *28*, 27–30.

(8) Cokelaer, T.; Pultz, D.; Harder, L. M.; Serra-Musach, J.; Saez-Rodriguez, J. *Bioinformatics* **2013**, *29*, 3241–3242.

(9) Morgan, X. C.; Tickle, T. L.; Sokol, H.; Gevers, D.; Devaney, K. L.; Ward, D. V.; Reyes, J. A.; Shah, S. A.; LeLeiko, N.; Snapper, S. B., et al. *Genome Biol* **2012**, *13*, R79.

(10) Greenblum, S.; Turnbaugh, P. J.; Borenstein, E. *Proceedings of the National Academy of Sciences* **2012**, *109*, 594–599.

(11) Tong, M.; Li, X.; Parfrey, L. W.; Roth, B.; Ippoliti, A.; Wei, B.; Borneman, J.; McGovern, D. P.; Frank, D. N.; Li, E., et al. *PloS one* **2013**, *8*, e80702.

(12) Erickson, A. R.; Cantarel, B. L.; Lamendella, R.; Darzi, Y.; Mongodin, E. F.; Pan, C.; Shah, M.; Halfvarson, J.; Tysk, C.; Henrissat, B., et al. *PloS one* **2012**, *7*, e49138.

(13) Wu, S.; Li, W.; Smarr, L.; Nelson, K.; Yooseph, S.; Torralba, M. In *Proceedings of the Conference on Extreme Science and Engineering Discovery Environment: Gateway to Discovery*, 2013, p 25.

(14) Cleynen, I.; Boucher, G.; Jostins, L.; Schumm, L. P.; Zeissig, S.; Ahmad, T.; Andersen, V.; Andrews, J. M.; Annese, V.; Brand, S., et al. *The Lancet* **2016**, *387*, 156–167.

(15) Zerbino, D. R.; Birney, E. *Genome research* **2008**, *18*, 821–829.

(16) Noguchi, H.; Taniguchi, T.; Itoh, T. *DNA research* **2008**, *15*, 387–396.

(17) Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. *Bioinformatics* **2012**, *28*, 3150–3152.

(18) Wu, S.; Zhu, Z.; Fu, L.; Niu, B.; Li, W. *BMC genomics* **2011**, *12*, 1.

(19) Liaw, A.; Wiener, M. *R News* **2002**, *2*, 18–22.

(20) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(21) Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. *arXiv preprint arXiv:1301.3781* **2013**.

(22) Pennington, J.; Socher, R.; Manning, C. D. In *EMNLP*, 2014; Vol. 14, pp 1532–43.

(23) Qin, J.; Li, R.; Raes, J.; Arumugam, M.; Burgdorf, K. S.; Manichanh, C.; Nielsen, T.; Pons, N.; Levenez, F.; Yamada, T., et al. *nature* **2010**, *464*, 59–65.

(24) Willing, B. P.; Dicksved, J.; Halfvarson, J.; Andersson, A. F.; Lucio, M.; Zheng, Z.; Järnerot, G.; Tysk, C.; Jansson, J. K.; Engstrand, L. *Gastroenterology* **2010**, *139*, 1844–1854.

(25) Mahowald, M. A.; Rey, F. E.; Seedorf, H.; Turnbaugh, P. J.; Fulton, R. S.; Wollam, A.; Shah, N.; Wang, C.; Magrini, V.; Wilson, R. K., et al. *Proceedings of the National Academy of Sciences* **2009**, *106*, 5859–5864.

(26) Chen, C.-S.; Sullivan, S.; Anderson, T.; Tan, A. C.; Alex, P. J.; Brant, S. R.; Cuffari, C.; Bayless, T. M.; Talor, M. V.; Burek, C. L., et al. *Molecular & Cellular Proteomics* **2009**, *8*, 1765–1776.

(27) Palmer, T.; Santini, C.-L.; Iobbi-Nivol, C.; Eaves, D. J.; Boxer, D. H.; Giordano, G. *Molecular microbiology* **1996**, *20*, 875–884.

(28) Eaves, D. J.; Palmer, T.; Boxer, D. H. *European Journal of Biochemistry* **1997**, *246*, 690–697.

(29) Kolios, G.; Valatas, V.; Ward, S. G. *Immunology* **2004**, *113*, 427–437.

(30) Yang, G.-Y.; Taboada, S.; Liao, J. *Inflammation and Cancer: Methods and Protocols: Volume 2: Molecular Analysis and Pathways* **2009**, 119–156.

(31) Winter, S. E.; Lopez, C. A.; Bäumler, A. J. *EMBO reports* **2013**, *14*, 319–327.

(32) Berry, M. W.; Browne, M. *Computational & Mathematical Organization Theory* **2005**, *11*, 249–264.

(33) Blei, D. M.; Ng, A. Y.; Jordan, M. I. *Journal of machine Learning research* **2003**, *3*, 993–1022.

Chapter 5

Depression in HIV and HCV Co-Infected Individuals is Associated with Systematic Differences in the Gut Microbiome and Metabolome

5.1 Abstract

Depression is influenced by the structure, diversity, and composition of the gut microbiome. Although depression has been described previously in human immunodeficiency virus (HIV) and hepatitis C (HCV) mono-infections, and to a lesser extent HIV-HCV coinfection, research on the interplay between depression and the gut microbiome in these disease states is limited. Here, we characterized the gut microbiome using 16S rRNA amplicon sequencing of fecal samples from 373 participants who underwent a comprehensive neuropsychiatric assessment, and the gut metabolome on a subset of these participants using untargeted metabolomics with liquid chromatography mass spectrometry. We observed that the gut microbiome and metabolome were distinct between HIV positive and negative individuals. HCV infection had a large association with the microbiome that was not confounded by drug use. Therefore, we

classified the participants by HIV and HCV infection status (HIV-monoinfected, HIV-HCV coinfected, or uninfected). The three groups significantly differed in their gut microbiome (unweighted UniFrac distances) and metabolome (BrayCurtis distances). Coinfected individuals also had lower alpha diversity. Within each of the three groups, we evaluated lifetime Major Depressive Disorder (MDD) and current Beck Depression Inventory II. We found that the gut microbiome differed between depression states only in coinfected individuals. Coinfected individuals with a lifetime history of MDD were enriched in primary and secondary bile acids, as well as taxa previously identified in people with MDD. Collectively, we observe persistent signatures associated with depression only in coinfected individuals, suggesting that HCV itself, or interactions between HCV and HIV, may drive HIV-related neuropsychiatric differences.

5.2 Importance

The human gut microbiome influences depression. Differences between the microbiomes of HIV infected and uninfected individuals have been described, but it is not known whether these are due to HIV itself, or to common HIV comorbidities such as HCV coinfection. Limited research has explored the influence of the microbiome on depression within these groups. Here, we characterized the microbial community and metabolome in the stool of 373 people, noting the presence of current or lifetime depression as well as their HIV and HCV infection status. Our findings provide additional evidence that individuals with HIV have different microbiomes which are further altered by HCV coinfection. In individuals coinfected with both HIV and HCV, we identified microbes and molecules that were associated with depression. These results suggest that the interplay of HIV and HCV and the gut microbiome may contribute to the HIV-associated neuropsychiatric problems.

5.3 Introduction

Disturbances in gut microbial communities may contribute to depression and neuropsychiatric disorders in HIV infection. For example, depletion of CD4+ T cells in gut lymphoid tissue occurs very early in HIV infection and is associated with dysbiosis and gut barrier dysfunction (leaky gut) [1, 2], which is not normalized by virologic suppression on antiretroviral therapy (ART) [3]. Leaky gut in HIV infection is associated with increased apoptosis, chronic inflammatory signals and reduced proliferation and repair of epithelial cells [1, 2, 4, 5] which may further introduce microbial metabolites known to impact brain activity [6–9]. Gut dysbiosis patterns in HIV-monoinfection may include greater proportions of Gram-negative bacteria, order Enterobacteriales [10], enrichment of Proteobacteria [11], depletion of Bacteroidia [12] and increased abundances of Prevotellaceae and Erysipelotrichaceae [13]. Some of these alterations involve pro-inflammatory species (e.g., *Prevotella*). Together dysbiosis and leaky gut render HIV-infected individuals more vulnerable to microbial antigen-driven effects on the central nervous system (CNS) via pro-inflammatory bacterial antigens such as LPS and flagellin [14, 15].

Dysbiosis-driven inflammation also may lead to depression, as suggested by existing literature [16, 17]. The gut microbiota may affect blood-brain barrier (BBB) integrity as well [18]. For example, germ-free mice have reduced expression of tight junction proteins on brain microvascular endothelial cells. BBB integrity was restored after gut colonization or by administration of butyrate [19]. BBB compromise may amplify entry of HIV and associated neurotoxins into the CNS [20]. These findings are of clinical importance, since interventions exist to restore normal gut microbes and barrier integrity (such as B. fragilis/B. thetaiotaomicron PSA [21], butyrate [22] and tryptophan metabolites [23]) with the potential to improve CNS function.

While no systematic research has been reported on the impact of HIV/HCV coinfection on the gut microbiota, a number of reports examining very different cohorts of patients with HCV monoinfection have evaluated alterations in the gut microbiota. A study of HCV patients with advanced liver disease showed increased abundance of Bacteroidetes and Firmicutes compared to healthy subjects [24]. The HCV patients had increased *Prevotella, Acinetobacter, Veillonella, Phascolarctobacterium* and *Faecalibacterium* and reduced *Ruminococcus, Clostridium*, and *Bifidobacterium* genus. Interpreting these findings is difficult, as these patients were likely treated with luminal antibiotics as prophylaxis against hepatic encephalopathy [24]. In another study of persons with HCV, bacterial diversity was lower compared with healthy individuals, with reduced Clostridiales and increased Streptococcus and Lactobacillus. Dysbiosis appeared very early, before cirrhotic changes [25]. In another report, gut microbiota alpha diversity was reduced in cirrhotic patients, but dysbiosis was significantly improved along with a reduction in serum cytokines and chemokines by the cure of HCV infection after treatment with direct-acting agents

[26]. However, another study showed that cirrhotic outpatients with HCV had similar microbiome and proinflammatory changes before and 1 year after HCV cure [27]. Thus, there is no consensus concerning changes in the gut microbiome associated with HCV, likely due to marked differences in the cohorts studied.

Abundant human and animal evidence link the gut microbiome to neuroinflammation and depressed mood. In rats treated with microbiota from rats vulnerable to social stress, there was higher microglial density and IL-1β expression in the ventral hippocampus, and higher depression-like behaviors relative to rats receiving microbiota from rats resistant to social stress, suggesting that the gut microbiome contributes to the depression-like behavior and inflammatory processes in the brain [28]. In HIV+ individuals, an abnormal microbiome in combination with leaky gut leads to high circulating levels of microbial antigens that provoke inflammation. This inflammation induces expression of indoleamine dioxygenase, which promotes depressed mood by shunting tryptophan away from serotonin synthesis [29].

Similarly, in humans without these infections, the gut microbiota can influence neuroinflammation and neuropsychiatric disorders by communication through the gutbrain axis [30]. For example, patients with major depressive disorder (MDD) showed increased Bacteroidetes, Protobacteria, and Actinobacteria, and less Firmicutes [31]. Interventions that affect the gut microbiota can be beneficial for neuropsychiatric dysfunction. For example, probiotics and prebiotics attenuated the physiological stress response. Colonizing germ-free male mice with Bifidobacterium infantis normalized their previously over-reactive hypothalamic-pituitary-adrenal axis in response to restraint stress [32]. Also, treatment with prebiotic fructo- and galacto-oligosaccharides (FOS/GOS) lowered pro-inflammatory cytokine levels in mice exposed to chronic psychosocial stress [33].

To address gaps in knowledge about the impact of co-infection with HIV and HCV on the gut microbiome, we performed 16S rRNA sequencing and metabolomics analyses on fecal samples from co-infected individuals and compared them to HIV monoinfected and HIV uninfected subjects. Despite the evident interplay between HIV infection and associated neurocognitive disorders, and between each of these and gut microbiome dysbiosis, prior work suggests that HIV infection and neurocognitive disorders are not associated with gut microbiome dysbiosis [34]. Here, we observe associations between gut microbiome dysbiosis and depression, a form of neurobehavioral disorder, only in HIV-HCV coinfected individuals.

These results suggest that HIV, HCV and the gut microbiome may work together to cause neuropsychiatric problems associated with HIV.

5.4 Results and Discussion

5.4.1 The gut microbiome and metabolome differs with HIV and HCV infection.

We first evaluated the gut microbiome and metabolome in the context of HIV infection status. As in previous studies, we found that beta diversity (i.e., between subject), but not alpha diversity (i.e., within subject), differed between HIV positive (n = 267) and negative individuals (n = 106) (Unweighted UniFrac distances [35–37]: permutational multivariate analysis of variance [PERMANOVA] pseudo-F-statistic (pseudo-F) = 4.24, Benjamini-Hochberg corrected (BH) p = 0.001). There was also a significant difference in the gut metabolome between HIV positive and negative individuals (Bray-Curtis, pseudo-F = 5.82, BH p = 0.001).

To characterize the impact of covariates on the microbiome, we performed regularized discriminant analysis [38] (RDA) to calculate the relative effect size of several covariates: sexual orientation; biological sex; HCV status; HIV status; BDI-II group; lifetime alcohol use disorders; lifetime MDD; and lifetime drug use disorders (including lifetime history of cocaine, methamphetamine, heroin, and sedative use disorders) in the unweighted UniFrac beta diversity principal coordinates analysis (PCoA). The lifetime drug use disorder categories were collinear in the PCoA, but no drug use disorder categories were collinear with HCV infection status. This suggests that history of drug use disorders does not confound HCV status. After merging colinear drug-use disorder covariates, we found that sexual orientation, HCV infection status, BDI group (mild 13, 13-19, mild; 20-28, moderate; or > 29 severely depressed current mood), and biological sex resulted in a significant RDA model (Fig. 5.1A).

Due to the large relative effect size of HCV status, we classified the participants by the presence or absence of both HIV and HCV infection (HIV-monoinfected, HIV-HCV coinfected, or uninfected; 5.1B).

5.4.2 Demographic and lifestyle comparisons between coinfected, HIV-monoinfected, and uninfected individuals.

To explore the relationship between the gut microbiome and depression in people with HIV monoinfection, HIV and HCV coinfection, or neither, we analyzed 16S rRNA gene amplicon sequencing data from a total of 571 fecal samples (Fig. 5.1B), 398 of which were from unique individuals. After filtering (see Materials and Methods), 373 samples from unique subjects (described in 5.1A) were retained for analysis. Participants were grouped according to their HIV and HCV infection state: coinfected individuals (n = 48) with both HIV and HCV, HIV-monoinfected (n = 219) with HIV but not HCV, and uninfected individuals (n = 106) with neither virus. A subset of these participants (coinfected, n = 27; HIV-monoinfected, n = 82; uninfected, n = 32) were additionally assessed using untargeted metabolomics by liquid chromatography mass spectrometry.

Sample characteristics of each infection group are included in Table 5.1. Biological sex, anal receptive intercourse, and age have been associated with differences in microbial communities [39–49]. The uninfected group had more women and significantly fewer bisexual and homosexual men (Chi2 = 76.9, $p < 0.0001$) than the other infection groups, but all three groups were similar in terms of age. The uninfected group had a higher estimated verbal IQ than the coinfected group. The uninfected group also had lower current depressive symptoms and fewer problems with activities of daily living than the HIV-monoinfected and coinfected group, and a higher rate of employment than the coinfected group. Lifetime substance use disorders were lower in the uninfected group while lifetime major depression showed a stair step pattern; with the uninfected group at 33%, the HIV-monoinfected group at 52%, and the coinfected group at 71% (all $p < 0.05$). In terms of HIV disease, the HIV-monoinfected and coinfected groups did not differ by AIDS status, current or nadir CD4, or plasma viral load detectability. While the coinfected group had been HIV positive an average of 4 years longer than the HIV-monoinfected group, they were less likely to be on ART at their study visit (85% vs 96% respectively, $p < 0.01$). The coinfected group also was more likely to be composed of minorities (specifically, African-Americans), but in all other respects (including history of substance use disorders) they were comparable to the HIV-monoinfected group.

Table 5.1: HIV and HCV group characteristics for both the full infection cohorts (A) and MSM filtered cohorts (B). T-tests were used for all normally distributed continuous variables (age, education, estimate verbal IQ, estimated duration HIV, Beck Depression Inventory II); Wilcoxon tests were used for nadir and current CD4; chi-square tests use for all nominal variables (% Caucasian, % AIDS, % undetectable HIV RNA, % cognitively impaired, % employed, % IADL dependent, % lifetime substance use disorder, % lifetime major depressive disorder, % bisexual and/or homosexual).

a.

	Uninfected (a)	HIV Mono-Infected (b)	Coinfected (c)	
n	106	219	48	
Age^{1}	51.2 (16.3)	51.7 (12.0)	53.8 (9.1)	
Education^{1}	14.4 (2.5)	14.2 (2.5)	13.5 (2.6)	
% Female	40%	12%	15%	a < b, c
% Caucasian	56%	58%	40%	b < c
Estimate Verbal IQ^{1}	104.8 (15.5)	102.0 (12.5)	98.7 (13.8)	a > c
Bisexual/Homosexual	28%	79%	74%	a < b, c
Heterosesxual	72%	20%	26%	
Other/Not Asked	0%	1%	0%	
% AIDS		59%	73%	
Est. Duration HIV+ (years)^{1}		17.5 (9.9)	21.5 (7.6)	b < c
Nadir CD4^{2}		178 [24-214]	149 [12-284]	
Current CD4^{2}		630 [456-840]	521 [420-794]	
% Undetectable HIV RNA (Plasma; on ART)		93%	89%	
% on ART		96%	85%	b > c
% Cognitively Impairment	44%	49%	56%	
Beck Depression Inventory^{1}	5.6 (7.3)	11.2 (10.8)	10.9 (10.7)	a < b, c
% Employed	43%	32%	19%	a > c
% IADL	10%	37%	48%	a < b, c
%Lifetime Substance Use Disorder	55%	74%	85%	a < b, c
% Lifetime Major Depressive Disorder	33%	52%	71%	a < b < c

b.

	Uninfected (a)	HIV Mono-Infected (b)	Coinfected (c)	
n	25	167	35	
Age^{1}	52.8 (17.1)	51.2 (12.7)	52.9 (9.0)	
Education^{1}	15.0 (2.4)	14.5 (2.4)	13.5 (2.5)	a > c
% Caucasian	64%	63%	46%	
Estimate Verbal IQ^{1}	110.7 (17.4)	103.0 (11.4)	101.2 (14.0)	a > b, c
Bisexual	20%	10%	24%	b < a, c
Homosexual	80%	90%	76%	
% AIDS		57%	71%	
Est. Duration HIV+ (years)^{1}		17.8 (10.4)	21.9 (8.2)	b < c
Nadir CD4^{2}		183 [40-343]	175 [14-300]	
Current CD4^{2}		627 [440-820]	508 [357-739]	
% Undetectable HIV RNA (Plasma; on ART)		92%	93%	
% on ART		95%	86%	b > c
% Cognitively Impairment	42%	48%	55%	
Beck Depression Inventory^{1}	6.7 (7.0)	11.3 (11.0)	10.3 (10.4)	
% Employed	38%	31%	23%	
% IADL	12%	34%	48%	a < b,c
%Lifetime Substance Use Disorder	76%	76%	86%	a < c
% Lifetime Major Depressive Disorder	32%	52%	63%	a < b < c

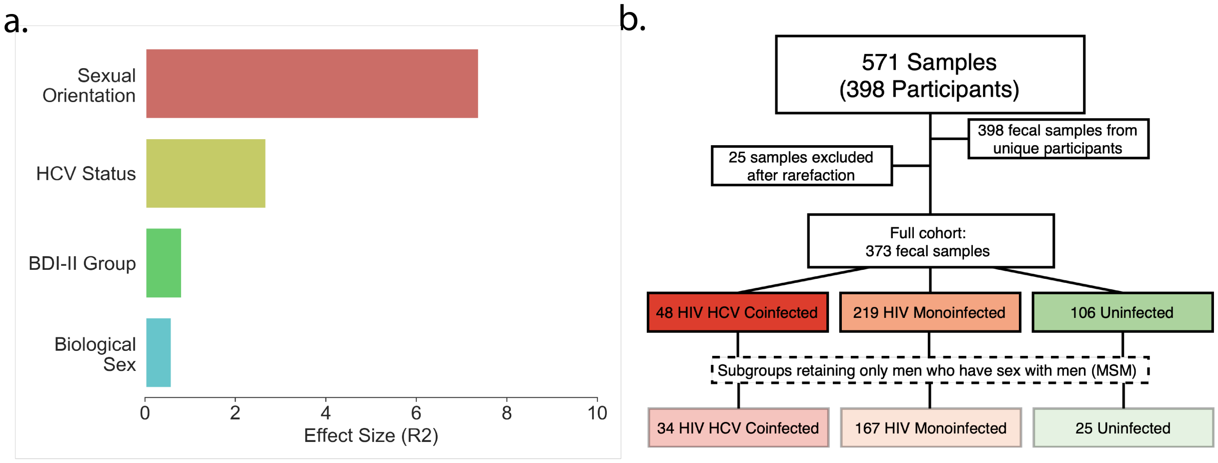


Figure 5.1: Cohort characteristics. A) Unweighted UniFrac relative effect sizes assessed using RDA in the full dataset. B) Sample selection pipeline. Coinfected groups are in red, HIV-monoinfected groups are in orange, and uninfected groups are in green. Lighter colors represent MSM subgroups.

Men who have sex with men (MSM) are known to have Prevotella-rich gut microbiomes, which is also a hallmark in HIV infection [42–49]. To account for this potentially confounding factor, we performed concerted microbiome analyses on 1) the full groups (coinfected, HIV-monoinfected, uninfected) and 2) the subgroups composed only of MSM (coinfected, n = 34; HIV-monoinfected, n = 167; uninfected, n = 25; Fig. 5.1).

When limited to MSMs, the uninfected group had somewhat higher education levels than the coinfected group, and higher premorbid IQ estimates than both infected groups; otherwise, demographic characteristics did not differ between the three subgroups (Table5.1B). The three groups differed in sexual behavior (Chi2 = 7.7, p = 0.02): in the coinfected subgroup, 24% reported sex with men and women and 76% reported sex with only men; for the HIV-monoinfected subgroup, 10% reported sex with men and women and 90% reported sex with only men. In the uninfected subgroup 20% reported sex with men and women and 80% reported sex with only men. As in the full group, the coinfected individuals in the MSM subgroup had longer estimated duration of HIV infection and a smaller percentage were on antiretroviral therapy in comparison to the HIV-monoinfected MSM subgroup. Additional cohort descriptors are included in Table 5.1B. Of the MSM dataset, 20 coinfected, 67 HIV-monoinfected, and 8 uninfected individuals were assessed using untargeted mass spectrometry.

5.4.3 The gut microbiome and metabolome are significantly different between coinfected, HIV-monoinfected, and uninfected individuals.

To understand how the gut microbiome and metabolome of the three infection groups differed from each other, we compared alpha and beta diversity between coinfected, HIV-monoinfected, and uninfected groups. After examining results with the full cohort, we then performed the same analyses on the MSM subgroups.

First, we compared coinfected to uninfected groups. In the full cohort, we observed a statistically significant difference in the overall gut microbial communities in unweighted UniFrac beta diversity distances between coinfected and uninfected individuals (Fig. 5.2A, PERMANOVA pseudo-F = 3.05, BH p = 0.001). Coinfected individuals also had lower alpha diversity than uninfected individuals (5.2B, Shannon index [50], Kruskal-Wallis H (KW-H) = 14.0, BH p = 0.0006). Coinfected and uninfected individuals were also significantly different in their overall gut metabolome (ig. 5.2C, beta diversity Bray-Curtis PERMANOVA pseudo-F = 7.57, BH p = 0.002). However, between coinfected and uninfected MSM subgroups, there were no differences in the overall composition of the gut microbiome and metabolome (unweighted UniFrac beta diversity PERMANOVA pseudo-F = 1.28, BH p = 0.20; Shannon index, KW-H = 2.85, BH p = 0.14; metabolomics beta diversity Bray-Curtis PERMANOVA pseudo-F = 0.98, BH p = 0.47).

Next, we compared HIV-monoinfected to uninfected groups. HIV-monoinfected individuals were also significantly different from uninfected individuals in unweighted UniFrac beta diversity distances (Fig. 5.2A, PERMANOVA pseudo-F = 4.4, BH p = 0.001), and in their overall gut metabolome (Fig. 5.2C, beta diversity Bray-Curtis PERMANOVA pseudo-F = 4.53, BH p = 0.002). Unlike coinfected individuals, however, there was no difference in alpha diversity between the HIV-monoinfected and uninfected groups (Shannon index, KW-H = 0.37, BH p = 0.55). Between the MSM subgroups of HIV-monoinfected and uninfected, there were no differences in the overall composition of the gut microbiome and metabolome (unweighted UniFrac beta diversity PERMANOVA pseudo-F = 1.18, BH p = 0.21; Shannon index, KW-H = 0.0006, BH p = 0.98; metabolomics beta diversity Bray Curtis PERMANOVA pseudo-F = 0.92, BH p = 0.47).

Finally, we compared coinfected to HIV-monoinfected groups. In the full cohorts, we observed a statistically significant difference in unweighted UniFrac beta diversity distances between coinfected and HIV-monoinfected individuals (Fig. 5.2A, PERMANOVA pseudo-F = 2.56, BH p = 0.001). Coinfected individuals also had lower alpha diversity than HIV-monoinfected individuals (Fig. 5.2B, Shannon index, KW-H = 12.5, BH p = 0.0006). Furthermore, coinfected and HIV-monoinfected were significantly different in their overall gut metabolome (Fig. 5.2C, beta diversity Bray-Curtis PERMANOVA pseudo-F = 3.416891, BH p = 0.004). In the MSM subgroups, the unweighted UniFrac beta diversity distances between the coinfected and HIV-monoinfected subgroups remained statistically significantly different (Fig. 5.2D, PERMANOVA pseudo-F = 1.73, BH p = 0.05). Again, the coinfected individuals had a lower alpha diversity than HIV-monoinfected individuals (Fig. 5.2E, Shannon index, KW-H = 6.38, BH p = 0.04). The differences in the overall gut metabolomes of the coinfected and HIV-monoinfected individuals also remained significant in the MSM cohort (Fig. 5.2F, beta diversity, Bray-Curtis PERMANOVA pseudo-F = 3.15, BH p = 0.03).

Alpha diversity does not correlate with immune biomarkers of disease progression in each cohort

Progression of untreated HIV infection is associated with worsening immune suppression, which is characterized by lower CD4+ T-cell counts and higher CD8+ T-cell counts [51], resulting in a low CD4/CD8 ratio. We did not observe any correlation between %CD4+, nadir CD4+, or absolute CD4+ T cells, and alpha diversity (Shannon) in any of the infection groups (SI Table 5.2). Likewise, there was no correlation between CD4/CD8 ratio and alpha diversity (Shannon index) (SI Table 5.2).

Elevated levels of the pro-inflammatory cytokine interleukin(IL)-6, even in the context of viral suppression on antiretroviral therapy (ART), are associated with adverse outcomes such as myocardial infarction and death [52–55]. There were no correlations between plasma IL-6 and alpha diversity (Shannon) in any of the infection groups or subgroups (Table 5.2).

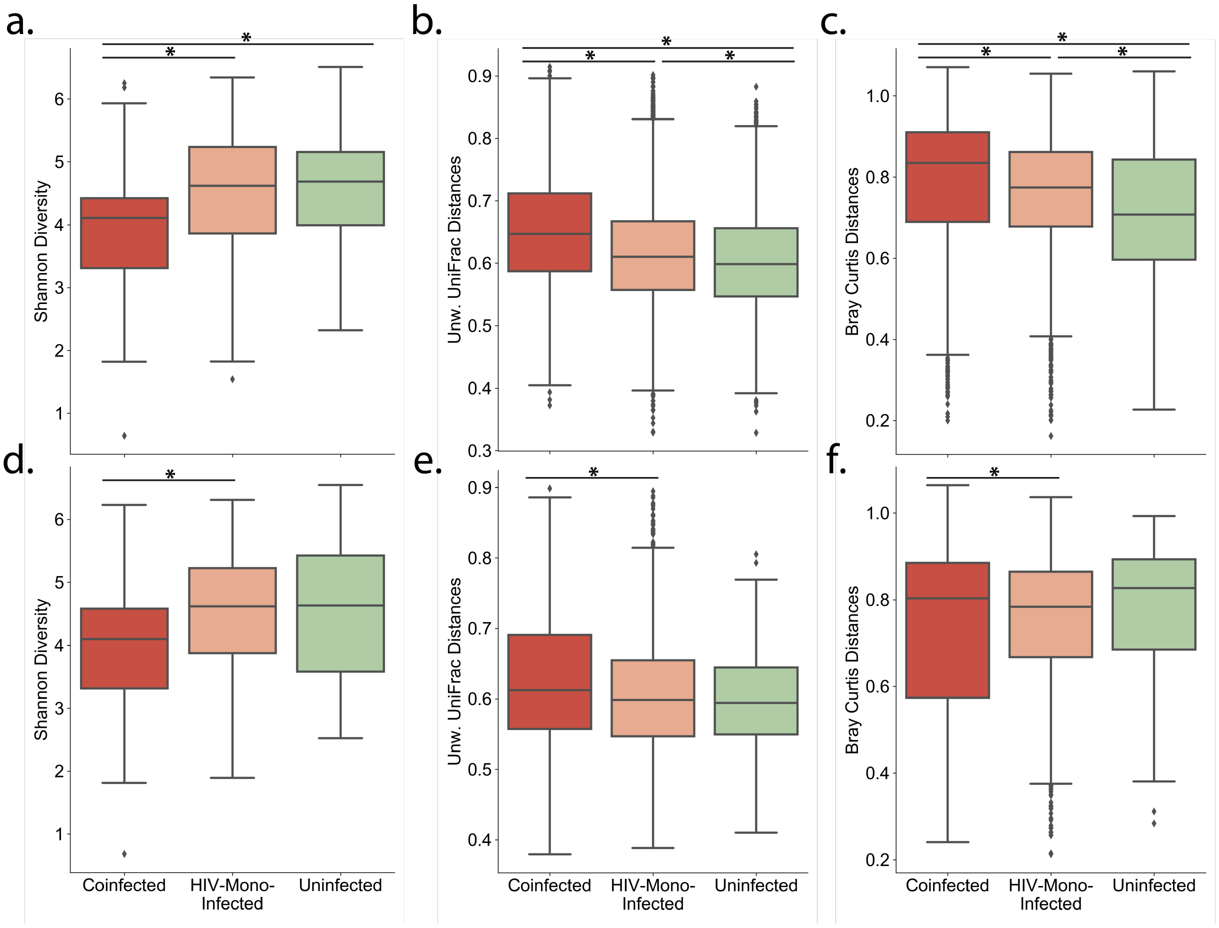


Figure 5.2: Comparison between coinfected (red), HIV-monoinfected (orange), and uninfected (green) groups. A-C) Full groups. D-F) MSM subgroups. A,D) between group alpha (Shannon index) diversity compared to the uninfected group, compared using Kruskal-Wallis test and FDR was controlled using Benjamini-Hochberg procedure; B,E) between group unweighted UniFrac distances of microbiome profiles, compared to the uninfected group, compared using pairwise PERMANOVA; C,F) between group Bray Curtis distances of metabolomic profiles compared to the uninfected group, compared using pairwise PERMANOVA.

5.4.4 Associations of gut microbiome and metabolome composition with current and lifetime depression within the coinfected, HIV-monoinfected, and uninfected cohorts.

We next evaluated each of the three infection groups separately to assess associations between the gut microbiome and depression. Participants underwent standardized assessments of lifetime Major Depressive Disorder using DSM-IV criteria (and current depressive symptoms using the Beck Depression Inventory-II) as described in Materials and Methods. Here, we evaluated the groups according to two assessments: occurrence of lifetime major depressive disorder (MDD) and current depressive symptoms of at least mild severity based on the Beck Depression Inventory (BDI-II 14).

The gut microbiome and metabolome are altered in coinfected individuals with depression. We first tested for association between the gut microbiome and BDI-II in any of the three groups. Individuals were considered currently depressed if they reported at least mild depressive symptoms; otherwise they were considered not depressed. In no infection cohorts was there a significant difference in alpha or beta diversity between individuals stratified by current depressive symptoms (SI Table 5.3). Consistent with prior research [56, 57], there also was no significant correlation between alpha diversity and continuous BDI-II severity in any of the cohorts (SI Table 5.2).

We also were interested in determining whether having MDD at any point (or multiple points) in an individuals life would be associated with gut microbiome differences, separately within the three infection groups. Only in the full coinfected group did we observe a statistically significant difference between those who met lifetime diagnostic criteria for MDD versus those who did not (SI Table 5.3, unweighted UniFrac PERMANOVA, pseudo-F = 1.6, BH p = 0.044). We found no significant differences in the HIV-monoinfected full group or MSM subgroup in unweighted UniFrac distances or Shannon diversity between MDD states (SI Table 5.3). Prior research also suggests that neurobehavioral disorders are not independently associated with gut microbiome dysbiosis in HIV infection [34]. We also found no significant differences between lifetime MDD status in the uninfected groups (SI Table 5.3).

In the metabolomics data, a partial least squares discriminant analysis (PLSDA) and random forest analysis were used to identify features of interest between lifetime MDD status within each group. In the metabolomics GNPS network, multiple compounds of interest were found to be annotated as bile acids.

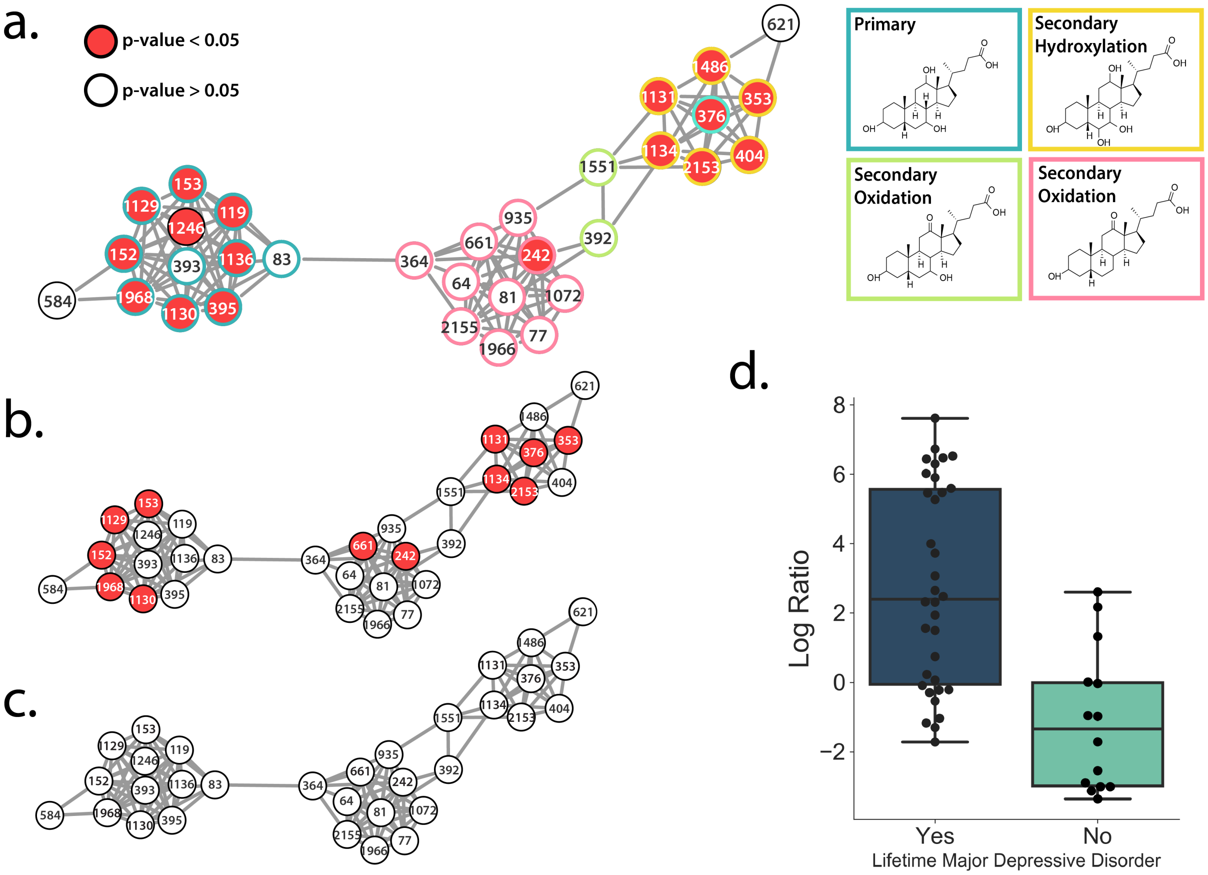


Figure 5.3: The gut microbiome and metabolome differ in coinfected individuals with a lifetime history of MDD. A)-C) Bile Acids networks. Red indicates that the bile acid was significantly higher ($p < 0.05$) in individuals who reported having a lifetime history of MDD versus those who never had MDD. The list of GNPS annotations for this network are available in SI Table 5.4. A) Full coinfected cohort; Primary and Secondary Bile Acids annotation of the network. B) Coinfected MSM subgroup. C) All other cohorts. D) Individuals who had lifetime MDD have a significantly higher log ratio of set 1 to set 2 (t-test, P = 9.1e-06, t = -5.21, df = 34.18, Cohens D = 1.43). The list of microbes in each set are available in SI Table 5.5.

Further analysis of all annotated bile acids revealed that in both the full coinfected group and the coinfected MSM subgroup, a cluster of primary and secondary bile acids were significantly increased (Dunns test, $p < 0.05$) in individuals with a lifetime history of MDD (5.3A,B; annotations in SI Table 5.4). This difference is not observed in the HIV-monoinfected or the uninfected groups (5.3C).

Bile acids and the gut microbiome exist in a dynamic equilibrium [58]. Primary bile acids are produced and conjugated in the liver, released in the biliary tract, and maintained through positive feedback antagonism of farnesoid X receptor (FXR) in the intestines and the liver [59]. Bile acids mediate anti-inflammatory immune responses by binding to receptors such as Takeda G protein-coupled receptor 5 [60, 61]. Primary bile acids are metabolized by gut microbes into secondary bile acids and passively absorbed into the portal circulation [62]. Secondary bile acids affect host physiology by binding and activating host nuclear receptors to a greater extent than primary bile acids [58]. Here, seven annotations of secondary bile acids were significantly increased in the full cohort of coinfected individuals with a lifetime history MDD, and six were significantly increased in the MSM subgroup (5.3A,B). This finding suggests increased metabolism of primary to secondary bile acids by gut microbes in individuals coinfected with both HIV and HCV.

Bile acid imbalances are known to result in pathological states such as liver disease, gastrointestinal cancers, and gallstones [58]. Shifts in bile acid homeostasis are associated with HCV infection [63] and chronic liver disease. Bile acid abundance and composition are also dysregulated in MDD [64]. Our observation of increased primary and secondary bile acids in coinfected individuals with a lifetime history of MDD compared to coinfected individuals without a lifetime history of MDD suggests that dysregulated bile acid metabolism by gut bacteria may be a mechanism that links HIV-HCV coinfection and MDD.

Due to our observation that overall microbiome composition, as measured by unweighted UniFrac, differed between coinfected individuals with or without lifetime MDD, we were also interested in determining whether specific groups of taxa may be driving the bile acid differences we observed in the gut metabolome. We used Songbird [65] to identify microbes that were associated with lifetime MDD in the full cohort of coinfected individuals. Songbird is a compositionally aware differential abundance method which provides rankings of features (suboperational taxonomic units [sOTUs]) based on their log fold-change with respect to covariates of interest. In this case, the formula we used described whether the individual had lifetime MDD or not. We selected the highest 10% (set 1, SI Table 5.5) and lowest 10% (set 2, SI Table 5.5) of the ranked sOTUs associated with lifetime MDD and used Qurro [66] to compute the log ratio of these sets of taxa (5.3D). Comparing the ratios of taxa in this way mitigates bias from the unknown total microbial load in each sample, and taking the log of this ratio gives equal weight to relative increases and decreases of taxa [65]. Evaluation of the Songbird model against a baseline model obtained a pseudo-Q2 value > 0, suggesting the model was not overfit. We found that coinfected individuals who had lifetime MDD had a significantly higher log ratio of set 1 to set 2 sOTUs than those who never had a MDD (t-test, p = 9.055e-06, t = -5.210, df = 34.183, Cohens D = 1.434), suggesting that they were associated with set 1. Several microbes that we found to be associated with coinfected individuals with a lifetime history of MDD (set 1 microbes, SI Table 5.5) have also been previously identified as enriched in MDD, including Enterobacteriaceae [31] and *Alistepes* species (here, *Alistepes onderdonkii*) [31, 67–69], Bacteroides [31, 68], and Parabacteroides (here, *Parabacteroides distasonis*) [31]. Likewise, coinfected individuals without lifetime MDD were enriched in several microbes (set 2, SI Table 5.5) that were previously identified as being decreased in uninfected individuals with MDD, including: *Dialister spp.* [31, 70], Lachnospiraceae [68], and *Ruminococcus spp.* [31].

5.5 Conclusions

This is to our knowledge the first study of the association between infection with HIV and HCV, depression, and the gut microbiome and metabolome. We performed 16S rRNA sequencing and liquid chromatography mass spectrometry using stool samples from nearly 400 individuals and evaluated the data with state-of-the-art tools. We observed that although the gut microbiome of HIV positive and negative individuals differed, HCV had a large effect on the microbiome which warranted consideration in our study. The infection groups differed from each other in terms of both alpha and beta diversity, in the full cohort as well as the MSM subgroups. Furthermore, we found that depression was associated with differences in the gut microbiome and metabolome only in HIV-HCV coinfected individuals. Coinfected individuals with a lifetime history of MDD were enriched in primary and secondary bile acids, as well as particular depression-related taxa. Since microbial diversity and function may be altered by interventions such as probiotic and prebiotic administration, our findings support further investigation of microbiome interventions for treatment of MDD, particularly in HIV-HCV coinfected individuals. Importantly, our results suggest that microbiome and metabolome investigations in HIV-infected cohorts should carefully consider possible effects of HCV coinfections, which are not uncommon among people living with HIV. While this study provides the foundation for more directed research, it has some limitations – particularly the lack of an HCV-monoinfected group, the small number of women, and the reduced sample sizes after subsetting for MSM. In future studies, it would also be of great interest to consider current MDD and other neurobehavioral or neuropsychiatric metrics in coinfected and monoinfected cohorts.

5.6 Acknowledgments

Chapter 5, in full, will be submitted for publication and may appear as “B. C. Taylor, K. C. Weldon, T. Groth, R. Ellis, E. Gentry, A. Tripathi, D. McDonald, G. Humphrey, M. Bryant, J. Toronczak, T. Schwartz, M. F. Oliveira, R. Heaton, S. Gianella, A. D. Swafford, P. C. Dorrestein, and R. Knight. *Depression in HIV and HCV Co-Infected Individuals is Associated with Systematic Differences in the Gut Microbiome and Metabolome*.” The dissertation author was the primary investigator and author of this work, contributing substantially to experimental design. Sample collection, 16S rRNA extraction, and sequencing/mass spectrometric analysis were performed by others. The dissertation author computationally processed the 16S rRNA sequence data, designed the analysis of the 16S rRNA sequence data and mass spectrometric data, interpreted the results of the analyses, designed and assembled the figures, performed the statistical analyses, and wrote the manuscript.

This work was supported by a Seed Grant from the Center for Microbiome Innovation. We would like to thank Lindsay DeRight-Goldasich, A. Cole Heale, and Karenina Sanders for sample processing, Gail Ackermann for assistance with metadata curation, and Jeff DeReus for data handling and processing.

The HIV Neurobehavioral Research Center (HNRC) is supported by Center award P30MH062512 from NIMH.

The San Diego HIV Neurobehavioral Research Center [HNRC] group is affiliated with the University of California, San Diego, the Naval Hospital, San Diego, and the Veterans Affairs San Diego

Healthcare System, and includes: Director: Robert K. Heaton, Ph.D., Co-Director: Igor Grant, M.D.; Associate Directors: J. Hampton Atkinson, M.D., Ronald J. Ellis, M.D., Ph.D., and Scott Letendre, M.D.; Center Manager: Jennifer Iudicello, Ph.D.; Donald Franklin, Jr.; Melanie Sherman; NeuroAssessment Core: Ronald J. Ellis, M.D., Ph.D. (P.I.), Scott Letendre, M.D., Thomas D. Marcotte, Ph.D, Christine Fennema-Notestine, Ph.D., Debra Rosario, M.P.H., Matthew Dawson; NeuroBiology Core: Cristian Achim, M.D., Ph.D. (P.I.), Ana Sanchez, Ph.D., Adam Fields, Ph.D.; NeuroGerm Core: Sara Gianella Weibel, M.D. (P.I.), David M. Smith, M.D., Rob Knight, Ph.D., Scott Peterson, Ph.D.; Developmental Core: Scott Letendre, M.D. (P.I.), J. Allen McCutchan; Participant Accrual and Retention Unit: J. Hampton Atkinson, M.D. (P.I.) Susan Little, M.D., Jennifer Marquie-Beck, M.P.H.; Data Management and Information Systems Unit: Lucila Ohno-Machado, Ph.D. (P.I.), Clint Cushman; Statistics Unit: Ian Abramson, Ph.D. (P.I.), Florin Vaida, Ph.D. (Co-PI), Anya Umlauf, M.S., Bin Tang, M.S.

The views expressed in this article are those of the authors and do not reflect the official policy or position of the Department of the Navy, Department of Defense, nor the United States Government.

M.F.O acknowledges support from R01AG061066 and (CNPq)-Brazil [245954/2012].

5.7 Materials and Methods

5.7.1 Participant recruitment, sample processing, and sample selection.

This was a cross-sectional prospective observational cohort study of persons with or without HIV infection recruited from community sources, who agreed to undergo comprehensive neuromedical and neurobehavioral evaluations for NIH-funded studies at the HIV Neurobehavioral Research Program (HNRP, https://hnrp.hivresearch.ucsd.edu/) including the HIV Neurobehavioral Research Center (HNRC) study (study details: HNRP [71, 72] at the University of California at San Diego. Those who also agreed to submit stool samples for microbiome studies were included in the current analyses. A subset of participants also had positive serology for hepatitis C virus. The UCSD's Human Research Protections Program (irb.ucsd.edu) approved all study procedures, and all participants provided written informed consent.

Exclusions were diagnoses of active substance use disorders and presence of an active, major psychiatric condition with current psychotic features or neurological conditions such as schizophrenia or epilepsy. If multiple stool samples were collected from participants, only the first time point was analyzed by 16S rRNA sequencing. A single time point per subject was additionally analyzed by HPLC-MS/MS. HIV and HCV infection was confirmed by a point-of-care vertical flow test (MedMira, Halifax, Nova Scotia). Participants were designated as a) HIV-monoinfected if they tested positive for HIV but not HCV, b) Coinfected if they tested positive for both HIV and HCV, or c) Uninfected if they tested positive for neither HIV or HCV.

5.7.2 Neuromedical and Laboratory Assessment.

All participants underwent a comprehensive neuromedical assessment, including a medical history that collected antiretroviral therapy (ART) and other medications, data to determine Centers for Disease Control (CDC) staging, and specimen collection (blood, stool). Routine clinical chemistry panels, complete blood counts, rapid plasma reagin, and CD4+ T cells (flow cytometry) were performed at a Clinical Laboratory Improvement Amendments (CLIA)certified medical center laboratory. HIV RNA were measured in plasma using reverse transcriptase polymerase chain reaction (Amplicor, Roche Diagnostics, Indianapolis, IN, with a lower limit of quantitation 40 copies/ml).

5.7.3 Evaluation of Depression

DSM-IV diagnosis of Lifetime Major Depressive Disorder was evaluated using the computer-assisted Composite International Diagnostic Interview (CIDI [73]), a structured instrument widely used in psychiatric research. Current self-reported depressed mood was assessed using the Beck Depression Inventory-II (BDI-II [74]). The BDI-II consists of 21 items that assess the severity of depression symptoms over the 2 weeks prior to assessment. The BDI-II total score ranges from 0-63 with higher scores denoting more severe depression symptoms. For analyses, we used the published cutoff of at least mild severity to define current self-reported depression [74].

5.7.4 16S rRNA Gene Sequencing.

DNA extraction and 16S rRNA amplicon sequencing were done using Earth Microbiome Project (EMP) standard protocols (http://www.earthmicrobiome.org/protocols-and-standards/16s). DNA was extracted with the Qiagen MagAttract PowerSoil DNA kit as previously described [75]. Amplicon PCR was performed on the V4 region of the 16S rRNA gene using the primer pair 515f to 806r with Golay error-correcting barcodes on the reverse primer. Amplicons were barcoded and pooled in equal concentrations for sequencing. The amplicon pool was purified with the MO BIO UltraClean PCR cleanup kit and sequenced on the Illumina MiSeq sequencing platform. Sequence data were demultiplexed and minimally quality filtered using the Qiita defaults.

5.7.5 16S marker gene data analysis.

QIIME 2 [76] was used to generate pairwise unweighted and weighted UniFrac distances [35, 37, 77]. Between group differences based on these distances were tested using PERMANOVA [78] and permuted t-tests in QIIME 2. Alpha diversity (Shannon diversity [50]) was compared with a Kruskal-Wallis test.

Songbird v1.0.1 [65] in QIIME 2 version 2020.2 was used to identify feature ranks (parameters: –p-epochs 10000 –batch-size 5 –learning-rate 1e-4 –min-sample-count 1000 –min-feature-count 0 –num-random-test-examples 10) and Qurro v0.4.0[66] was used to compute log-ratios of these ranked features. T-tests and Cohens D were calculated to assess the significance (alpha=0.05) and effect size of the log-ratios.

5.7.6 LC-MS/MS data acquisition.

Human fecal samples were transferred to clean 2 mL sample tubes (Qiagen Cat No.990381) and the weights were recorded. The samples were then extracted in a solution of 1:1 methanol to water spiked with an internal standard of 1 μM sulfamethazine, using a 1:10 sample weight (milligram) to solvent volume (microliter) ratio. Using a Tissuelyser II (Qiagen), the samples were homogenized for 5 minutes at 25 hertz. This was followed by a 15 minute centrifugation at 14,000 rpm. From the supernatant, 400 μL was transferred to a pre-labeled 96-Well DeepWell plate and the plates were concentrated using a CentriVap Benchtop Vacuum Concentrator (Labconco) for approximately 4 hours. The dry plates were placed into the -80 degrees Celcius until time for analysis.

The plates were resuspended in 150 μL of a 1:1 methanol to water with a 1 μM sulfadimethoxine internal standard solution. For metabolomics analysis, an ultra-high performance liquid chromatography system (Thermo Dionex Ultimate 3000 UHPLC) coupled to an ultra-high resolution quadrapole time of flight (qToF) mass spectrometer (Bruker Daltonics MaXis HD). For chromatographic separation, a Phenomenex Kinetex column (C18 1.7 μM, 2.1 mm x 50 mm) was used. The mobile phase consisted of solvent A: 100 percent LC-MS grade water with 0.1 percent formic acid and solvent B: 100 percent acetonitrile with 0.1 percent formic acid. Each sample was injected at a volume of 5 μL into a flow rate of 0.5 mL for the entire analysis. The 12 minute chromatographic gradient began at 5 percent solvent B for the first minute, an increase to 100 percent solvent B from minute 1 to minute 11, a hold at 100 percent B until minute 11.5, and back down to 5 percent B reached at minute 11.5. All data was collected using electrospray ionization in positive mode.

5.7.7 LC-MS/MS data analysis.

The raw data in Bruker (.d) format were lock mass corrected using hexakis ((1H, 1H, 2H-difluoroethoxy)phosphazene (Synquest Laboratories, Alachua, FL) and were exported as .mzXML files using the Bruker Data Analysis software. Both the raw .d and the .mzXML files were uploaded to the UC San Diego mass spectrometry data repository MassIVE (https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp). Feature detection was completed using MZmine version 2.37 software [79]. Parameters can be found in supplemental 5.6. The resulting feature tables were exported as both a quantification file (.csv) and a spectral information file (.mgf) for analysis using the Global Natural Products Social Molecular Networking (GNPS) platform [80].

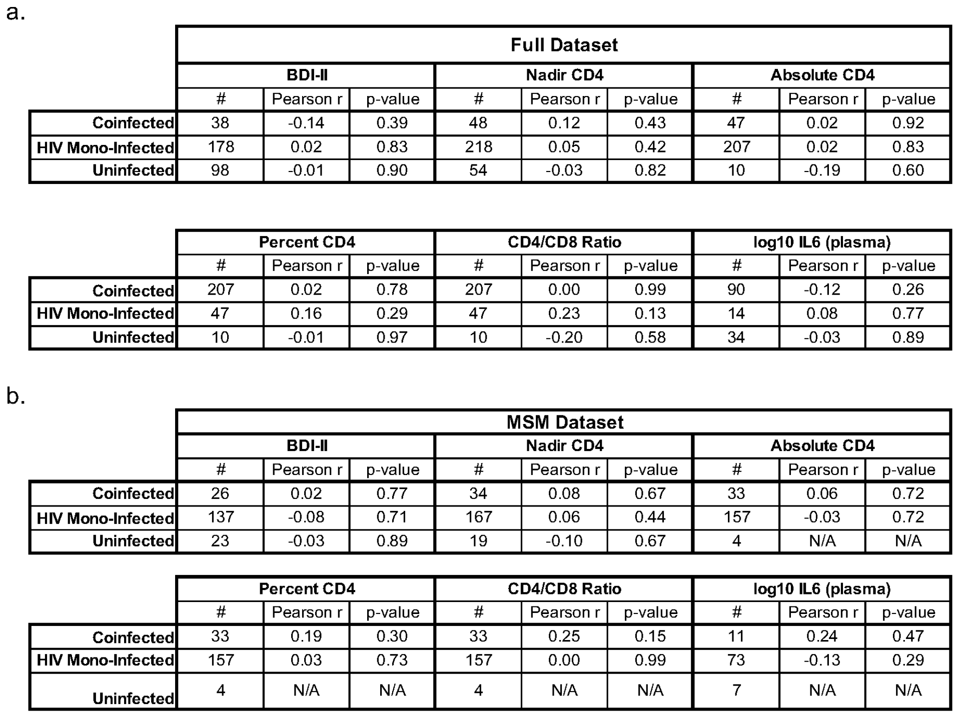
The quantification table and the spectral information was analyzed using the GNPS feature based molecular networking workflow [81]. Parameters can be viewed via the job results page (https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=350392e8e24c41f2b84fde04f9183fc4). The results reflect MSI level 2 or 3 annotations [82]. For the statistical analyses, the MZmine-produced feature abundance table containing peak areas was inputted into the web-based MetaboAnalyst Software [83]. The data was normalized following the metabolomics data analysis protocols outlined in the previous metabolomics project [84], a normalization by quantile normalization and an auto scale. The Metaboanlyst software was used to produce the partial least square discriminant analysis (PLS-DA) and random forest accuracy results. The normalized data was used to calculate a squareform matrix based on the bray curtis distance metric which was inputted into a .qza format for use in QIIME2. All PERMANOVAs were run using the QIIME2 beta group significance command [76]. The Cytoscape v3.7.2 software was used for all molecular networking visualizations [85]. Individual feature level comparisons were completed using a Dunns test.

5.7.8 Data availability.

The data generated in this study are available publicly in Qiita under the study ID 11135. Sequence data associated with this study can be found under EBI accession TBD. The GNPS feature based molecular networking job is available at: https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=350392e8e24c41f2b84fde04f9183fc4. The raw experimental data are available at MassIVE (https://massive.ucsd.edu/), dataset MSV000083664.

5.8 Supplementary Material

Table 5.2: Alpha (Shannon index) diversity correlations with continuous BDI-II and continuous biomarkers in the infection groups and MSM subgroups.



a.

	Full Dataset								
	BDI-II			Nadir CD4			Absolute CD4		
	#	Pearson r	p-value	#	Pearson r	p-value	#	Pearson r	p-value
Coinfected	38	-0.14	0.39	48	0.12	0.43	47	0.02	0.92
HIV Mono-Infected	178	0.02	0.83	218	0.05	0.42	207	0.02	0.83
Uninfected	98	-0.01	0.90	54	-0.03	0.82	10	-0.19	0.60

	Percent CD4			CD4/CD8 Ratio			log10 IL6 (plasma)		
	#	Pearson r	p-value	#	Pearson r	p-value	#	Pearson r	p-value
Coinfected	207	0.02	0.78	207	0.00	0.99	90	-0.12	0.26
HIV Mono-Infected	47	0.16	0.29	47	0.23	0.13	14	0.08	0.77
Uninfected	10	-0.01	0.97	10	-0.20	0.58	34	-0.03	0.89

b.

	MSM Dataset								
	BDI-II			Nadir CD4			Absolute CD4		
	#	Pearson r	p-value	#	Pearson r	p-value	#	Pearson r	p-value
Coinfected	26	0.02	0.77	34	0.08	0.67	33	0.06	0.72
HIV Mono-Infected	137	-0.08	0.71	167	0.06	0.44	157	-0.03	0.72
Uninfected	23	-0.03	0.89	19	-0.10	0.67	4	N/A	N/A

	Percent CD4			CD4/CD8 Ratio			log10 IL6 (plasma)		
	#	Pearson r	p-value	#	Pearson r	p-value	#	Pearson r	p-value
Coinfected	33	0.19	0.30	33	0.25	0.15	11	0.24	0.47
HIV Mono-Infected	157	0.03	0.73	157	0.00	0.99	73	-0.13	0.29
Uninfected	4	N/A	N/A	4	N/A	N/A	7	N/A	N/A

Table 5.3: Alpha and beta diversity between categorical BDI-II and LT MDD groups in all infection groups. Beta (unweighted UniFrac) and alpha (Shannon) diversity in the 16S data, and Beta (Bray Curtis) diversity in the metabolomics data.

a.

	Full Dataset							
	BDI-II				LT MDD			
	Minimal	> Minimal	pseudo-F	p-value	LT MDD	No MDD	pseudo-F	p-value
Coinfected	27	11	1.07	0.30	34	14	1.60	0.044
HIV Mono-Infected	122	55	0.68	0.96	113	104	1.28	0.12
Uninfected	84	14	1.37	0.078	34	69	0.81	0.81

b.

	MSM Filtered Dataset							
	BDI-II				LT MDD			
	Minimal	> Minimal	pseudo-F	p-value	LT MDD	No MDD	pseudo-F	p-value
Coinfected	19	7	1.12	0.27	21	13	0.98	0.45
HIV Mono-Infected	94	42	0.92	0.56	86	79	1.43	0.06
Uninfected	20	3	1.00	0.36	8	17	0.82	0.77

c.

	Full Dataset							
	BDI-II				LT MDD			
	Minimal	> Minimal	KW-H	BH p-value	LT MDD	No MDD	KW-H	BH p-value
Coinfected	27	11	2.05	48	34	14	2.59	0.11
HIV Mono-Infected	122	55	0.005	0.94	113	104	0.0002	0.99
Uninfected	84	14	3.19	0.22	34	69	0.42	0.52

d.

	MSM Filtered Dataset							
	BDI-II				LT MDD			
	Minimal	> Minimal	KW-H	BH p-value	LT MDD	No MDD	KW-H	BH p-value
Coinfected	19	7	0.80	0.37	21	13	1.25	0.26
HIV Mono-Infected	94	42	0.10	0.76	86	79	0.03	0.87
Uninfected	20	3	0.30	0.58	8	17	0.12	0.73

e.

	Full Dataset							
	BDI-II				LT MDD			
	Minimal	> Minimal	pseudo-F	p-value	LT MDD	No MDD	pseudo-F	p-value
Coinfected	15	8	2.09	0.17	20	7	1.14	0.28
HIV Mono-Infected	48	21	0.78	0.66	46	36	0.74	0.70
Uninfected	26	4	1.80	0.087	10	22	1.03	0.37

f.

	MSM Filtered Dataset							
	BDI-II				LT MDD			
	Minimal	> Minimal	pseudo-F	p-value	LT MDD	No MDD	pseudo-F	p-value
Coinfected	11	5	1.83	0.17	14	6	0.93	0.39
HIV Mono-Infected	39	19	0.82	0.63	38	30	0.87	0.51
Uninfected	6	2	0.45	0.84	2	6	0.45	0.81

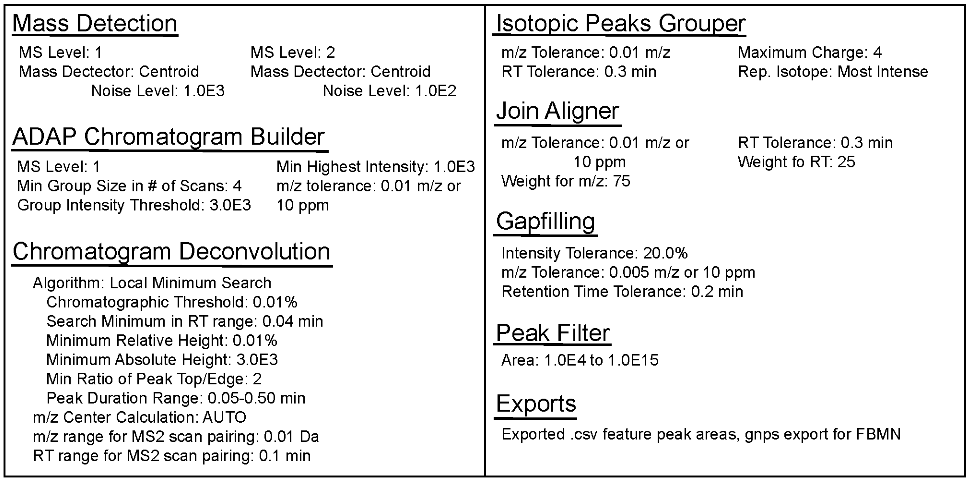
Table 5.4: GNPS annotations (MSI level 3) for the bile acid networks shown in all figures. The annotation indicates the closest spectral match found in the GNPS libraries for the feature.

Cluster Index	m/z	Retention Time (min)	GNPS Annotation
64	391.2836	5.579	12-Ketodeoxycholic Acid
77	391.2835	5.516	12-Ketodeoxycholic Acid
81	373.2728	5.581	(4R)-4-((3S,5R,9S,10S,13R,14S,17R)-3-hydroxy-10,13-dimethyl-12-oxohexadecahydro-1H-cyclopenta[a]phenanthren-17-yl)pentanoic acid
83	373.2731	5.914	(R)-4-((3S,5S,7S,8R,9S,10S,12S,13R,14S,17R)-3,7,12-trihydroxy-10,13-dimethylhexadecahydro-1H-cyclopenta[a]phenanthren-17-yl)pentanoic acid
119	373.2729	4.632	(R)-4-((3S,5S,7R,8R,9S,10S,12S,13R,14S,17R)-3,7,12-trihydroxy-10,13-dimethylhexadecahydro-1H-cyclopenta[a]phenanthren-17-yl)pentanoic acid
152	373.2730	5.154	(R)-4-((3R,5S,7R,8R,9S,10S,12S,13R,14S,17R)-3,7,12-trihydroxy-10,13-dimethylhexadecahydro-1H-cyclopenta[a]phenanthren-17-yl)pentanoic acid
153	373.2729	4.715	(R)-4-((3S,5S,7R,8R,9S,10S,12S,13R,14S,17R)-3,7,12-trihydroxy-10,13-dimethylhexadecahydro-1H-cyclopenta[a]phenanthren-17-yl)pentanoic acid
242	391.2833	5.246	(4R)-4-((3S,5R,9S,10S,13R,14S,17R)-3-hydroxy-10,13-dimethyl-12-oxohexadecahydro-1H-cyclopenta[a]phenanthren-17-yl)pentanoic acid
353	389.2684	4.844	(4R)-4-((3R,5R,6S,7R,9S,10R,12S,13R,14S,17R)-3,6,7,12-tetrahydroxy-10,13-dimethylhexadecahydro-1H-cyclopenta[a]phenanthren-17-yl)pentanoic acid
364	373.2732	5.545	(4R)-4-((3S,5R,9S,10S,13R,14S,17R)-3-hydroxy-10,13-dimethyl-12-oxohexadecahydro-1H-cyclopenta[a]phenanthren-17-yl)pentanoic acid
376	371.2577	4.855	(R)-4-((3R,5S,8R,9S,10S,13R,14S,17R)-3-hydroxy-10,13-dimethyl-7-oxohexadecahydro-1H-cyclopenta[a]phenanthren-17-yl)pent-2-enoic acid
392	389.2688	5.642	(4R)-4-((3S,5S,7S,9S,10S,13R,14S,17R)-3,7-dihydroxy-10,13-dimethyl-12-oxohexadecahydro-1H-cyclopenta[a]phenanthren-17-yl)pentanoic acid
393	373.2738	4.303	(R)-4-((3S,5S,7S,8R,9S,10S,12S,13R,14S,17R)-3,7,12-trihydroxy-10,13-dimethylhexadecahydro-1H-cyclopenta[a]phenanthren-17-yl)pentanoic acid
395	391.2837	4.606	(R)-4-((3R,5S,7S,8R,9S,10S,12S,13R,14S,17R)-3,7,12-trihydroxy-10,13-dimethylhexadecahydro-1H-cyclopenta[a]phenanthren-17-yl)pentanoic acid
404	389.2684	4.595	(4R)-4-((3R,5R,6S,7R,9S,10R,12S,13R,17R)-3,6,7,12-tetrahydroxy-10,13-dimethylhexadecahydro-1H-cyclopenta[a]phenanthren-17-yl)pentanoic acid
584	415.3197	5.995	(2S,6R)-2-met-yl-6-((3R,5S,7R,9S,10S,12S,13R,14S,17R)-3,7,12-trihydroxy-10,13-dimethylhexadecahydro-1H-cyclopenta[a]phenanthren-17-yl)heptanoic acid
621	431.3147	5.664	Not Annotated
661	373.2728	5.438	(4R)-4-((3S,5R,9S,10S,13R,14S,17R)-3-hydroxy-10,13-dimethyl-12-oxohexadecahydro-1H-cyclopenta[a]phenanthren-17-yl)pentanoic acid
935	355.2622	5.441	(4R)-4-((3S,5R,9S,10S,13R,14S,17R)-3-hydroxy-10,13-dimethyl-12-oxohexadecahydro-1H-cyclopenta[a]phenanthren-17-yl)pentanoic acid
1072	373.2737	7.102	(4R)-4-((3S,5R,9S,10S,13R,14S,17R)-3-hydroxy-10,13-dimethyl-12-oxohexadecahydro-1H-cyclopenta[a]phenanthren-17-yl)pentanoic acid
1129	355.2627	4.59	Cholic Acid
1130	373.2734	4.588	(4R)-4-((3R,5S,7S,9S,10S,12S,13R,17R)-3,7,12-trihydroxy-10,13-dimethylhexadecahydro-1H-cyclopenta[a]phenanthren-17-yl)pentanoic acid
1131	407.2793	4.809	(R)-4-((1R,3S,5S,7R,8S,9S,10S,12S,13R,14S,17R)-1,3,7,12-tetrahydroxy-10,13-dimethylhexadecahydro-1H-cyclopenta[a]phenanthren-17-yl)pentanoic acid
1134	407.2791	4.444	(4R)-4-((1R,3S,5S,7R,9S,10S,12S,13R,14S,17R)-1,3,7,12-tetrahydroxy-10,13-dimethylhexadecahydro-1H-cyclopenta[a]phenanthren-17-yl)pentanoic acid
1136	355.2632	4.412	Cholic Acid
1246	375.2812	5.17	Not Annotated
1486	389.2690	4.523	(R)-4-((3R,5R,6R,7R,8R,9S,10R,12S,13R,14S,17R)-3,6,7,12-tetrahydroxy-10,13-dimethylhexadecahydro-1H-cyclopenta[a]phenanthren-17-yl)pentanoic acid
1551	371.2578	5.644	(R)-4-((3S,5S,7R,8R,9S,10S,13R,14S,17R)-3,7-dihydroxy-10,13-dimethyl-12-oxohexadecahydro-1H-cyclopenta[a]phenanthren-17-yl)pentanoic acid
1966	391.2839	5.561	12-Ketodeoxycholic Acid
1968	373.2734	5.145	(R)-4-((3R,5S,7R,8R,9S,10S,12S,13R,14S,17R)-3,7,12-trihydroxy-10,13-dimethylhexadecahydro-1H-cyclopenta[a]phenanthren-17-yl)pentanoic acid
2153	407.2806	4.731	(4R)-4-((1R,3S,5S,7R,9S,10S,12S,13R,14S,17R)-1,3,7,12-tetrahydroxy-10,13-dimethylhexadecahydro-1H-cyclopenta[a]phenanthren-17-yl)pentanoic acid
2155	373.2794	4.893	(4R)-4-((3S,5R,9S,10S,13R,14S,17R)-3-hydroxy-10,13-dimethyl-12-oxohexadecahydro-1H-cyclopenta[a]phenanthren-17-yl)pentanoic acid

Table 5.5: Sets of taxa identified using Songbird and used in log ratio calculations.

Sets	**Taxa**
Set 1	Enterobacteriaceae, *Alistipes onderdonkii, Anaerotruncus spp., Blautia spp., Bacteroides spp., Bacteroides caccae, Blautia producta, Clostridium aldenense, Clostridium ramosum, Clostridium spiroforme, Clostridium symbiosum, Dialister spp., Oribacterium spp., Oscillospira spp., Parabacteroides distasonis, Prevotella spp., [Ruminococcus] gnavus, [Ruminococcus] torques, Sutterella spp., and Veillonella parvula*.
Set 2	*Faecalibacterium prausnitzii, Akkermansia muciniphila*, families Rikenellaceae, Ruminococcaceae, Coriobacteriaceae, [Mogibacteriaceae], Lachnospiraceae, *Haemophilus parainfluenzae, Blautia spp.*, order Clostridiales, *Streptococcus spp., Ruminococcus spp., Clostridium celatum, Lachnospira spp., Roseburia spp.*

Table 5.6: MZmine version 2.37 parameters used for feature detection on the metabolomics data.



Mass Detection

MS Level: 1	MS Level: 2
Mass Dectector: Centroid	Mass Dectector: Centroid
Noise Level: 1.0E3	Noise Level: 1.0E2

ADAP Chromatogram Builder

MS Level: 1	Min Highest Intensity: 1.0E3
Min Group Size in # of Scans: 4	m/z tolerance: 0.01 m/z or
Group Intensity Threshold: 3.0E3	10 ppm

Chromatogram Deconvolution

Algorithm: Local Minimum Search
Chromatographic Threshold: 0.01%
Search Minimum in RT range: 0.04 min
Minimum Relative Height: 0.01%
Minimum Absolute Height: 3.0E3
Min Ratio of Peak Top/Edge: 2
Peak Duration Range: 0.05-0.50 min
m/z Center Calculation: AUTO
m/z range for MS2 scan pairing: 0.01 Da
RT range for MS2 scan pairing: 0.1 min

Isotopic Peaks Grouper

m/z Tolerance: 0.01 m/z	Maximum Charge: 4
RT Tolerance: 0.3 min	Rep. Isotope: Most Intense

Join Aligner

m/z Tolerance: 0.01 m/z or 10 ppm	RT Tolerance: 0.3 min
Weight for m/z: 75	Weight fo RT: 25

Gapfilling

Intensity Tolerance: 20.0%
m/z Tolerance: 0.005 m/z or 10 ppm
Retention Time Tolerance: 0.2 min

Peak Filter

Area: 1.0E4 to 1.0E15

Exports

Exported .csv feature peak areas, gnps export for FBMN

References

(1) Maingat, F.; Halloran, B.; Acharjee, S.; Marle, G.; Church, D.; Gill, M. J.; Uwiera, R. R. E.; Cohen, E. A.; Meddings, J.; Madsen, K.; Power, C. *The FASEB Journal* **2011**, DOI: 10.1096/fj.10-175992.

(2) Douek, D. *Topics in HIV medicine : a publication of the International AIDS Society, USA* **2007**.

(3) Asmuth, D. M.; Thompson, C. G.; Chun, T. W.; Ma, Z. M.; Mann, S.; Sainz, T.; Serrano-Villar, S.; Utay, N. S.; Garcia, J. C.; Troia-Cancio, P.; Pollard, R. B.; Miller, C. J.; Landay, A.; Kashuba, A. D. *Journal of Infectious Diseases* **2017**, DOI: 10.1093/infdis/jix418.

(4) Yoseph, B. P.; Klingensmith, N. J.; Liang, Z.; Breed, E. R.; Burd, E. M.; Mittal, R.; Dominguez, J. A.; Petrie, B.; Ford, M. L.; Coopersmith, C. M. *Shock* **2016**, DOI: 10.1097/SHK.0000000000000565.

(5) De Medeiros, R. M.; Valverde-Villegas, J. M.; Junqueira, D. M.; Gräf, T.; Lindenau, J. D.; De Mello, M. G.; Vianna, P.; Almeida, S. E.; Chies, J. A. B. *PLoS ONE* **2016**, DOI: 10.1371/journal.pone.0156163.

(6) Zheng, P.; Zeng, B.; Zhou, C.; Liu, M.; Fang, Z.; Xu, X.; Zeng, L.; Chen, J.; Fan, S.; Du, X.; Zhang, X.; Yang, D.; Yang, Y.; Meng, H.; Li, W.; Melgiri, N. D.; Licinio, J.; Wei, H.; Xie, P. *Molecular Psychiatry* **2016**, DOI: 10.1038/mp.2016.44.

(7) Cryan, J. F.; Dinan, T. G. *Nature Reviews Neuroscience* **2012**, DOI: 10.1038/nrn3346.

(8) Hanstock, T. L.; Mallet, P. E.; Clayton, E. H. *Physiology and Behavior* **2010**, DOI: 10.1016/j.physbeh.2010.09.018.

(9) Wikoff, W. R.; Anfora, A. T.; Liu, J.; Schultz, P. G.; Lesley, S. A.; Peters, E. C.; Siuzdak, G. *Proceedings of the National Academy of Sciences of the United States of America* **2009**, DOI: 10.1073/pnas.0812874106.

(10) Ellis, C. L.; Ma, Z. M.; Mann, S. K.; Li, C. S.; Wu, J.; Knight, T. H.; Yotter, T.; Hayes, T. L.; Maniar, A. H.; Troia-Cancio, P. V.; Overman, H. A.; Torok, N. J.; Albanese, A.; Rutledge, J. C.; Miller, C. J.; Pollard, R. B.; Asmuth, D. M. *Journal of Acquired Immune Deficiency Syndromes* **2011**, DOI: 10.1097/QAI.0b013e31821a603c.

(11) Nowak, P.; Troseid, M.; Avershina, E.; Barqasho, B.; Neogi, U.; Holm, K.; Hov, J. R.; Noyan, K.; Vesterbacka, J.; Svärd, J.; Rudi, K.; Sönnerborg, A. *AIDS* **2015**, DOI: 10.1097/QAD.0000000000000869.

(12) Vujkovic-Cvijin, I.; Dunham, R. M.; Iwai, S.; Maher, M. C.; Albright, R. G.; Broadhurst, M. J.; Hernandez, R. D.; Lederman, M. M.; Huang, Y.; Somsouk, M.; Deeks, S. G.; Hunt, P. W.; Lynch, S. V.; McCune, J. M. *Science Translational Medicine* **2013**, DOI: 10.1126/scitranslmed.3006438.

(13) Ling, Z.; Jin, C.; Xie, T.; Cheng, Y.; Li, L.; Wu, N. *Scientific reports* **2016**, DOI: 10.1038/srep30673.

(14) Fukui, H. *Cellular & Molecular Medicine: Open access* **2016**, DOI: 10.21767/2573-5365.100023.

(15) Fukui, H. *Inflammatory Intestinal Diseases* **2016**, DOI: 10.1159/000447252.

(16) Hong, S.; Banks, W. A. *Brain, Behavior, and Immunity* **2015**, DOI: 10.1016/j.bbi.2014.10.008.

(17) Alford, K.; Vera, J. H. *British Medical Bulletin* **2018**, DOI: 10.1093/bmb/ldy019.

(18) Logsdon, A. F.; Erickson, M. A.; Rhea, E. M.; Salameh, T. S.; Banks, W. A. *Experimental Biology and Medicine* **2018**, DOI: 10.1177/1535370217743766.

(19) Braniste, V.; Al-Asmakh, M.; Kowal, C.; Anuar, F.; Abbaspour, A.; Tóth, M.; Korecka, A.; Bakocevic, N.; Guan, N. L.; Kundu, P.; Gulyás, B.; Halldin, C.; Hultenby, K.; Nilsson, H.; Hebert, H.; Volpe, B. T.; Diamond, B.; Pettersson, S. *Science Translational Medicine* **2014**, DOI: 10.1126/scitranslmed.3009759.

(20) Persidsky, Y.; Ramirez, S. H.; Haorah, J.; Kanmogne, G. D. *Journal of Neuroimmune Pharmacology* **2006**, DOI: 10.1007/s11481-006-9025-3.

(21) Hsiao, E. Y.; McBride, S. W.; Hsien, S.; Sharon, G.; Hyde, E. R.; McCue, T.; Codelli, J. A.; Chow, J.; Reisman, S. E.; Petrosino, J. F.; Patterson, P. H.; Mazmanian, S. K. *Cell* **2013**, DOI: 10.1016/j.cell.2013.11.024.

(22) Kelly, C. J.; Zheng, L.; Campbell, E. L.; Saeedi, B.; Scholz, C. C.; Bayless, A. J.; Wilson, K. E.; Glover, L. E.; Kominsky, D. J.; Magnuson, A.; Weir, T. L.; Ehrentraut, S. F.; Pickel, C.; Kuhn, K. A.; Lanis, J. M.; Nguyen, V.; Taylor, C. T.; Colgan, S. P. *Cell Host and Microbe* **2015**, DOI: 10.1016/j.chom.2015.03.005.

(23) Roager, H. M.; Licht, T. R. *Nature Communications* **2018**, DOI: 10.1038/s41467-018-05470-4.

(24) Aly, A. M.; Adel, A.; El-Gendy, A. O.; Essam, T. M.; Aziz, R. K. *Gut Pathogens* **2016**, DOI: 10.1186/s13099-016-0124-2.

(25) Inoue, T.; Nakayama, J.; Moriya, K.; Kawaratani, H.; Momoda, R.; Ito, K.; Iio, E.; Nojiri, S.; Fujiwara, K.; Yoneda, M.; Yoshiji, H.; Tanaka, Y. *Clinical Infectious Diseases* **2018**, DOI: 10.1093/cid/ciy205.

(26) Ponziani, F. R.; Putignani, L.; Paroni Sterbini, F.; Petito, V.; Picca, A.; Del Chierico, F.; Reddel, S.; Calvani, R.; Marzetti, E.; Sanguinetti, M.; Gasbarrini, A.; Pompili, M. *Alimentary Pharmacology and Therapeutics* **2018**, DOI: 10.1111/apt.15004.

(27) Bajaj, J. S.; Sterling, R. K.; Betrapally, N. S.; Nixon, D. E.; Fuchs, M.; Daita, K.; Heuman, D. M.; Sikaroodi, M.; Hylemon, P. B.; White, M. B.; Ganapathy, D.; Gillevet, P. M. *Alimentary Pharmacology and Therapeutics* **2016**, DOI: 10.1111/apt.13732.

(28) Pearson-Leary, J.; Zhao, C.; Bittinger, K.; Eacret, D.; Luz, S.; Vigderman, A. S.; Dayanim, G.; Bhatnagar, S. *Molecular Psychiatry* **2019**, DOI: 10.1038/s41380-019-0380-x.

(29) Ciesla, J. A.; Roberts, J. E. *American Journal of Psychiatry* **2001**, DOI: 10.1176/appi.ajp.158.5.725.

(30) Carabotti, M.; Scirocco, A.; Maselli, M. A.; Severi, C. *Annals of Gastroenterology* **2015**.

(31) Jiang, H.; Ling, Z.; Zhang, Y.; Mao, H.; Ma, Z.; Yin, Y.; Wang, W.; Tang, W.; Tan, Z.; Shi, J.; Li, L.; Ruan, B. *Brain, Behavior, and Immunity* **2015**, DOI: 10.1016/j.bbi.2015.03.016.

(32) Sudo, N.; Chida, Y.; Aiba, Y.; Sonoda, J.; Oyama, N.; Yu, X. N.; Kubo, C.; Koga, Y. *Journal of Physiology* **2004**, DOI: 10.1113/jphysiol.2004.063388.

(33) Burokas, A.; Arboleya, S.; Moloney, R. D.; Peterson, V. L.; Murphy, K.; Clarke, G.; Stanton, C.; Dinan, T. G.; Cryan, J. F. *Biological Psychiatry* **2017**, DOI: 10.1016/j.biopsych.2016.12.031.

(34) Zhang, F.; Yang, J.; Ji, Y.; Sun, M.; Shen, J.; Sun, J.; Wang, J.; Liu, L.; Shen, Y.; Zhang, R.; Chen, J.; Lu, H. *Frontiers in Microbiology* **2019**, DOI: 10.3389/fmicb.2018.03352.

(35) Lozupone, C.; Knight, R. *Applied and Environmental Microbiology* **2005**, DOI: 10.1128/AEM.71.12.8228-8235.2005.

(36) Lozupone, C.; Hamady, M.; Knight, R. *BMC Bioinformatics* **2006**, DOI: 10.1186/1471-2105-7-371.

(37) McDonald, D.; Vázquez-Baeza, Y.; Koslicki, D.; McClelland, J.; Reeve, N.; Xu, Z.; Gonzalez, A.; Knight, R. *Nature Methods* **2018**, DOI: 10.1038/s41592-018-0187-8.

(38) Falony, G.; Joossens, M.; Vieira-Silva, S.; Wang, J.; Darzi, Y.; Faust, K.; Kurilshikov, A.; Bonder, M. J.; Valles-Colomer, M.; Vandeputte, D.; Tito, R. Y.; Chaffron, S.; Rymenans, L.; Verspecht, C.; Sutter, L. D.; Lima-Mendez, G.; D'hoe, K.; Jonckheere, K.; Homola, D.; Garcia, R.; Tigchelaar, E. F.; Eeckhaudt, L.; Fu, J.; Henckaerts, L.; Zhernakova, A.; Wijmenga, C.; Raes, J. *Science* **2016**, DOI: 10.1126/science.aad3503.

(39) Huang, S.; Haiminen, N.; Carrieri, A.-P.; Hu, R.; Jiang, L.; Parida, L.; Russell, B.; Allaband, C.; Zarrinpar, A.; Vázquez-Baeza, Y.; Belda-Ferre, P.; Zhou, H.; Kim, H.-C.; Swafford, A. D.; Knight, R.; Xu, Z. Z. *mSystems* **2020**, DOI: 10.1128/msystems.00630-19.

(40) De la Cuesta-Zuluaga, J.; Kelley, S. T.; Chen, Y.; Escobar, J. S.; Mueller, N. T.; Ley, R. E.; McDonald, D.; Huang, S.; Swafford, A. D.; Knight, R.; Thackray, V. G. *mSystems* **2019**, DOI: 10.1128/msystems.00261-19.

(41) McDonald, D. et al. *mSystems* **2018**, DOI: 10.1128/msystems.00031-18.

(42) Armstrong, A. J.; Shaffer, M.; Nusbacher, N. M.; Griesmer, C.; Fiorillo, S.; Schneider, J. M.; Preston Neff, C.; Li, S. X.; Fontenot, A. P.; Campbell, T.; Palmer, B. E.; Lozupone, C. A. *Microbiome* **2018**, DOI: 10.1186/s40168-018-0580-7.

(43) Neff, C. P.; Krueger, O.; Xiong, K.; Arif, S.; Nusbacher, N.; Schneider, J. M.; Cunningham, A. W.; Armstrong, A.; Li, S.; McCarter, M. D.; Campbell, T. B.; Lozupone, C. A.; Palmer, B. E. *EBioMedicine* **2018**, DOI: 10.1016/j.ebiom.2018.03.024.

(44) Noguera-Julian, M.; Rocafort, M.; Guillén, Y.; Rivera, J.; Casadellà, M.; Nowak, P.; Hildebrand, F.; Zeller, G.; Parera, M.; Bellido, R.; Rodríguez, C.; Carrillo, J.; Mothe, B.; Coll, J.; Bravo, I.; Estany, C.; Herrero, C.; Saz, J.; Sirera, G.; Torrela, A.; Navarro, J.; Crespo, M.; Brander, C.; Negredo, E.; Blanco, J.; Guarner, F.; Calle, M. L.; Bork, P.; Sönnerborg, A.; Clotet, B.; Paredes, R. *EBioMedicine* **2016**, DOI: 10.1016/j.ebiom.2016.01.032.

(45) Kelley, C. F.; Kraft, C. S.; De Man, T. J.; Duphare, C.; Lee, H. W.; Yang, J.; Easley, K. A.; Tharp, G. K.; Mulligan, M. J.; Sullivan, P. S.; Bosinger, S. E.; Amara, R. R. *Mucosal Immunology* **2017**, DOI: 10.1038/mi.2016.97.

(46) Lozupone, C. A.; Li, M.; Campbell, T. B.; Flores, S. C.; Linderman, D.; Gebert, M. J.; Knight, R.; Fontenot, A. P.; Palmer, B. E. *Cell Host and Microbe* **2013**, DOI: 10.1016/j.chom.2013.08.006.

(47) Mutlu, E. A.; Keshavarzian, A.; Losurdo, J.; Swanson, G.; Siewe, B.; Forsyth, C.; French, A.; DeMarais, P.; Sun, Y.; Koenig, L.; Cox, S.; Engen, P.; Chakradeo, P.; Abbasi, R.; Gorenz, A.; Burns, C.; Landay, A. *PLoS Pathogens* **2014**, DOI: 10.1371/journal.ppat.1003829.

(48) Dillon, S. M.; Lee, E. J.; Kotter, C. V.; Austin, G. L.; Dong, Z.; Hecht, D. K.; Gianella, S.; Siewe, B.; Smith, D. M.; Landay, A. L.; Robertson, C. E.; Frank, D. N.; Wilson, C. C. *Mucosal Immunology* **2014**, DOI: 10.1038/mi.2013.116.

(49) Vázquez-Castellanos, J. F.; Serrano-Villar, S.; Latorre, A.; Artacho, A.; Ferrús, M. L.; Madrid, N.; Vallejo, A.; Sainz, T.; Martínez-Botas, J.; Ferrando-Martínez, S.; Vera, M.; Dronda, F.; Leal, M.; Del Romero, J.; Moreno, S.; Estrada, V.; Gosalbes, M. J.; Moya, A. *Mucosal Immunology* **2015**, DOI: 10.1038/mi.2014.107.

(50) Shannon, C. E. *Bell System Technical Journal* **1948**, DOI: 10.1002/j.1538-7305.1948.tb01338.x.

(51) Serrano-Villar, S.; Deeks, S. G. *The Lancet HIV* **2015**, DOI: 10.1016/S2352-3018(15)00018-1.

(52) Grund, B.; Baker, J. V.; Deeks, S. G.; Wolfson, J.; Wentworth, D.; Cozzi-Lepri, A.; Cohen, C. J.; Phillips, A.; Lundgren, J. D.; Neaton, J. D. *PLoS ONE* **2016**, DOI: 10.1371/journal.pone.0155100.

(53) Duprez, D. A.; Neuhaus, J.; Kuller, L. H.; Tracy, R.; Belloso, W.; De Wit, S.; Drummond, F.; Lane, H. C.; Ledergerber, B.; Lundgren, J.; Nixon, D.; Paton, N. I.; Prineas, R. J.; Neaton, J. D. *PLoS ONE* **2012**, DOI: 10.1371/journal.pone.0044454.

(54) Ross, A. C.; Rizk, N.; O'Riordan, M. A.; Dogra, V.; ElBejjani, D.; Storer, N.; Harrill, D.; Tungsiripat, M.; Adell, J.; McComsey, G. A. *Clinical Infectious Diseases* **2009**, DOI: 10.1086/605578.

(55) Ford, E. S.; Greenwald, J. H.; Richterman, A. G.; Rupert, A.; Dutcher, L.; Badralmaa, Y.; Natarajan, V.; Rehm, C.; Hadigan, C.; Sereti, I. *AIDS* **2010**, DOI: 10.1097/QAD.0b013e32833ad914.

(56) Chahwan, B.; Kwan, S.; Isik, A.; van Hemert, S.; Burke, C.; Roberts, L. *Journal of Affective Disorders* **2019**, DOI: 10.1016/j.jad.2019.04.097.

(57) Kleiman, S. C.; Bulik-Sullivan, E. C.; Glenny, E. M.; Zerwas, S. C.; Huh, E. Y.; Tsilimigras, M. C.; Fodor, A. A.; Bulik, C. M.; Carroll, I. M. *PLoS ONE* **2017**, DOI: 10.1371/journal.pone.0170208.

(58) Ridlon, J. M.; Kang, D. J.; Hylemon, P. B.; Bajaj, J. S. *Current Opinion in Gastroenterology* **2014**, DOI: 10.1097/MOG.0000000000000057.

(59) Chiang, J. Y. *Comprehensive Physiology* **2013**, DOI: 10.1002/cphy.c120023.

(60) Pols, T. W.; Noriega, L. G.; Nomura, M.; Auwerx, J.; Schoonjans, K. *Journal of Hepatology* **2011**, DOI: 10.1016/j.jhep.2010.12.004.

(61) Broeders, E. P.; Nascimento, E. B.; Havekes, B.; Brans, B.; Roumans, K. H.; Tailleux, A.; Schaart, G.; Kouach, M.; Charton, J.; Deprez, B.; Bouvy, N. D.; Mottaghy, F.; Staels, B.; Van Marken Lichtenbelt, W. D.; Schrauwen, P. *Cell Metabolism* **2015**, DOI: 10.1016/j.cmet.2015.07.002.

(62) Arab, J. P.; Karpen, S. J.; Dawson, P. A.; Arrese, M.; Trauner, M. *Hepatology* **2017**, DOI: 10.1002/hep.28709.

(63) Chhatwal, P.; Bankwitz, D.; Gentzsch, J.; Frentzen, A.; Schult, P.; Lohmann, V.; Pietschmann, T. *PloS one* **2012**, DOI: 10.1371/journal.pone.0036029.

(64) Caspani, G.; Kennedy, S.; Foster, J. A.; Swann, J. *Microbial Cell* **2019**, DOI: 10.15698/mic2019.10.693.

(65) Morton, J. T.; Marotz, C.; Washburne, A.; Silverman, J.; Zaramela, L. S.; Edlund, A.; Zengler, K.; Knight, R. *Nature Communications* **2019**, DOI: 10.1038/s41467-019-10656-5.

(66) Fedarko, M. W.; Martino, C.; Morton, J. T.; González, A.; Rahman, G.; Marotz, C. A.; Minich, J. J.; Allen, E. E.; Knight, R. *bioRxiv* **2019**, DOI: 10.1101/2019.12.17.880047.

(67) Peirce, J. M.; Alviña, K. *Journal of Neuroscience Research* **2019**, DOI: 10.1002/jnr.24476.

(68) Naseribafrouei, A.; Hestad, K.; Avershina, E.; Sekelja, M.; Linløkken, A.; Wilson, R.; Rudi, K. *Neurogastroenterology and Motility* **2014**, DOI: 10.1111/nmo.12378.

(69) Bastiaanssen, T. F.; Cussotto, S.; Claesson, M. J.; Clarke, G.; Dinan, T. G.; Cryan, J. F. *Harvard Review of Psychiatry* **2020**, DOI: 10.1097/HRP.0000000000000243.

(70) Kelly, J. R.; Borre, Y.; O' Brien, C.; Patterson, E.; El Aidy, S.; Deane, J.; Kennedy, P. J.; Beers, S.; Scott, K.; Moloney, G.; Hoban, A. E.; Scott, L.; Fitzgerald, P.; Ross, P.; Stanton, C.; Clarke, G.; Cryan, J. F.; Dinan, T. G. *Journal of Psychiatric Research* **2016**, DOI: 10.1016/j.jpsychires.2016.07.019.

(71) Heaton, R. K.; Franklin, D. R.; Ellis, R. J.; McCutchan, J. A.; Letendre, S. L.; LeBlanc, S.; Corkran, S. H.; Duarte, N. A.; Clifford, D. B.; Woods, S. P.; Collier, A. C.; Marra, C. M.; Morgello, S.; Rivera Mindt, M.; Taylor, M. J.; Marcotte, T. D.; Atkinson, J. H.; Wolfson, T.; Gelman, B. B.; McArthur, J. C.; Simpson, D. M.; Abramson, I.; Gamst, A.; Fennema-Notestine, C.; Jernigan, T. L.; Wong, J.; Grant, I. *Journal of NeuroVirology* **2011**, DOI: 10.1007/s13365-010-0006-1.

(72) Marquine, M. J.; Heaton, A.; Johnson, N.; Rivera-Mindt, M.; Cherner, M.; Bloss, C.; Hulgan, T.; Umlauf, A.; Moore, D. J.; Fazeli, P.; Morgello, S.; Franklin, D.; Letendre, S.; Ellis, R.; Collier, A. C.; Marra, C. M.; Clifford, D. B.; Gelman, B. B.; Sacktor, N.; Simpson, D.; McCutchan, J. A.; Grant, I.; Heaton, R. K. *Journal of the International Neuropsychological Society* **2018**, DOI: 10.1017/S1355617717000832.

(73) Composite International Diagnostic Interview, version 2.1. Geneva., 1997.

(74) Beck, A.; Steer, R.; GK., B. Beck Depression Inventory: Second Edition manual., 1996.

(75) Marotz, C.; Amir, A.; Humphrey, G.; Gaffney, J.; Gogul, G.; Knight, R. *BioTechniques* **2017**, DOI: 10.2144/000114559.

(76) *Nature Biotechnology* **2019**, DOI: 10.1038/s41587-019-0209-9.

(77) McDonald, D.; Kaehler, B.; Gonzalez, A.; DeReus, J.; Ackermann, G.; Marotz, C.; Huttley, G.; Knight, R. *mSystems* **2019**, DOI: 10.1128/msystems.00215-19.

(78) Anderson, M. J. *Austral Ecology* **2005**, DOI: 10.1139/cjfas-58-3-626.

(79) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orei, M. *BMC Bioinformatics* **2010**, DOI: 10.1186/1471-2105-11-395.

(80) Wang, M. et al. *Nature Biotechnology* **2016**, DOI: 10.1038/nbt.3597.

(81) Nothias, L. F.; Petras, D.; Schmid, R.; Dührkop, K.; Rainer, J.; Sarvepalli, A.; Protsyuk, I.; Ernst, M.; Tsugawa, H. *biorXiv* **2019**.

(82) Sumner, L. W.; Amberg, A.; Barrett, D.; Beale, M. H.; Beger, R.; Daykin, C. A.; Fan, T. W.; Fiehn, O.; Goodacre, R.; Griffin, J. L.; Hankemeier, T.; Hardy, N.; Harnly, J.; Higashi, R.; Kopka, J.; Lane, A. N.; Lindon, J. C.; Marriott, P.; Nicholls, A. W.; Reily, M. D.; Thaden, J. J.; Viant, M. R. *Metabolomics* **2007**, DOI: 10.1007/s11306-007-0082-2.

(83) Xia, J.; Sinelnikov, I. V.; Han, B.; Wishart, D. S. *Nucleic Acids Research* **2015**, DOI: 10.1093/nar/gkv380.

(84) Taylor, B. C.; Lejzerowicz, F.; Poirel, M.; Shaffer, J. P.; Jiang, L.; Aksenov, A.; Litwin, N.; Humphrey, G.; Martino, C.; Miller-Montgomery, S.; Dorrestein, P. C.; Veiga, P.; Song, S. J.; McDonald, D.; Derrien, M.; Knight, R. *mSystems* **2020**, DOI: 10.1128/msystems.00901-19.

(85) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. *Genome Research* **2003**, DOI: 10.1101/gr.1239303.

Chapter 6

Consumption of Fermented Foods Is Associated with Systematic Differences in the Gut Microbiome and Metabolome

6.1 Abstract

Lifestyle factors, such as diet, strongly influence the structure, diversity, and composition of the microbiome. While we have witnessed over the last several years a resurgence of interest in fermented foods, no study has specifically explored the effects of their consumption on gut microbiota in large cohorts. To assess whether the consumption of fermented foods is associated with a systematic signal in the gut microbiome and metabolome, we used a multi-omic approach (16S rRNA amplicon sequencing, metagenomic sequencing, and untargeted mass spectrometry) to analyze stool samples from 6,811 individuals from the American Gut Project, including 115 individuals specifically recruited for their frequency of fermented food consumption for a targeted four-week longitudinal study. We observed subtle but statistically significant differences between consumers and non-consumers in beta diversity as well as differential taxa between the two groups. We found that the metabolome of fermented food consumers was enriched with conjugated linoleic acid (CLA), a putatively health-promoting molecule. Cross-omic

analyses between metagenomic sequencing and mass spectrometry suggest that CLA may be driven by taxa associated with fermented food consumers. Collectively, we found modest yet persistent signatures associated with fermented food consumption that appear present in multiple -omic types which motivate further investigation of how different types of fermented food impact the gut microbiome and overall health.

6.2 Importance

Public interest in the effects of fermented food on the human gut microbiome is high but limited studies have explored the association between fermented food consumption and the gut microbiome in large cohorts. Here we used a combination of omics-based analyses to study the relationship between the microbiome and fermented food consumption in thousands of people using both cross-sectional and longitudinal data. We found that fermented food consumers have subtle differences in their gut microbiota structure, which is enriched in conjugated linoleic acid, thought to be beneficial. The results suggest that further studies of specific kinds of fermented food and their impacts on the microbiome and health will be useful.

6.3 Introduction

Fermentation is an ancient process of food preparation dating from the introduction of agriculture and animal husbandry during the Neolithic period approximately 10,000 years ago. Advantages of food fermentation include improvements in food preservation, food safety, nutritional value, and organoleptic quality resulting from the activity of microbial ecosystems (bacteria and yeast) [1]. Fermentation can be applied to a range of food types including meat, fish, milk, vegetables, beans, cereals, and fruits, and occurs spontaneously from the original ingredients or environment, or is controlled by the addition of specific starters such as lactic acid bacteria (LAB) [2]. These bacteria are commonly detected in fermented food, mostly including *Lactobacillus*, *Streptococcus*, *Lactococcus* and *Leuconostoc*; but other bacteria as well as yeast and fungi are also involved in food fermentations [3]. In addition to microbial diversity, the number of microorganisms present in fermented foods vary between food type, process, and storage. A survey of diverse fermented food products suggested that the count of viable lactic acid bacteria usually reaches at least 10^6cells/mL [4]. Recovery of viable bacterial and fungal species ingested through fermented food has been observed in subjects who consume an animal-based diet [5]. Moreover, metabolites generated from fermentation, including lactic acid, vitamins, and exopolysaccharides, are thought to exert health benefits [6]. A recent study reported that D-phenyllactic acid, produced by LAB, interacts with the human host through the activation of hydroxycarboxylic acid receptor 3 (HCA3), and is involved in the regulation of immune functions and energy homeostasis under changing metabolic and dietary conditions [7].

Due to their supposed health benefits [6], there has been a resurgence of interest in consumption of fermented foods in Western society. To date, many of the studies focused on the health benefits of fermented food intake have been mostly focused on yogurt, consumption of which is associated with better metabolic parameters in large American cohorts [8, 9]. Similarly, high intake of fermented foods has been associated with a lower prevalence of atopic dermatitis in a Korean population [10], and another study found consumption of miso and natto to be inversely associated with high blood pressure in a Japanese population [11].

While we know that both short and long-term dietary intake affects the structure, function, and activity of the human gut microbiome [5, 12–16], and a few studies have explored the response of gut microbiota to a single type of fermented food (recently reviewed by [17]), no study has explored the functional capacity of the gut microbiota of fermented food consumers. Intervention studies, which are often underpowered for analysis of the gut microbiome response, are complemented by studies of population-based cohorts, which due to large sample sizes have the advantage of capturing large amounts of microbial variation and enable us to disentangle the contributions of host and environmental factors such as diet [18–21].

To address the hypothesis that fermented food consumption is associated with compositional or functional changes in the human gut microbiome, we analysed a subset of the American Gut Project (AGP) cohort based on self-reported consumption of fermented foods, and in particular, fermented plants. We also explored the longitudinal stability and function of the gut microbiota using untargeted HPLC-MS/MS mass spectrometry and 16S rRNA amplicon sequencing, as well as shotgun sequencing on a subset of subjects at a single time point.

6.4 Results

6.4.1 Demographic and dietary assessments of fermented plant consumers and non-consumers.

To explore the differences in the gut microbiome between fermented consumers and non-consumers, we analyzed 16S rRNA sequencing data from 28,114 samples from 21,464 individuals in the AGP (FIG 6.1A). After filtering (see Materials and Methods), 6,811 participants were retained, and hereafter are referred to as the cross-sectional cohort (FIG 6.1A). 115 of these participants were initially recruited for a concurrent longitudinal assessment which is discussed in detail below. Participants were identified as "consumers" or "non-consumers" depending on the frequency of fermented plants they reported consuming. The fermented plant frequency question is in the standard AGP questionnaire that every participant answered, and while the language may not have allowed for the capture of all fermented foods, this represented the most efficient way to delineate consumers and non-consumers. We considered consumers to be those who reported eating fermented plants "Daily", "Regularly (3-5 times/week)", or "Occasionally (1-2 times/week)", and non-consumers as those who reported eating fermented plants either "Rarely (less than once/week)" or "Never" (FIG 6.1B). 30.5% of participants were considered consumers, of which most (45.3%) were occasional consumers. Consumer and non-consumer cohorts were composed of slightly differing demographic groups. For example, while consumers were significantly younger than non-consumers, the difference was modest (47 versus 47.61 years respectively), with a higher proportion of participants in their 30s (23.0% vs 19.4%; Chi2 = 11.08, p = 0.03) (FIG 6.1B). Similarly, the consumer group was composed of a modestly higher proportion of females (56.8% vs 52.6%; Chi2 = 9.60, p = 0.002), and a higher proportion of participants with a normal BMI between 18.5 and 25 (65.6% vs 59.3%; Chi2 = 35.93, p « 0.001), with an average BMI of 23.9 and 24.8 respectively. Consumers also reported eating a greater diversity of plants (>20) (29.7% vs 24.5%; Chi2 = 126.96,p « 0.001). In addition, because alcohol may be an end product of a fermentation process and might be a confounding factor associated with gut microbiota variation, we verified that alcohol consumption was not associated with fermented plant consumption (81.7% vs 82.6%, Chi2 = 0.76, p-value = 0.38).

Statistically significant differences in mean total carbohydrate and fat intake (g/d and % of energy)

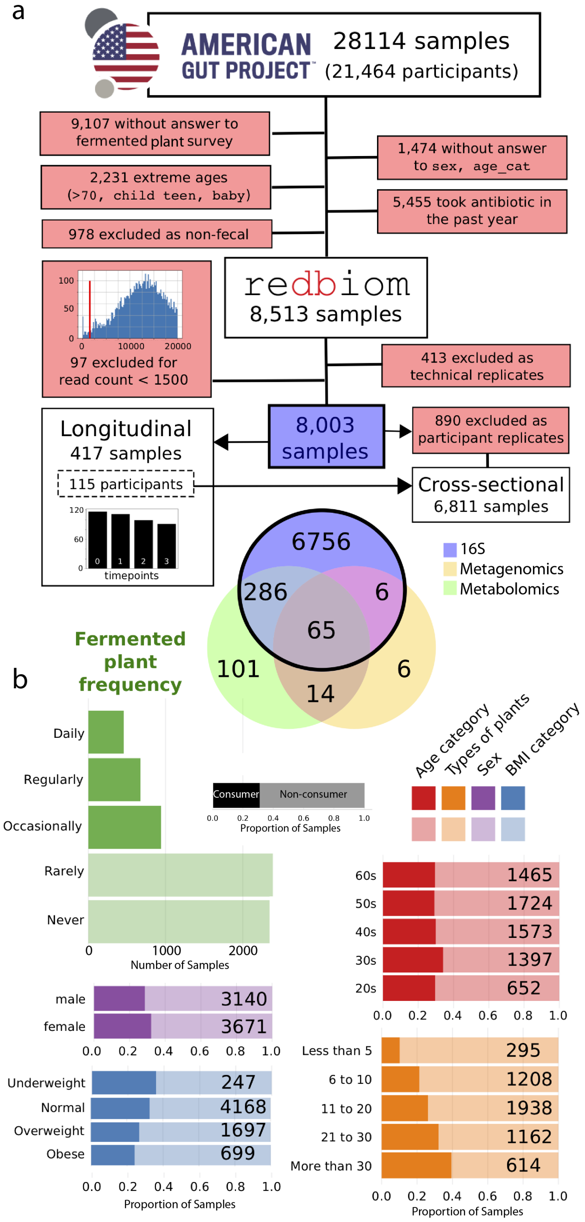


Figure 6.1: Cohort overview, sample filtering and metadata exploration. (A). Data filtering process and the number of samples analyzed by 16S rRNA gene sequencing, metabolomics, and shotgun metagenomics and the resulting number of samples in the cross-sectional and longitudinal cohorts. (B). Distribution of some metadata categories (demographic and diet) in the cross-sectional cohort between consumers and non-consumers:. Darker colors denote consumers, lighter colors denote non-consumers. The consumer and non-consumer groups were defined by the “Fermented plant frequency” questionnaire.

and percentage of energy from protein, as estimated by the FFQ, were observed between fermented plant consumers and non-consumers, while total energy (kcal/d), dietary fiber (g/d) and protein (g/d) intake did not differ (Table S1). There was no significant difference in overall diet quality observed, as assessed by the Healthy Eating Index (HEI-2010; Mann-Whitney, U = 223409, p value = 0.094; FIG S1A), despite the differences in the consumption of fermented plants and number of plant types between consumers and non-consumers, this non-significant difference in total HEI-2010 scores between consumers and non-consumers (71.29 vs. 71.53, respectively) suggests similar intake of dietary patterns relatively high in quality. It should be noted that the mean total HEI-2010 score for both consumers and non-consumers is above the national average (58.27) for U.S. adults aged 18-64 years based on 2011-2012 National Health and Nutrition Examination (NHANES) data [22]. This suggests that the cohort in our study has a diet pattern that better aligns to the Dietary Guidelines for Americans than average American adults. Additionally, it has been shown that higher HEI scores are associated with higher income and education levels [23, 24], thereby suggesting that the higher total HEI scores observed in this AGP cohort may reflect higher than average socioeconomic status and education level as previously observed [25].

6.4.2 Gut microbiome composition in fermented plant consumers and non-consumers.

Examining unweighted UniFrac distances [26], we observed a statistically significant difference in the overall gut microbial communities between consumers and non-consumers (FIG S1B, PERMANOVA pseudo-F statistic = 3.677, p=0.001). The comparison of non-consumers with occasional consumers results in a weaker group separation (F-statistic = 2.233, p-value = 0.001) than with regular or daily consumers (F-statistics = 3.512 and 3.246, respectively, p-values = 0.001), suggesting a dose-dependence for the frequency of fermented plant consumption on the gut microbiome. However there was no dose-dependence with frequency of types of plants between consumers and non-consumers (Unweighted Unifrac distances between consumers and non-consumers versus the number of types of plants frequency, $R^2 = 0.0065$). There was no difference in alpha diversity between the two groups (Faiths PD, Shannon Diversity, nor Observed OTUs richness, FIG S1B) and also no difference when groups were stratified by consumption frequency (Table S2).

Next, we used Songbird [27] to identify specific microbes that were associated with consumers or non-consumers. Songbird is a compositionally-aware differential abundance method which provides rankings of features (sub-operational taxonomic units, or sOTUs) based on their log-fold change with respect to covariates of interest. In this case, the formula we used described whether the subject consumed fermented plants or not. We selected the 20 highest (“Set 1”, Table S3) and 20 lowest (“Set 2”, Table S3) ranked sOTUs associated with fermented plant consumption and used Qurro[28] to compute the log-ratio of these sets of taxa (FIG S1C). Comparing the ratio of taxa in this way mitigates bias from the unknown total microbial load in each sample, and taking the log of this ratio gives equal weight to relative increases and decreases of taxa [27]. Evaluation of the Songbird model for fermented plant consumption against a baseline model obtained a Q^2-value of -5.4249, suggesting possible overfitting related to the subtlety of the differences between fermented plant consumption groups. In order to verify the log-ratios chosen by Songbird ranks, we performed a permutation test by taking 1000 random permutations of log-ratios with 20 non-overlapping features in the numerator and denominator. The rank orders, compared to the random permutation, was 16 corresponding to a p-value of 0.0159 (FIG S2A), suggesting that the log-ratio based on the Songbird ranks is non-random. We found that consumers have a significantly higher log-ratio of Set 1 / Set 2 than non-consumers (t-test, p = 0.00065, t = 3.6367) suggesting that they are associated with *Bacteroides* spp., *Pseudomonas* spp., *Dorea* spp., Lachnospiraceae, *Prevotella* spp., *Alistipes putredinis*, *Oscillospira* spp., Enterobacteriaceae, *Fusobacterium* spp., *Actinomyces* spp., *Achromobacter* spp., *Clostridium clostridioforme*, *Faecalibacterium prausnitzii*, *Bacteroides uniformis*, Clostridiales, and *Delftia* spp.

6.4.3 Gut microbiome composition in frequent and rare fermented food consumers.

115 participants were recruited for a longitudinal study in order to assess the gut microbiome over time and at a finer resolution by using untargeted mass spectrometry in addition to 16S rRNA sequencing (FIG 6.1A). We targeted participants who self-identified as frequent consumers or very rare consumers. Consumers were identified using the same definition as in the cross-sectional cohort: consumers ate fermented plants “Daily”, “Regularly (3-5 times/week)”, or “Occasionally (1-2 times/week)”; Non-consumers “Rarely (less than once/week)” or “Never” (FIG S3). The longitudinal cohort was designed to have a higher proportion of consumers who reported eating fermented plants “Daily” and “Regularly”

versus “Occasionally” when compared to the cross-sectional cohort (FIG S4). Similarly, the non-consumer group in the longitudinal cohort had a higher proportion of participants who reported eating them “Never” and “Rarely” (FIG S4) compared to non-consumers in the cross-sectional study.

A separate fermented food questionnaire was provided to these 115 participants to characterize additional types of fermented food consumed, and to evaluate the proxy of fermented plant consumption for general fermented food consumption. Briefly, the major fermented foods consumed were beer, kimchi, kombucha, pickled vegetables, sauerkraut, and yogurt. More consumers reported eating fermented foods than non-consumers (FIG S3B). Only 7.0% of participants (8 / 115) who stated that they never consumed fermented plants reported consuming another type of fermented food. Of these eight participants, two reported that they consumed wine or beer; one participant reported consuming yogurt, cider, wine, and beer; and five participants reported consuming unspecified fermented foods. We also observed that fermented plant consumers more frequently ate fermented dairy products (yogurt, sour cream/creme fraiche, kefir milk, and cottage cheese) than non-consumers (FIG S3B). Therefore, we further identified them as “fermented food consumers”, in contrast to the cross-sectional cohort.

Within the 16S data, we did not observe a difference in alpha diversity (Shannons Index [30]) and Faiths Phylogenetic Diversity [31] between consumers and non-consumers (FIG S1B). We further applied a Sparse Functional Principal Components Analysis[32], which explicitly factors in the longitudinal component, and did not observe a significant difference in alpha diversity (Shannons Index, Wilcoxon p = 0.20) suggesting that the stability of alpha diversity in the microbiome over four weeks is consistent for consumers and non-consumers.

A subset of 100 samples were sequenced by shotgun metagenomics to provide a finer resolution of the taxonomic differences between the two groups. First, we verified whether the gut microbiota of self-reported fermented food consumers were associated with fermented food associated species. We computed a log-ratio using Qurro[28] of fermented food associated taxa according to Marco et al. 2017 [6] (“Set 3”, Table S3) compared to a set of taxa that were present across all samples (“Set 4”, Table S3) (FIG 6.2B). Eight species were detected in our dataset and were used to compute this log-ratio: *Lactobacillus acidophilus*, *Lactobacillus brevis*, *Lactobacillus fermentum*, *Lactococcus lactis*, *Leuconostoc mesenteroides*, *Lactobacillus paracasei*, *Lactobacillus plantarum*, and *Lactobacillus rhamnosus* (FIG

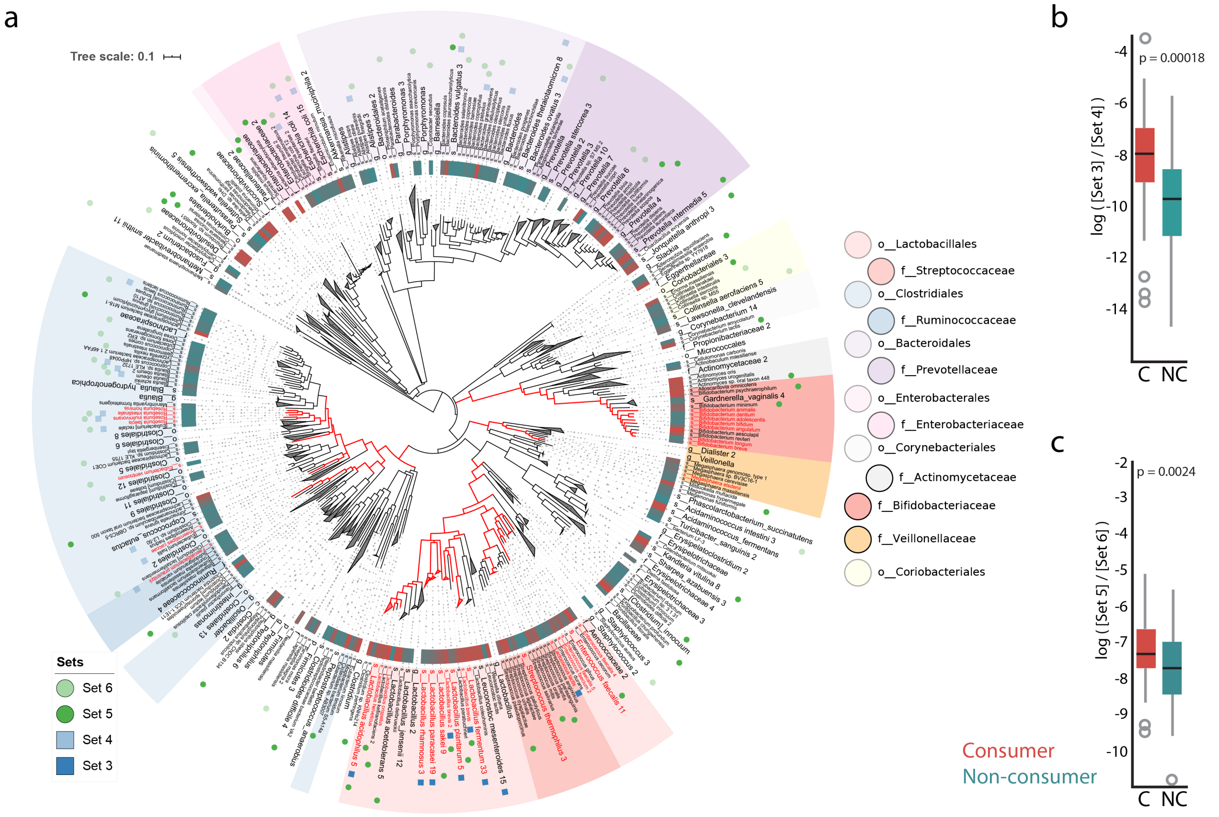


Figure 6.2: Phylogenetic and log-ratio differences between consumers and non-consumers in gut metagenomes. (A) Phylogenetic diversity captured in the metagenomes. The species taxa hits to the rep82 database were mapped to the species taxa of the Web of Life database[29] (97.8% mapped) in order to represent the phylogenetic distances computed from full genomes. The species known to produce CLA are indicated in red. The species sets used for log-ratio calculations are labeled using opaque (numerators) and transparent (denominators) colors. Set 3 is composed of microbes identified from [6] and Set 4 contains the most prevalent microbes across all samples (blue squares). Set 5 and Set 6 are derived from Songbird (green circles). (B) Consumers have a significantly higher ratio of [Set 3] / [Set 4] (t-test, p-value = 0.0001838, t = 3.9386, Cohens D = 0.851). (C) The log-ratio of [Set 5] / [Set 6] is significantly different (Boxplot, t-test p = 0.0024, t = 3.15, Cohens D = 0.692). The lists of microbes in each set are available in Table S3.

6.2A). We found that consumers had a significantly higher log-ratio of Set 3 / Set 4 than non-consumers (t-test, p-value = 0.0001838, t = 3.9386, Cohens D = 0.851), suggesting that consumers were associated with some taxa derived from fermented foods.

We then used Songbird [27] to test whether there was a broader set of microbial features associated with consumers or non-consumers. We selected the 40 highest-ranked ("Set 5", Table S3) and 40 lowest-ranked ("Set 6", Table S3) microbes associated with fermented plant consumption and used Qurro to compute the log-ratio of these sets of taxa (FIG 6.2C); these were the smallest sets of features that provided meaningful differences between consumers and non-consumers. Again, because evaluation of the Songbird models for fermented plant consumption against a baseline model suggested overfitting (Q^2-value of -0.12), we further verified the log-ratios chosen by Songbird ranks by performing a permutation test of taking 1000 random permutations of log-ratios with 20 non-overlapping features in the numerator and denominator. The rank orders, compared to the random permutation, was 2 corresponding to a p-value of 0.0019 (FIG S2B), suggesting that the log-ratio based on the Songbird ranks is non-random. This analysis at the species-level showed that consumers have a significantly higher log-ratio of Set 5 / Set 6 than non-consumers (t-test p = 0.0024, t = 3.15, Cohens D = 0.692).

Several microbes of relevance to fermented foods were also associated with consumers, including *Lactobacillus acidophilus*, *Lactobacillus brevis*, *Lactobacillus kefiranofaciens*, *Lactobacillus parabuchneri*, *Lactobacillus helveticus*, and *Lactobacillus sakei* [6, 33–36] (6.2A). Consumers were also associated with several other microbes unrelated to fermented foods, including *Streptococcus dysgalactiae*, *Prevotella melaninogenica*, *Enorma massiliensis*, *Prevotella multiformis*, *Enterococcus cecorum*, and *Bacteroides paurosaccharolyticus*. The microbes that distinguish consumers and non-consumers in the cross-sectional and longitudinal datasets may not fully overlap because the longitudinal cohort was intentionally composed of participants in the more "extreme" ends of consumption (individuals who consume "Daily" and "Regularly" versus individuals who "Never" consume fermented plants), because the cohorts were analyzed using different sequencing methods (16S vs. metagenomics), or a combination of these aspects.

6.4.4 The functional profile of the gut microbiome differs with consumption of fermented food.

To assess the functional profile of the gut microbiome of specifically recruited fermented food consumers and non-consumers, we performed untargeted HPLC-MS/MS analysis on all longitudinal samples (115 subjects, 417 samples, with up to 4 samples per subject, collected weekly for four weeks) (FIG 6.1A). We explored the longitudinal stability using both the 16S and mass spectrometry data and found that the taxa and metabolites remained stable (Spearmans rho ranging from 0.42 to 0.68; $p < 0.001$) between time points within both consumers and non-consumers (FIG S5). The correlation coefficients for metabolites tended to be lower than for the taxa, suggesting more volatility in the observed metabolic features. This is expected since the metabolome is driven in large part by the diet which changes day-to-day.

Using Partial Least Squares Discriminatory Analysis (PLS-DA) we found that notable differences exist between consumers and non-consumers when all time points were taken into account (FIG 6.3A, FIG S6A). The majority of the top discriminating features appeared to be lipids, several of which have broad natural distributions and thus are likely common. In particular, one compound was identified as octadecadienoic acid, then determined specifically to be an isomer of conjugated linoleic acid (CLA). At a single time point, we found that this isomer of CLA (denoted "CLA4", the exact configuration is unknown) was enriched in consumers (Wilcox test, p value = 0.04) whereas the unconjugated Linoleic Acid (LA) was not significantly different between the two groups (Wilcox test, p value = 0.52) (FIG 6.3B). As CLA has also been found as one of discriminating features in samples from subjects who consume a large number of types of plants[25], it might suggest that the difference between consumers and non-consumers could be partly explained by the number of types of plants consumed. However, in this study CLA abundances were not significantly different between the two extreme groups of types of plant consumption: less than 10 types of plants vs. more than 30 types of plants (Wilcoxon rank sums test, p value = 0.98). From the food frequency questionnaire, we found that dietary consumption of total LA (18:2 n-6; g/d) and total CLA (g/d) did not differ significantly between consumers and non-consumers (FIG S6B) suggesting that the elevated level of CLA in the fecal samples of consumers are likely derived from an endogenous process or microbial origin.

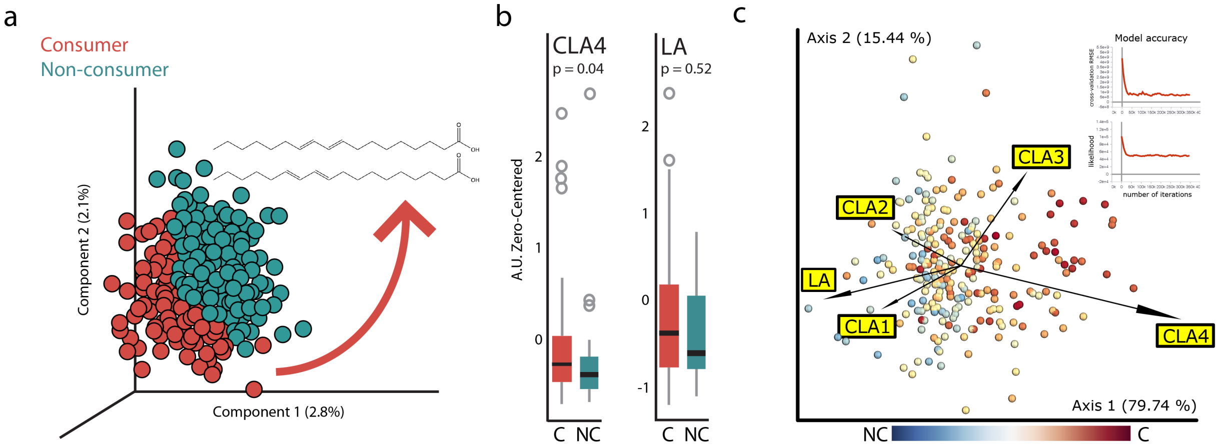


Figure 6.3: Conjugated Linoleic Acid is significantly higher in consumers than non-consumers. (A) Partial Least Squares Discriminatory Analysis (PLS-DA) of untargeted mass spectrometry data identified CLA as one of the discriminating features in fermented consumer samples. (B) Zero-centered counts of MS1 features annotated with a CLA isomer (denoted "CLA4") and with the unconjugated Linoleic Acid (LA) between consumers and non-consumers. "CLA4" is enriched in the consumer group (Wilcox test, p value = 0.04), but not LA (Wilcox test, p value = 0.52). (C) Integrative analysis of metagenomics and mass spectrometry datasets using mmvec. Genome features (dots) are labelled according to their strength of change with respect to fermentation food consumption (red is associated with consumption, blue with non-consumption). The metabolites are represented by arrows indicating their co-occurrences with the genomes.

A total of 79 samples were analyzed using both metagenomic sequencing and mass spectrometry (FIG 6.1A). We used mmvec [37] to integrate these data to assess co-occurrence patterns between genomic features (species) and the LA and CLA metabolites. We found that "CLA4", which was significantly enriched in consumers, co-occurs with the species (previously identified using Songbird) that were most strongly associated with consumers. Additionally, we found that linoleic acid (LA) co-occurs with the microbes that are most strongly associated with non-consumers (FIG 6.3C). Of the top 50 taxa that had the highest probability of co-occurring with "CLA", 14 are known CLA producers. These include *Eubacterium rectale*, *Faecalibacterium prausnitzii*, *Eubacterium siraeum*, *Eubacterium hallii*, *Bifidobacterium adolescentis*, and genera *Roseburia*, *Anaerostipes*, *Eubacterium*, *Ruminococcus*, and *Clostridium* (Table S4) [38–41]. 48 out of these top 50 taxa were more abundant in consumers than non-consumers (Table S4).

6.5 Discussion

In this study we explored the gut microbiome of fermented plant consumers and non-consumers in the American Gut Project [25], an extensive collection of sample contributions from tens of thousands of citizen scientists. Gut microbiome profiles, but not overall microbial diversity, differed slightly between the groups, suggesting that small but systematic compositional differences may occur based on a dietary choice to consume fermented plants. In a concurrent targeted longitudinal study we found that fermented-food related taxa as well as a putatively health-associated molecule were associated with consumers. Several microbes that were found to be associated with fermented consumers include microbes known to be derived from fermented foods, including fermented milk products (*Lactobacillus acidophilus* [6], *Lactobacillus brevis* [6], *Lactobacillus kefiranofaciens* [33]), *Lactobacillus parabuchneri* [34], *Lactobacillus helveticus* [35]) and fermented meat (*Lactobacillus sakei* [36]). This is consistent with other metagenomic studies from population-based cohorts that detected species related to starters such as *Leuconostoc mesenteroides* and *Lactococcus lactis* in subjects who consumed a specific fermented milk product (buttermilk) in the Dutch cohort Lifeline DEEP [20].

Analysis of the metabolomics data using PLS-DA identified that shifts in lipid metabolism were associated with consumption of fermented plants, since the majority of the top discriminating metabolites appeared to be lipids. Of those that could be identified, CLA was particularly notable. The abundance of the CLA isomer “CLA4”is significantly increased in consumers compared to non-consumers. CLA is known to be produced during ruminal bacterial fermentation and impacts the fatty acid composition of meat and dairy products from ruminants that represent the major dietary sources of CLA in humans [41]. Due to their possible health benefits [42, 43], CLA is also often consumed as a nutritional supplement. However, CLA fecal recovery did not correlate with dietary CLA intake derived mainly from meat, full-fat dairy and egg sources as determined by the food frequency questionnaire (FFQ). Moreover, dietary consumption of total CLA (g/d) did not differ between consumers and non-consumers. Thus, it is possible that CLA is being produced by resident or transient bacteria derived from fermented foods.

Indeed, diet-related bacteria, such as LAB, bifidobacteria and propionibacteria, have been shown to produce CLA [40]. Intestinal bacteria belonging to the families Lachnospiraceae and Ruminococcaceae have also been shown to metabolize LA into products that can be precursors of CLA [38], and two of these Lachnospiraceae were also found to be associated with consumers. The order Lactobacillales includes the largest diversity of previously reported CLA producers, and notably, seven out of the eight species previously identified as associated with fermented foods (Set 3) are CLA-producing Lactobacillus species that we found to be associated with fermented food consumers: *L. acidophilus*, *L. brevis*, *L. fermentum*, *L. helveticus*, *L. paracasei*, *L. plantarum* and *L. sakei* (for reviews, see [39–41, 44]. However, increased CLA in consumers cannot be fully attributed to production by fermented food associated bacteria. For example, some members of the order Clostridiales previously reported to produce CLA in human feces (including four Roseburia species: *R. inulinivorans*, *R. hominis*, *R. intestinalis* and *R. faecis* [38]) were found to be associated with non-consumers, along with *Anaerostipes caccae*, *Eubacterium ventriosum* (L2-12) and *Faecalibacterium prausnitzii* which are also known to metabolize LA.

We detected seven *Bifidobacterium* species previously reported to produce CLA using LA as a precursor [39, 40], including *Bifidobacterium animalis*, *B. longum* [45], and *B. breve*, which has been considered for CLA enrichment in commercial foods such as yogurt due to its CLA-producing ability [46]. Yet none of these were found to be associated with the fermented food consumers. Rather, two other *Bifidobacterium* species not known to produce CLA (*B. aesculapii* and *B. reuteri*) were found to be associated with fermented food consumers, with *B. reuteri* growth actually inhibited at high concentration of LA precursor [47]. Moreover, of the top 50 taxa that were identified as having the highest probability of co-occurring with “CLA4”, only 14 were known CLA producers (Table S4). Future investigation into metabolic pathways in larger datasets would allow the identification of species that explain the higher abundance of “CLA4” in consumers compared to non-consumers.

This is to our knowledge the largest study of the association between fermented food (specifically, fermented plant) consumption and the human gut microbiome, with nearly seven thousand individuals at one time point and over one hundred individuals across four weeks of sampling. We took a multi-omic approach – a combination of 16S rRNA sequencing, shotgun metagenomics, and mass spectrometry – coupled with state-of-the-art tools to evaluate the data. We find that the consumption of fermented plants and more broadly, fermented foods, is associated with quite subtle microbiome variation in healthy individuals. While this explorative study provides the foundation for more directed research, such as randomized placebo-controlled studies, it has some limitations: particularly that consumers were categorized according to self-reported frequency of fermented plant consumption. First, self-reported dietary information can be flawed with measurement errors [48]. Second, although our data suggests fermented plant consumption may be a reasonable proxy for consumption of fermented food more generally, it does not explicitly take into account other food types, such as fermented dairy products. Additionally, this study is mostly limited to participants living in the United States, who may consume a lower diversity of fermented foods than populations living in other countries; expanding this study to a wider range of populations would allow us to capture a greater diversity of fermented food types and associated microbial communities. Due to a combination of these factors, we may be underestimating the potential effects of fermented food consumption on the gut microbiome. Yet notably, the recovery of LAB and fermented food derived microbes in the stool of self-reported consumers suggests that data from stool may be used to help verify the reliability of self-reported dietary information. It would therefore be of great relevance to evaluate not only the associations between specific types of food and the microbiome, but also our ability to detect consumption of specific fermented foods in future studies.

6.6 Acknowledgments

Chapter 6, in full, is a modified reprint of the material as it appears in “B. C. Taylor, F. Lejzerowicz, M. Poirel, J. Schaffer, L. Jiang, A. Aksenov, G. Humphrey, C. Martino, S. Miller-Montgomery, P. Dorrestein, P. Veiga, S. J. Song, D. McDonald, M. Derrien, and R. Knight. *Consumption of Fermented Foods Is Associated with Systematic Differences in the Gut Microbiome and Metabolome*. mSystems (2020). DOI: 10.1128/mSystems.00901-19.” The dissertation author was the primary investigator and author of this work, contributing substantially to experimental design. Sample collection, 16S rRNA extraction, and sequencing/mass spectrometric analysis were performed by others. The dissertation author computationally processed the 16S rRNA sequence data and shotgun metagenomics data; designed the analyses for and interpreted the 16S rRNA sequence data, shotgun metagenomics data, and mass spectrometric data; designed and assembled the figures; performed the statistical analyses; and wrote the manuscript.

We thank Aurélie Cotillard for helpful discussion on statistical analysis. We thank Embriette Hyde and Elaine Wolfe for their help in setting up the study. We also thank James Morton for his valuable input and discussion. Sequence data associated with this study can be found under EBI accession ERP012803. This work was supported in part by Danone Nutricia Research and the Center for Research on Intelligent Storage and Processing in-memory.

6.7 Materials and Methods

6.7.1 Participant recruitment, sample processing, and sample selection.

This research was performed in accordance with the University of Colorado Boulders Institutional Review Board protocol number 12-0582 and the University of California San Diegos Human Research Protection Program protocol number 141853. In order to investigate the effect of fermented plant and food consumption on the gut microbiome, a retrospective analysis was performed on the American Gut Project dataset [25]. An additional cohort of 115 subjects were recruited to explore the effect of fermented food consumption or non-consumption over a period of 4 weeks; the samples from the longitudinal cohort were processed and sequenced in accordance with AGP protocol and integrated into the AGP dataset. The time point with the highest read count from each of the 115 recruited individuals was added to the concurrent cross-sectional assessment. The longitudinal cohort also responded to a specific fermented food questionnaire.

The entire AGP dataset was subset using the metadata version accessed August 8th, 2019 for stool samples from adult participants (age >19 and <70 years) who answered the “fermented plant frequency” question from the AGP questionnaire. Participants were excluded if they took antibiotics in the last year; or if they had outlier values for their body mass index (<15 or >50), height (<48cm or >210cm), or weight (<2.5 kg or >200 kg). If biological replicates were present, the replicate with the lower number of reads was removed (with the exception of the 115 participants that constitute the longitudinal cohort). Based on the AGP questionnaire, participants were considered consumers if they reported “Daily”, “Frequent” and “Occasional” fermented plant consumption (i.e. >1-2 times per week), and as non-consumers if they reported “Rarely” and “Never”.

6.7.2 Diet quality and intake assessment.

Overall diet quality was assessed by the Healthy Eating Index 2010 (HEI-2010) as described elsewhere [49]. Briefly, the HEI-2010 is a valid, reliable measure of diet quality that assesses how an individuals diet pattern adheres to the 2010-2015 Dietary Guidelines for Americans (DGA). HEI-2010 includes 12 dietary components, nine of which are classified as adequacy components that should be included regularly in the diet (total fruit, whole fruit, total vegetables, greens and beans, whole grains, dairy, total protein foods, seafood and plant proteins, fatty acids), and 3 "moderation" components (refined grains, sodium, empty calories) that should be limited in the diet. Individual dietary components are scored from 0 to 5, 10 or 20 points with maximum points indicating higher consumption of adequacy components and lower consumption of moderation components. Total HEI-2010 scores (range: 0-100) were calculated as the sum of the 12 components with a higher total score indicating better/optimal diet quality and greater adherence to the DGA. HEI-2010 scores, as well as total energy, carbohydrate, fat, protein and fiber intake were calculated from individuals in the AGP cohort who completed the VioScreen food frequency questionnaire (FFQ). We compared the total HEI score and mean nutrient intakes between consumers and non-consumers using the Mann-Whitney U test.

Daily total consumption of CLA and LA (g/d) was estimated from the VioScreen FFQ reports. Total CLA consumption was deduced from the following food sources: beef and other meat such as fish and turkey, full-fat dairy products (e.g., milk, butter, cheese, yogurt), and eggs. Total LA consumption was obtained from the following reported foods: vegetable oil (e.g. canola and olive), salad dressings containing vegetable oils, butter, eggs, meat (beef, chicken, turkey, pork), potatoes (e.g., French Fries/fried white potatoes, potato chips), nuts, nut butters and seeds, mixed Mexican dishes and meat dishes such as stews and casseroles.

6.7.3 16S rRNA gene sequencing.

DNA extraction and 16S rRNA amplicon sequencing were done using Earth Microbiome Project (EMP) standard protocols (http://www.earthmicrobiome.org/protocols-and-standards/16s). DNA was extracted with the Qiagen MagAttract PowerSoil DNA kit as previously described [50]. Amplicon PCR was performed on the V4 region of the 16S rRNA gene using the primer pair 515f to 806r with Golay error-correcting barcodes on the reverse primer. Amplicons were barcoded and pooled in equal concentrations for sequencing. The amplicon pool was purified with the MO BIO UltraClean PCR cleanup kit and sequenced on the Illumina MiSeq sequencing platform. Based on the filtering noted above, a feature table representing the 16S V4 rRNA gene sequence data was obtained from Qiita [51] using redbiom [52] from the Deblur-Illumina-16S-V4-150nt-780653 context. This table was composed of 8,513 samples. Prior to extraction from Qiita, the AGP data had been trimmed to 150 bases, and processed using Deblur v1.0.4 [53] using the Qiita default parameters (i.e., setting –min-reads 1) to generate sOTUs. Technical replicates of samples were excluded as to only keep the most sequenced version of each sample. After previously recognized bloom sequences were removed [54], samples with fewer than 1500 reads were omitted. Taxonomies for sOTUs were assigned using the sklearn-based taxonomy classifier trained on the GreenGenes reference database 13_8 [55] clustered at 99% similarity (feature classifier plug-in of QIIME 2 v2019.1 [56]. The sOTU table was rarefied to a depth of 1,500 sequences/sample to control for sequencing effort [57] and sOTUs totaling 5 reads across samples. The deblurred sequence fragments were inserted into the Greengenes 13_8 phylogenetic tree using SATé-enabled phylogenetic placement [58, 59].

6.7.4 16S marker gene data analysis.

QIIME 2 v2019.1 [56] was used to generate pairwise unweighted and weighted UniFrac distances [52, 60]. Between group differences based on these distances were tested using PERMANOVA [61] and permuted t-tests in QIIME 2. Alpha diversity (Faith's PD [31], Shannon diversity [30], and observed OTU richness) between consumers and non-consumers (as a whole and when stratified by consumption frequency) was generated using QIIME 2 [56] and compared with a Kruskal-Wallis test. Wilcoxon signed-rank [62] and Mann-Whitney U tests were respectively used to assess alpha diversity between successive time points within consumers and non-consumers, and within time point between consumers and non-consumers in the longitudinal cohort. Songbird v1.0.1 [27] was used to identify feature ranks corresponding to consumers and non-consumers (parameters: –epochs 5000 –batch-size 5 –learning-rate 1e-4 –min-sample-count 1000 –min-feature-count 0 –num-random-test-examples 10) and Qurro v0.4.0[28] was used to compute log-ratios of these ranked features. T-tests and Cohens D were calculated to assess the significance (alpha=0.05) and effect size of the log-ratios. The stability of the participants microbiomes was assessed by comparing samples log-ratios in consecutive timepoints, for both the 16S and metabolomic datasets. The 40 highest and lowest ranked features were used in order to compute enough log-ratios for Spearmans rank correlation coefficients across all samples, and for Ordinary Least Squares regression (FIG S5).

6.7.5 LC-MS/MS data acquisition.

The untargeted metabolomics analysis using high-performance LC-MS/MS (HPLC-MS) was carried out as described previously [25]. The chromatography was performed on a Dionex UltiMate 3000 Thermo Fisher Scientific high-performance liquid chromatography system (Thermo Fisher Scientific, Waltham, MA) coupled to a Bruker Impact HD quadrupole time of flight (qTOF) mass spectrometer. The chromatographic separation was carried out on a reverse phase (RP) Kinetex C18 1.7-ţm, 100-Å ultrahigh-performance liquid chromatography (UHPLC) column (50 mm by 2.1 mm) (Phenomenex, Torrance, CA), held at 40řC during analysis. A total of 5 ţl of each sample was injected. Mobile phase A was water, and mobile phase B was acetonitrile, both with added 0.1% (vol/vol) formic acid. The solvent gradient table was set as follows: initial mobile phase composition was 5% B for 1 min, increased to 40% B over 1 min and then to 100% B over 6 min, held at 100% B for 1 min, and decreased back to 5% B in 0.1 min, followed by a washout cycle and equilibration for a total analysis time of 13 min. The scanned m/z range was 80 to 2,000, the capillary voltage was 4,500 V, the nebulizer gas pressure was 2 Œ 105 Pa, the drying gas flow rate was 9 liters/min, and the temperature was 200řC. Each full MS scan was followed by tandem MS (MS/MS) using collision-induced dissociation (CID) fragmentation of the seven most abundant ions in the spectrum. For MS/MS, the collision cell collision energy was set at 3 eV and the collision energy was stepped 50%, 75%, 150%, and 200% to obtain optimal fragmentation for differentially sized ions. The scan rate was 3 Hz. An HP-921 lock mass compound was infused during the analysis to carry out postprocessing mass correction. All of the raw data are publicly available at the UCSD Center for Computational Mass Spectrometry (https://massive.ucsd.edu/, the data set ID MassIVE MSV000081171). To determine the specific isomer of the annotations for octadecadienoic acid isomers, authentic standards for linoleic acid (LA; Spectrum Laboratory Products, Inc., USA) and conjugated linoleic acid (CLA; mixture of 4 isomers:

9,11 and 10,12 isomers, E and Z) (Sigma-Aldrich, USA) were compared by retention times (RTs) and MS/MS spectra. This brings these annotations to the level 1 identifications (authentic compound was analyzed under identical experimental conditions with orthogonal physical property compared).

6.7.6 LC-MS/MS data analysis.

The collected data were processed as described in [63]. Briefly, the feature tables were obtained using MZmine2 [64]. The collected HPLC-MS raw data were converted from Bruker's .d to .mzXML format. The data were then batch-processed with the following settings for each step:

Mass detection:

- Noise level 1000
- Chromatogram builder
- Minimum time span 0.01 min
- Minimum peak height 3000
- m/z tolerance 0.1 m/z or 20 ppm

Chromatogram deconvolution - Baseline cutoff

- Minimum peak height 3000
- Peak duration range (min) 0.01 - 3.00
- Baseline level 300

Deisotopisation - Isotopic peak grouper

- m/z tolerance 0.1 m/z or 20 ppm
- RT tolerance 0.1 min
- Maximum charge 4

Peak alignment - Join aligner

- m/z tolerance 0.1 m/z or 20 ppm
- Weight for m/z 75
- Weight for RT 25
- RT tolerance (min) 0.1

Peak filtering - Peak list raw filter

- Minimum peak in a row 3
- Minimum peak in an isotope pattern 2

The metadata were added into the resulting extracted feature table and used as input for the MetaboAnalyst software [65, 66]. The feature tables were filtered with interquantile ranges to remove outliers, the data imputed, normalized by the quantile normalization, and autoscaled (mean centring and dividing by the standard deviation for each feature). Partial least squares discriminant analysis (PLS-DA) was used to explore and visualize variance within data and differences among experimental categories. The CLA and LA metabolite features were identified manually based on GNPS [67] and MZmine2 [64] processing pipelines (see link to feature based molecular networking). Wilcox rank sums test (Mann-Whitney U test) was used to assess the significance of difference between the consumers and non-consumers for the levels of identified CLA and LA metabolites (alpha=0.05).

The annotations and visualizations of chemical distributions were explored on GNPS using molecular networking [67] as follows. MS/MS spectra were window filtered by choosing only the top 6 peaks in the 50-Da window throughout the spectrum. The MS spectra were then clustered with a parent mass tolerance of 0.02 Da and an MS/MS fragment ion tolerance of 0.02; consensus spectra that contained fewer than 4 spectra were discarded. Network was created where with edges filtered to have a cosine score above 0.65 and more than 5 matched peaks. The edges between two nodes are kept in the network if and only if each of the nodes appeared in each other's respective top 10 most similar nodes. The required library matches were set to have a score above 0.7 and at least 6 matched peaks when searched the spectra in the network against GNPS spectral libraries. All resulting annotations are at level 2/3 according to the proposed minimum standards in metabolomics [68]. The GNPS results are located at https://gnps.ucsd.edu/ProteoSAFe/ status.jsp?task=420a545b5b164d10a20f62c0ec0ce7e7. Feature-based molecular network [69] results can be found at https://gnps.ucsd.edu/ProteoSAFe/ status.jsp?task=9ce1517e83a94d9a8cd9d79f3e16eea0. The CLA and LA metabolite features were initially identified based on GNPS library search [67] and then their annotation further confirmed via use of authentic standards. Wilcoxon rank sums test was used to assess the significance of difference between the consumers and non-consumers for the levels of identified CLA and LA metabolites (alpha=0.05).

6.7.7 Metagenomic sequencing.

Extracted DNA was quantified with PicoGreen dsDNA Assay Kit, and 5 nanograms of input, or maximum 3.5 microliters, gDNA was used in a 1:10 miniaturized Kapa HyperPlus protocol. Per sample libraries were quantified and pooled at equal nanomolar concentration. The pooled library was cleaned with the QIAquick PCR Purification Kit and size selected for fragments between 300 and 700 bp on the Sage Science PippinHT. The pooled library was sequenced as a paired-end 150-cycle run on an Illumina HiSeq2500 v2 in Rapid Run mode at the UCSD IGM Genomics Center, with a target depth of c.a. 20 million reads per sample. The sequencing adapter and short reads were first removed using Atropos v1.1.21 (-q 15 –minimum-length 100 –pair-filter any) as well as reads aligning to the human genome using bowtie2 (–very-sensitive). The pass-filter reads were then concatenated per sample, excluding 1 biological duplicate and 8 samples from participants exposed to antibiotics, in order to obtain 91 pairs of fastq files.

6.7.8 Metagenomic data analysis.

On each separate sample fastq files, paired-end reads were merged using FLASH v1.2.11 [70], and then processed for taxonomic profiling using SHOGUN v1.0.6 [71] with bowtie2 v2.3.4.3 [72] to align reads to the 85,626 prokaryotic genomes covering 12,977 species from the NCBI RefSeq database release 82 [73]. The read counts for the genome features identified in each sample were merged into one genome-per-sample table that was then filtered to keep genomes with a per-sample relative mapped read abundance of at least 0.01%. The features labeled at the sub-species level were sum-collapsed at the species level; taxonomy was used as a proxy for a phylogeny. As with the 16S cross-sectional data, Songbird (Songbird v1.0.1, [27]) was used for regression modelling on our binary fermented consumption variable to identify features associated with consumption and non-consumption (parameters as above). Qurro v0.4.0[28] was used to compute log-ratios of these ranked features. T-tests and Cohens D were calculated to assess the significance (alpha=0.05) and effect size of the log-ratios.

6.7.9 Multi-omics data analysis.

In order to identify microbial features associated with fermented food consumption and the metabolites they might be producing, we measured probabilities of co-occurrence between observed species (based on metagenomic data) and either all metabolites, or a set of five linoleic and isomers of conjugated linoleic acids discernable in the data (as informed by the metabolomic analysis). For this analysis, we used mmvec v1.0.2 [37], a neural network solution inspired from natural language processing to build a log-transformed conditional probability matrix from each cross-omics features pairs and apply singular value decomposition in order to represent co-occurrence in the form of biplots. We chose the model where accuracy was highest for different initialization conditions for the gradient descent algorithm (–batch-size of 1000, 2500 and 5000 and –learn-rate of 1e-4 and 1e-5), with low cross-validation error and model likelihood. To evaluate the fitness of the mmvec microbe-metabolite interactions, we compared the latent representation to the observed Songbird differentials. The relationship between the microbial first principal component learned from mmvec and the log fold change of the microbes between fermented food consumption was significantly negatively correlated (Pearsons r=-0.651, P=4.63e-22, n=249 microbes; FIG S2C), suggesting that the mmvec microbe-metabolite relationship to fermented food consumption is a valid comparison. We used EMPeror v2019.1.0 [74] to visualize features-features biplots along with overlying genomes differential abundance ranks for our fermented food consumption model.

6.7.10 Data availability.

The data generated in this study are available publicly in Qiita under the study ID 10317. Sequence data associated with this study can be found under EBI accession ERP012803. The metabolomics analysis is available at: https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=420a545b5b164d10a20f62c0ec0ce7e7 (classical molecular networking) and: https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=9ce1517e83a94d

9a8cd9d79f3e16eea0 (feature-based molecular networking). The raw experimental data are available at MassIVE (https://massive.ucsd.edu/), dataset MSV000081171.

6.8 Supplemental Material

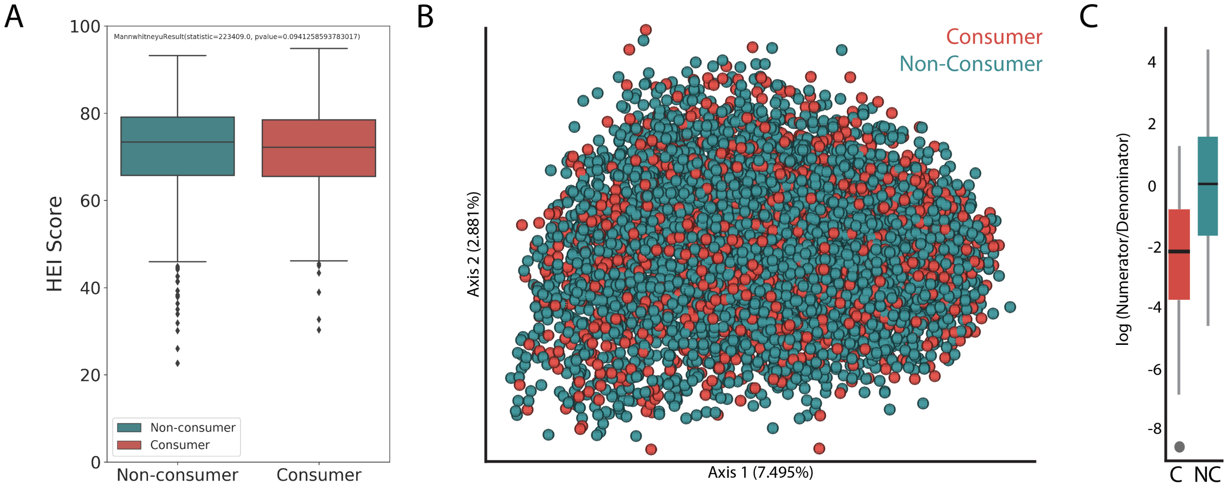


Figure 6.4: (A) Healthy Eating Index-2010 scores of consumers (N=483) and non-consumers (N=966). (B) Taxonomic differences (16S) between Consumers and Non-Consumers in the Cross-sectional Cohort. Unweighted UniFrac PCoA colored by consumers (red) and non-consumers (teal). Beta diversity (PERMANOVA): Unweighted UniFrac, pseudo-F test statistic = 3.677, p = 0.001; Weighted UniFrac: pseudo-F test statistic = 2.163, p = 0.058. Alpha diversity (Kruskal Wallis): Faith PD: H = 4908194, p-value = 0.9565; Shannon: H = 4947661, p-value = 0.6356; Richness: H = 4831862, p-value = 0.793 (C) Differentially abundant taxa associated with Consumers ("Set 1 Microbes", Table S3) and Non-Consumers ("Set 2 Microbes", Table S3) identified by Songbird. The log-ratio of [Set 1] / [Set 2] is significantly different (t-test p=0.00065, t = 3.6367, df = 50.76, Cohens D = 0.9998).

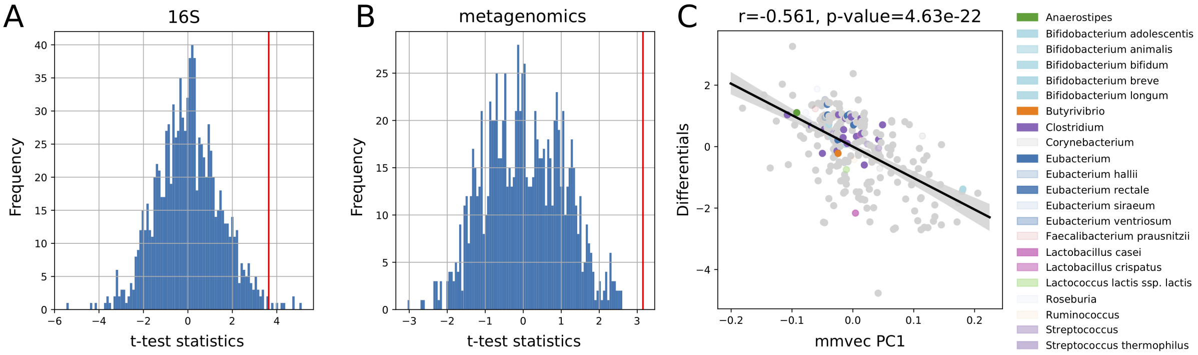


Figure 6.5: Validation of log-ratio usage for differentially-abundant microbe selections (Songbird) and microbe-metabolite co-occurrence interpretation (mmvec). (A-B) Distribution of the t-test statistics values obtained for the comparison of consumer vs. non-consumer based on the log-ratios of each of 1000 random microbe selections. One caveat to this type of permutation is the dependence caused by the re-use of features in the generation of the null distribution, with only 3100 and 153 features employed for the 16S and metagenomic datasets, respectively. (C) Relationship between the location of the metagenomic microbial features along the first axis of the mmvec biplot (see 6.3C) and their differential abundance ranks (low values = most associated with consumers). Pearson's correlation r score and p-value are indicated for 249 microbes, with CLA-producers coloured per taxon.

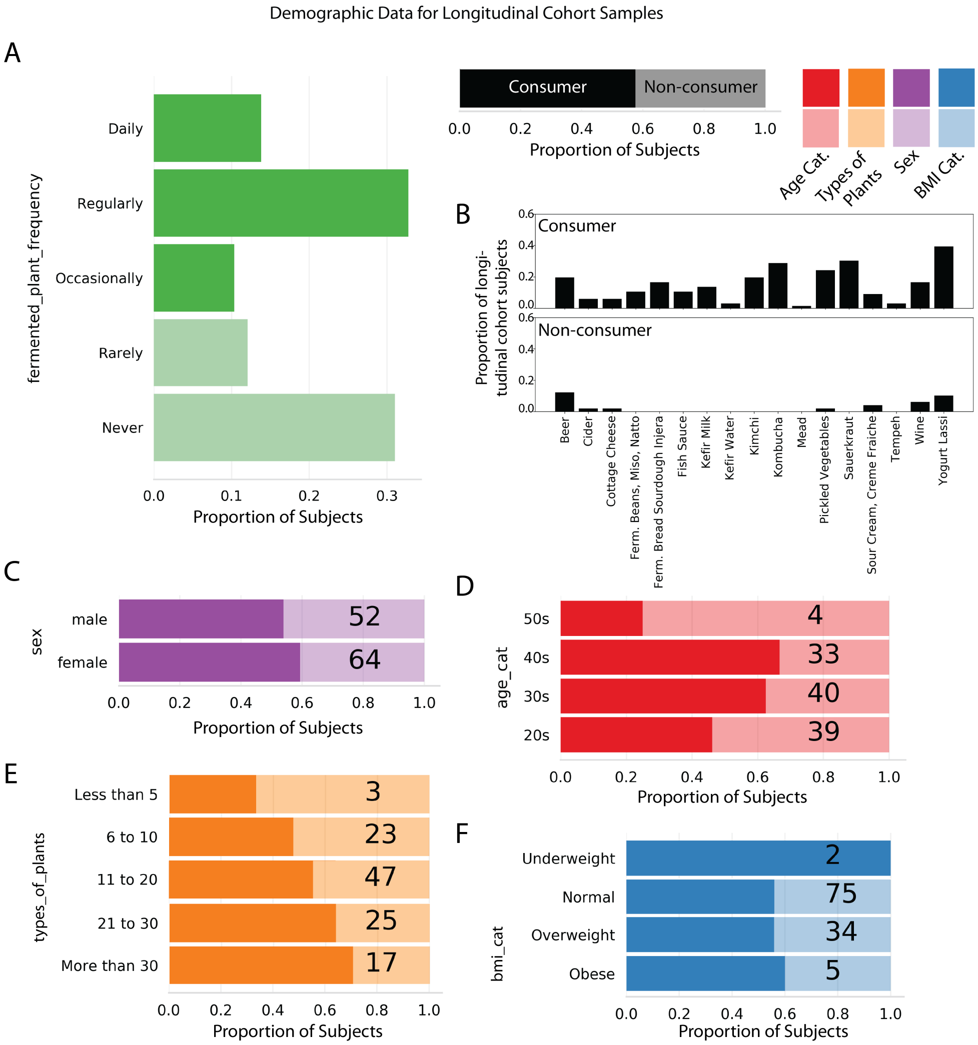


Figure 6.6: Demographic and selected dietary data for longitudinal cohort samples. (A) Fermented plant consumption frequency within the longitudinal cohort subjects. (B) The other types of fermented food that consumers and non-consumers reported eating, aside from fermented plants. (C-F) Proportion of subjects in the longitudinal cohort between consumers and non consumers, described by sex (chi2 = 1.31, p-value = 2.53e-01), age category (chi2 = 23.85, p-value = 2.68e-05), number of types of plants consumed (chi2 = 19.61, p-value = 5.97e-04), and BMI category (chi2 = .22, p-value = 8.97e-01; underweight subjects were removed in this comparison because there are no underweight non-consumers in the longitudinal cohort). The total number of individuals is displayed on each bar.

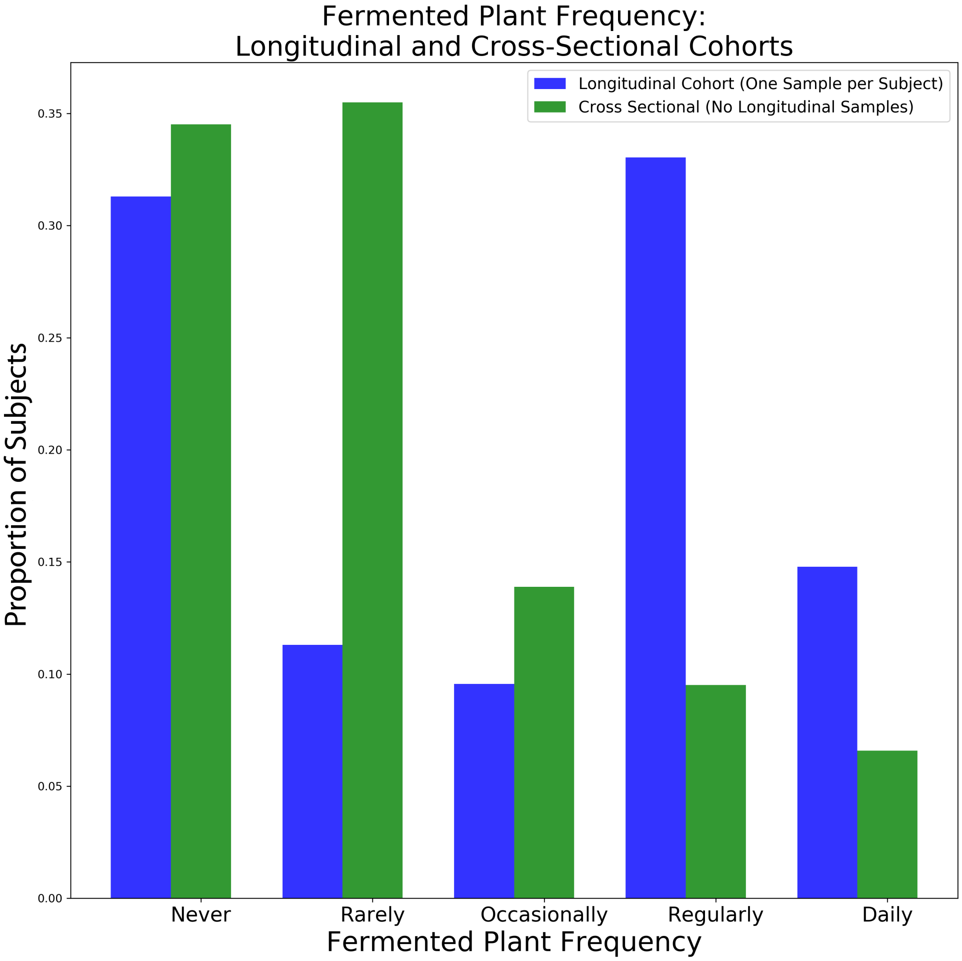


Figure 6.7: Proportion of fermented plant consumption in the (single time point) longitudinal and cross-sectional cohorts. Proportion of subjects within the longitudinal vs. the cross-sectional cohorts who consume fermented plants at frequencies of Never, Rarely, Occasionally, Regularly, and Daily.

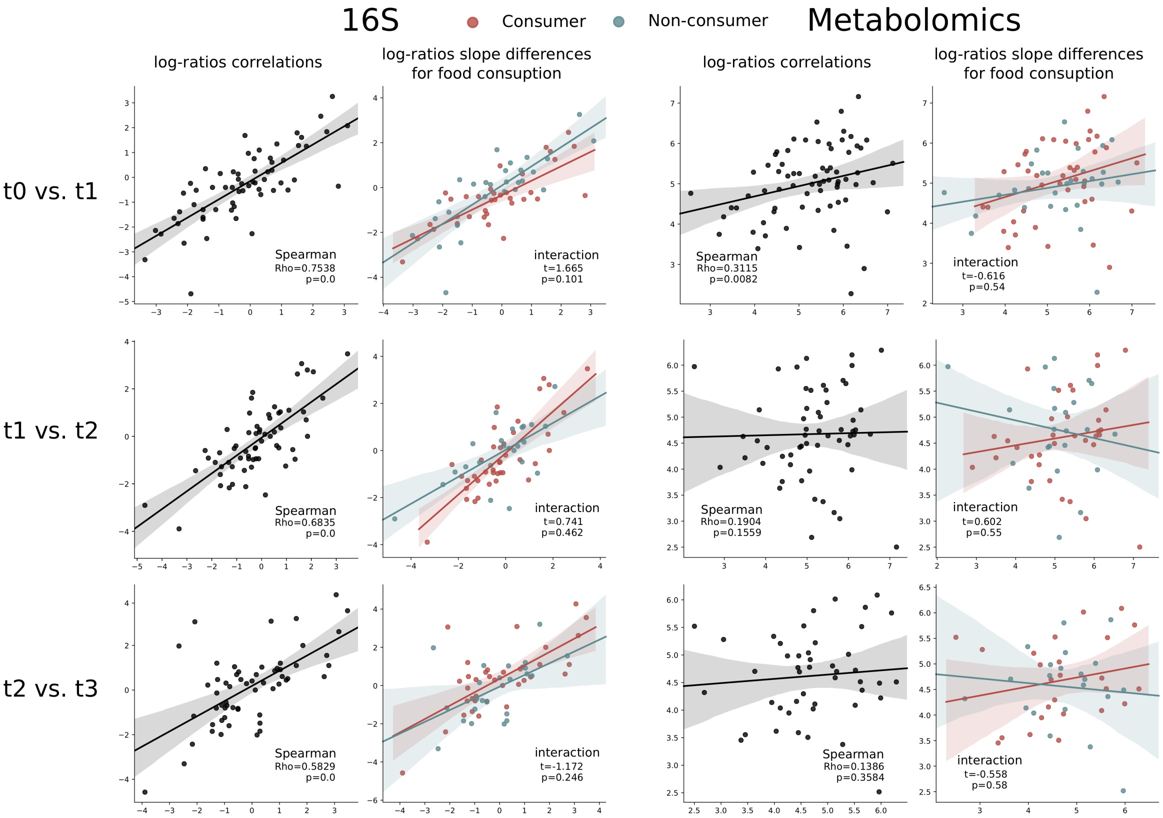


Figure 6.8: Longitudinal stability of microbes and metabolites (correlations were assessed by Spearman rho). The first and third columns are based on the Songbird 16S results and metabolomics results, respectively, for all longitudinal samples. The second and fourth columns consider consumers and non-consumers separately. First column: The log-ratios of the sets of taxa identified using Songbird on the 16S data between timepoints. 16S signatures, within participants, and within datatype, appear consistent over time based on Spearmans rho. The second column plots these log-ratios for consumers and non-consumers separately. Third column: The log-ratios of the sets of annotated mass spectrometry features identified using Songbird between timepoints. The fourth column plots these log-ratios for consumers and non-consumers separately.

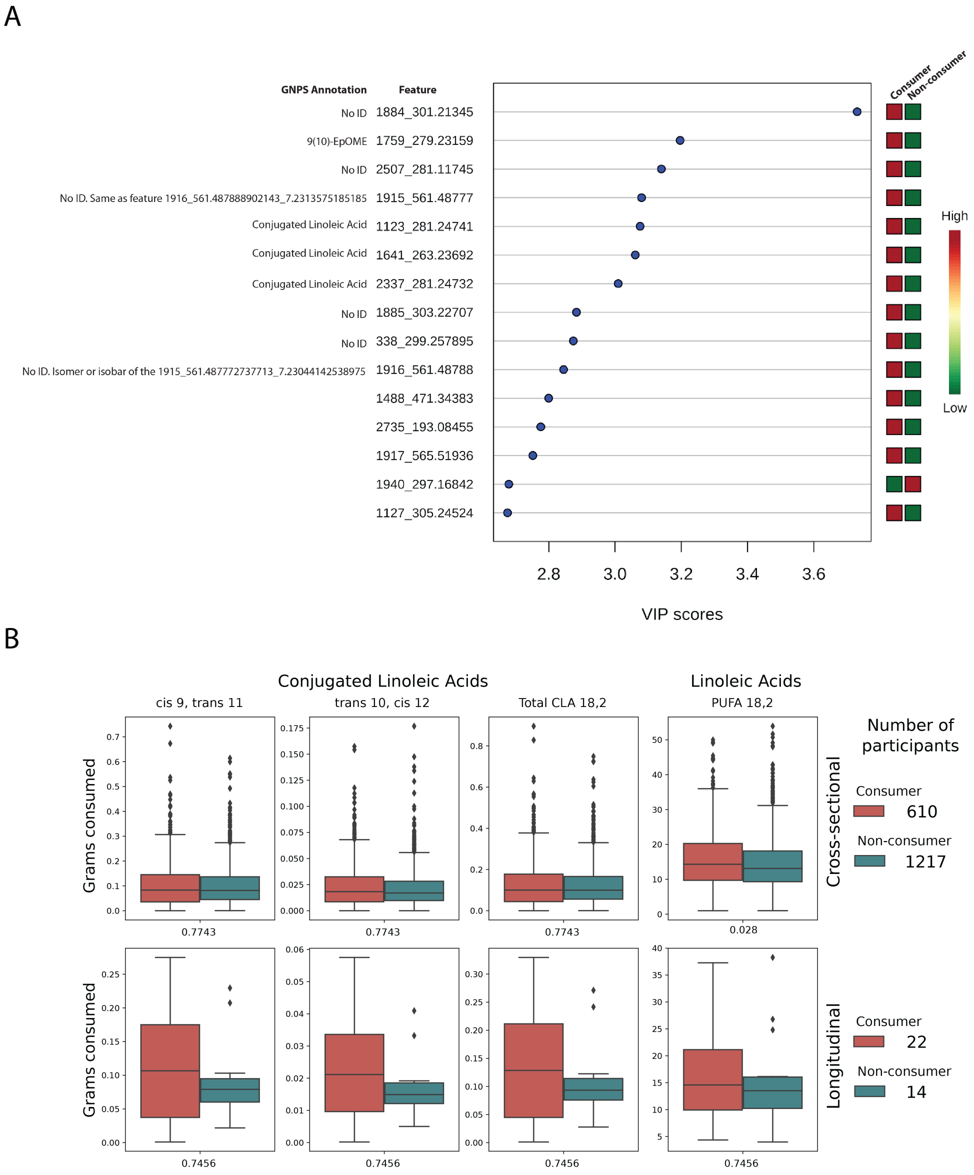


Figure 6.9: (A) Variable Importance in the Projection (VIP) scores of the top 15 features discriminating between consumers and non-consumers. Abundance across categories is represented by color marking on the right (green is low, red is high). The values under "feature" refer to the detected m/z. The putative annotations from GNPS library search of the top ten features discriminating Consumer and Non-consumer categories are included next to the features. (B) CLA and LA consumption from the Food Frequency Questionnaire does not differ between consumers and non-consumers. Comparison of dietary CLA and LA in consumers and non-consumers (Wilcoxon rank sums test with Bonferroni-Hochberg correction for multiple testing) for both cross-sectional and longitudinal datasets. Sample sizes indicated on the right-hand side. Test p-values are indicated under each panel.

Table 6.1: Estimated daily nutrient intake for fermented plant consumers and non-consumers. Data are presented as group mean ± SD; (min, max). P-values were obtained via Mann-Whitney U test. Abbreviations: g, gram; d, day; NS, non-significant; kcal, kilocalories.

	Consumer (n = 147)	Non-consumer (n = 163)	p-value
Total energy, kcal/d	1845 ± 615 (605, 3634)	1855 ± 728 (467, 5672)	NS
Total Carbohydrate, *g/d*	173 ± 78 (33, 406)	197 ± 93 (8, 552)	0.012
Total Carbohydrate, % of energy	39 ± 15 (9, 94)	43 ± 13 (1.4, 81)	<0.001
Total Fat, *g/d*	90 ± 44 (5, 239)	80 ± 40 (12, 266)	0.014
Total Fat, % of energy	42 ± 7 (7, 63)	37 ± 5 (23, 48)	<0.001
Total Protein, *g/d*	79 ± 34 (15, 195)	77 ± 45 (18, 457)	NS
Total Protein, % of energy	17 ± 2 (10, 22)	16 ± 2 (15, 32)	<0.001
Total Dietary Fiber, *g/d*	27 ± 12	26 ± 11 (0, 59)	NS

Table 6.2: Alpha diversity comparisons (Kruskal-Wallis) between different frequencies of fermented plant consumption.

Consumer	Non-consumer	Alpha Diversity Metric	H-statistic	p-value
Daily	Rarely	Faith’s PD Shannon ObservedOTUs	1.2998 1.4434 1.328	0.2543 0.2296 0.2492
Daily	Never	Faith’s PD Shannon ObservedOTUs	0.3275 0.4501 2.1045	0.5671 0.5023 0.1469
Regularly	Never	Faith’s PD Shannon ObservedOTUs	1.4185 0.3848 0.0451	0.2336 0.535 0.8317
Occasionally	Never	Faith’s PD Shannon ObservedOTUs	0.2723 0.0 3.1067	0.6018 0.9975 0.078

Table 6.3: Sets of taxa used in Songbird analyses.

Sets	Taxa
Set 1	Lachnospiraceae, *Coprococcus eutacus*, *Cronobacter sakazakii*, *Enterobacteriaceae*, *Pseudomonas stutzeri*, Cyanobacteria order YS2, *Prevotella* spp., *Prevotella copri*, Alphaproteobacteria order RF32, Fusobacteriaceae, *Bacteroides coprophilus*, *Bacteroides* spp., Bacteroidales family S24-7, *Coprococcus* spp.
Set 2	*Bacteroides* spp., *Pseudomonas* spp., *Dorea* spp., Lachnospiraceae, *Prevotella* spp., *Alistipes putredinis*, *Oscillospira* spp., Enterobacteriaceae, *Fusobacterium* spp., *Actinomyces* spp., *Achromobacter* spp., *Clostridium clostridioforme*, *Faecalibacterium prausnitzii*, *Bacteroides uniformis*, Clostridiales, *Delftia* spp.
Set 3	*Lactobacillus acidophilus*, *Lactobacillus brevis*, *Lactobacillus fermentum*, *Lactococcus lactis*, *Leuconostoc mesenteroides*, *Lactobacillus paracasei*, *Lactobacillus plantarum*, and *Lactobacillus rhamnosus*.
Set 4	*Blautia obeum*, *[Eubacterium] hallii*, *Ruminococcus faecis*, *Blautia sp._KLE_1732*, *Clostridium sp.L2-50, Faecalibacterium prausnitzii, Blautia obeum, Roseburia inulinivorans, Roseburia intestinalis, Bacteroides thetaiotaomicron, [Eubacterium] rectale, Bacteroides ovatus, Bacteroides plebeius, Alistipes putredinis, Bacteroides uniformis, Bacteroides vulgatus, Escherichia coli*.
Set 5	*Streptococcus dysgalactiae, Lachnospiraceae bacterium oral taxon 500, Prevotella melaninogenica, Megamonas hypermegale, Proteus mirabilis, Actinomyces sp. oral taxon 448, Leclercia adecarboxylata, Comamonas kerstersii, Pontibacillus chungwhensis, Citrobacter freundii, Enorma massiliensis, Rhizobium sp. Root651, Enterococcus cecorum, Pseudomonas xanthomarina, Peptoniphilus sp. ChDC B134, Clostridioides difficile, Prevotella multiformis, Lactobacillus kefiranofaciens, Neisseria mucosa, Bifidobacterium psychraerophilum, Lactobacillus brevis, Eubacterium sp. AB3007, Lactobacillus parabuchneri, Bifidobacterium minimum, Prevotella intermedia, Jonquetella anthropi, Lachnospiraceae bacterium M18-1, Lactobacillus acidophilus, Clostridium perfringens, [Clostridium] sporosphaeroides, Streptococcus infantis, Lactobacillus acetotolerans, Bacteroides paurosaccharolyticus, Corynebacterium lactis, Streptococcus mitis, Eggerthella sp. YY7918, Lactobacillus helveticus, Kandleria vitulina, Serratia liquefaciens, Lactobacillus sakei*.
Set 6	*Bacteroides coprophilus, Bacteroides ovatus, Sutterella wadsworthensis, Prevotella bivia, Prevotella copri, Acidaminococcus intestini, Coprobacter fastidiosus, Bacteroides cellulosilyticus, Prevotella stercorea, [Ruminococcus] torques, Dorea longicatena, Sutterella wadsworthensis, Flavonifractor plautii, Finegoldia magna, [Eubacterium] rectale, Roseburia intestinalis, Bacteroides vulgatus, Porphyromonas crevioricanis, Blautia sp. KLE 1732, Escherichia coli, Blautia obeum, Roseburia inulinivorans, [Clostridium] lactatifermentans, Faecalicatena contorta, Alistipes finegoldii, [Clostridium] bolteae, Collinsella intestinalis, Campylobacter hominis, Coprococcus comes, Coprobacter secundus, Eubacterium ventriosum, Bacteroides plebeius, Methanobrevibacter smithii, Parabacteroides distasonis, Tyzzerella nexilis, Bacteroides salanitronis, Clostridium sp. KLE 1755, Collinsella aerofaciens, Kluyvera intermedia, Staphylococcus aureus*.

Table 6.4: mmvec: species and CLA4 co-occurrence. Species (rep82 features) are listed in the column "Feature" and are sorted by "mmvecRank", the ranked conditional probabilities of co-occurrence between rep82 features and "CLA4". Species known to produce CLA appear in column "CLA-producer". The column "mmvecPC1" contains the associated taxa features and their PC1 axis values from Figure 3C. The last two columns display the numbers of consumers and non-consumers who have the feature.

CLA-producer	Feature	mmvec PC1	mmvec Rank	#Cons.	#Non-cons.
other	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__Bacteroides_vulgatus	-0.057	6.223	43	36
other	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__Bacteroides_uniformis	-0.031	5.617	43	36
other	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacterales; f__Enterobacteriaceae; g__Escherichia; s__Escherichia_coli	-0.146	5.265	43	36
other	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Rikenellaceae; g__Alistipes; s__Alistipes_putredinis	-0.181	4.931	43	36
other	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Prevotellaceae; g__Prevotella; s__Prevotella_copri	-0.099	4.777	24	25
other	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__Bacteroides_ovatus	-0.050	4.643	43	36
other	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__Bacteroides_stercoris	-0.042	4.526	42	36
Eubacterium rectale	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__; s__[Eubacterium]_rectale	-0.042	4.264	43	36
other	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__Bacteroides_thetaiotaomicron	-0.004	4.244	43	36
other	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Rikenellaceae; g__Alistipes; s__Alistipes_finegoldii	-0.099	3.744	43	35
other	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__Bacteroides_plebeius	-0.046	3.687	43	36
other	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__Blautia; s__Blautia_obeum	-0.137	3.640	43	36
other	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__Bacteroides_coprocola	-0.068	3.616	41	36
Roseburia	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__Roseburia; s__Roseburia_intestinalis	-0.059	3.611	43	36
other	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Rikenellaceae; g__Alistipes; s__Alistipes_shahii	-0.065	3.579	42	35
other	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Tannerellaceae; g__Parabacteroides; s__Parabacteroides_distasonis	-0.050	3.573	42	35
Faecalibacterium prausnitzii	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae; g__Faecalibacterium; s__Faecalibacterium_prausnitzii	-0.062	3.569	43	36
other	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__Bacteroides_cellulosilyticus	-0.027	3.499	43	36
other	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Barnesiellaceae; g__Barnesiella; s__Barnesiella_intestinihominis	-0.028	3.462	35	32
Roseburia	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__Roseburia; s__Roseburia_inulinivorans	-0.047	3.419	43	36
Anaerostipes	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__Anaerostipes; s__Anaerostipes_hadrus	-0.092	3.398	42	35
Roseburia	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__Roseburia; s__Roseburia_faecis	-0.074	3.389	43	35
other	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Odoribacteraceae; g__Odoribacter; s__Odoribacter_splanchnicus	0.004	3.329	42	36
other	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__Bacteroides_fragilis	0.009	3.289	42	33
Ruminococcus	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae; g__Ruminococcus; s__Ruminococcus_bicirculans	-0.047	3.192	37	30
other	k__Bacteria; p__Verrucomicrobia; c__Verrucomicrobiae; o__Verrucomicrobiales; f__Akkermansiaceae; g__Akkermansia; s__Akkermansia_muciniphila	0.045	3.162	26	12
other	k__Bacteria; p__Proteobacteria; c__Betaproteobacteria; o__Burkholderiales; f__Sutterellaceae; g__Sutterella; s__Sutterella_wadsworthensis	-0.116	3.145	17	29
other	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Tannerellaceae; g__Parabacteroides; s__Parabacteroides_johnsonii	-0.086	2.896	41	36

Table 6.5: (Continued) mmvec: species and CLA4 co-occurrence.

CLA-producer	Feature	mmvec PC1	mmvec Rank	#Cons.	#Non-cons.
Eubacterium siraeum	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae; g__Ruminiclostridium; s__[Eubacterium]_siraeum	0.001	2.881	22	16
other	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Oscillospiraceae; g__Oscillibacter; s__Oscillibacter_sp._ER4	0.018	2.857	43	35
other	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__Bacteroides_salyersiae	-0.025	2.841	41	35
Eubacterium	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Eubacteriaceae; g__Eubacterium; s__[Eubacterium]_eligens	-0.007	2.816	38	32
other	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Rikenellaceae; g__Alistipes; s__Alistipes_obesi	-0.043	2.805	39	35
other	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__Blautia; s__Blautia_sp._KLE_1732	-0.066	2.748	43	36
other	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__; s__Lachnospiraceae_bacterium_TF01-11	-0.021	2.738	37	34
Roseburia	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__Roseburia; s__Roseburia_hominis	-0.077	2.589	43	35
other	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Oscillospiraceae; g__Oscillibacter; s__Oscillibacter_sp._KLE_1745	-0.042	2.566	42	35
Ruminococcus	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae; g__Ruminococcus; s__Ruminococcus_faecis	-0.028	2.554	43	36
Eubacterium hallii	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Eubacteriaceae; g__Eubacterium; s__[Eubacterium]_hallii	-0.046	2.539	43	36
other	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__Coprococcus; s__Coprococcus_comes	-0.051	2.419	43	35
other	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Prevotellaceae; g__Prevotella; s__Prevotella_sp._109	0.001	2.404	41	35
other	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__Dorea; s__Dorea_longicatena	-0.038	2.403	41	35
other	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__Tyzzerella; s__Tyzzerella_nexilis	-0.058	2.388	43	35
Bifidobacterium adolescentis	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae; g__Bifidobacterium; s__Bifidobacterium_adolescentis	-0.035	2.342	30	25
other	k__Bacteria; p__; c__; o__; f__; g__; s__bacterium_LF-3	-0.045	2.281	39	32
other	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__Bacteroides_coprophilus	-0.011	2.169	42	35
other	k__Bacteria; p__Proteobacteria; c__Betaproteobacteria; o__Burkholderiales; f__Sutterellaceae; g__Parasutterella; s__Parasutterella_excrementihominis	-0.029	2.130	33	28
other	k__Bacteria; p__Firmicutes; c__Negativicutes; o__Veillonellales; f__Veillonellaceae; g__Dialister; s__Dialister_invisus	-0.034	2.114	14	12
Clostridium	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Clostridiaceae; g__Clostridium; s__Clostridium_sp._L2-50	-0.108	2.113	43	36
Eubacterium	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Eubacteriaceae; g__Eubacterium; s__Eubacterium_ramulus	-0.041	1.975	43	34

References

(1) Bourdichon, F.; Casaregola, S.; Farrokh, C.; Frisvad, J. C.; Gerds, M. L.; Hammes, W. P.; Harnett, J.; Huys, G.; Laulund, S.; Ouwehand, A.; Powell, I. B.; Prajapati, J. B.; Seto, Y.; Ter Schure, E.; Van Boven, A.; Vankerckhoven, V.; Zgoda, A.; Tuijtelaars, S.; Hansen, E. B. *International Journal of Food Microbiology* **2012**, DOI: 10.1016/j.ijfoodmicro.2011.12.030.

(2) Tamang, J. P.; Cotter, P. D.; Endo, A.; Han, N. S.; Kort, R.; Liu, S. Q.; Mayo, B.; Westerik, N.; Hutkins, R. *Comprehensive Reviews in Food Science and Food Safety* **2020**, DOI: 10.1111/1541-4337.12520.

(3) Tamang, J. P.; Watanabe, K.; Holzapfel, W. H. *Frontiers in Microbiology* **2016**, DOI: 10.3389/fmicb.2016.00377.

(4) Rezac, S.; Kok, C. R.; Heermann, M.; Hutkins, R. *Frontiers in Microbiology* **2018**, DOI: 10.3389/fmicb.2018.01785.

(5) David, L. A.; Maurice, C. F.; Carmody, R. N.; Gootenberg, D. B.; Button, J. E.; Wolfe, B. E.; Ling, A. V.; Devlin, A. S.; Varma, Y.; Fischbach, M. A.; Biddinger, S. B.; Dutton, R. J.; Turnbaugh, P. J. *Nature* **2014**, DOI: 10.1038/nature12820.

(6) Marco, M. L.; Heeney, D.; Binda, S.; Cifelli, C. J.; Cotter, P. D.; Foligné, B.; Gänzle, M.; Kort, R.; Pasin, G.; Pihlanto, A.; Smid, E. J.; Hutkins, R. *Current Opinion in Biotechnology* **2017**, DOI: 10.1016/j.copbio.2016.11.010.

(7) Peters, A.; Krumbholz, P.; Jäger, E.; Heintz-Buschart, A.; Çakir, M. V.; Rothemund, S.; Gaudl, A.; Ceglarek, U.; Schöneberg, T.; Stäubert, C. *PLoS Genetics* **2019**, DOI: 10.1371/journal.pgen.1008145.

(8) Mozaffarian, D.; Hao, T.; Rimm, E. B.; Willett, W. C.; Hu, F. B. *New England Journal of Medicine* **2011**, DOI: 10.1056/NEJMoa1014296.

(9) Chen, M.; Sun, Q.; Giovannucci, E.; Mozaffarian, D.; Manson, J. A. E.; Willett, W. C.; Hu, F. B. *BMC Medicine* **2014**, DOI: 10.1186/s12916-014-0215-1.

(10) Park, S.; Bae, J. H. *Nutrition Research* **2016**, DOI: 10.1016/j.nutres.2015.11.011.

(11) Nozue, M.; Shimazu, T.; Sasazuki, S.; Charvat, H.; Mori, N.; Mutoh, M.; Sawada, N.; Iwasaki, M.; Yamaji, T.; Inoue, M.; Kokubo, Y.; Yamagishi, K.; Iso, H.; Tsugane, S. *The Journal of Nutrition* **2017**, DOI: 10.3945/jn.117.250282.

(12) Wu, G. D.; Chen, J.; Hoffmann, C.; Bittinger, K.; Chen, Y. Y.; Keilbaugh, S. A.; Bewtra, M.; Knights, D.; Walters, W. A.; Knight, R.; Sinha, R.; Gilroy, E.; Gupta, K.; Baldassano, R.; Nessel, L.; Li, H.; Bushman, F. D.; Lewis, J. D. *Science* **2011**, DOI: 10.1126/science.1208344.

(13) Muegge, B. D.; Kuczynski, J.; Knights, D.; Clemente, J. C.; González, A.; Fontana, L.; Henrissat, B.; Knight, R.; Gordon, J. I. *Science* **2011**, DOI: 10.1126/science.1198719.

(14) Duncan, S. H.; Belenguer, A.; Holtrop, G.; Johnstone, A. M.; Flint, H. J.; Lobley, G. E. *Applied and Environmental Microbiology* **2007**, DOI: 10.1128/AEM.02340-06.

(15) Ley, R. E.; Turnbaugh, P. J.; Klein, S.; Gordon, J. I. *Nature* **2006**, DOI: 10.1038/4441022a.

(16) Walker, A. W.; Ince, J.; Duncan, S. H.; Webster, L. M.; Holtrop, G.; Ze, X.; Brown, D.; Stares, M. D.; Scott, P.; Bergerat, A.; Louis, P.; McIntosh, F.; Johnstone, A. M.; Lobley, G. E.; Parkhill, J.; Flint, H. J. *ISME Journal* **2011**, DOI: 10.1038/ismej.2010.118.

(17) Dimidi, E.; Cox, S. R.; Rossi, M.; Whelan, K. *Nutrients* **2019**, DOI: 10.3390/nu11081806.

(18) Grieneisen, L. E.; Blekhman, R. *mSystems* **2018**, DOI: 10.1128/msystems.00060-18.

(19) Falony, G.; Joossens, M.; Vieira-Silva, S.; Wang, J.; Darzi, Y.; Faust, K.; Kurilshikov, A.; Bonder, M. J.; Valles-Colomer, M.; Vandeputte, D.; Tito, R. Y.; Chaffron, S.; Rymenans, L.; Verspecht, C.; Sutter, L. D.; Lima-Mendez, G.; D'hoe, K.; Jonckheere, K.; Homola, D.; Garcia, R.; Tigchelaar, E. F.; Eeckhaudt, L.; Fu, J.; Henckaerts, L.; Zhernakova, A.; Wijmenga, C.; Raes, J. *Science* **2016**, DOI: 10.1126/science.aad3503.

(20) Zhernakova, A.; Kurilshikov, A.; Bonder, M. J.; Tigchelaar, E. F.; Schirmer, M.; Vatanen, T.; Mujagic, Z.; Vila, A. V.; Falony, G.; Vieira-Silva, S.; Wang, J.; Imhann, F.; Brandsma, E.; Jankipersadsing, S. A.; Joossens, M.; Cenit, M. C.; Deelen, P.; Swertz, M. A.; Weersma, R. K.; Feskens, E. J.; Netea, M. G.; Gevers, D.; Jonkers, D.; Franke, L.; Aulchenko, Y. S.; Huttenhower, C.; Raes, J.; Hofker, M. H.; Xavier, R. J.; Wijmenga, C.; Fu, J. *Science* **2016**, DOI: 10.1126/science.aad3369.

(21) Rothschild, D.; Weissbrod, O.; Barkan, E.; Kurilshikov, A.; Korem, T.; Zeevi, D.; Costea, P. I.; Godneva, A.; Kalka, I. N.; Bar, N.; Shilo, S.; Lador, D.; Vila, A. V.; Zmora, N.; Pevsner-Fischer, M.; Israeli, D.; Kosower, N.; Malka, G.; Wolf, B. C.; Avnit-Sagi, T.; Lotan-Pompan, M.; Weinberger, A.; Halpern, Z.; Carmi, S.; Fu, J.; Wijmenga, C.; Zhernakova, A.; Elinav, E.; Segal, E. *Nature* **2018**, DOI: 10.1038/nature25973.

(22) CDC Center for Disease Control and Prevention, National Health and Nutrition Examination Survey Data, 2011-2012.

(23) Wang, D. D.; Leung, C. W.; Li, Y.; Ding, E. L.; Chiuve, S. E.; Hu, F. B.; Willett, W. C. *JAMA Internal Medicine* **2014**, DOI: 10.1001/jamainternmed.2014.3422.

(24) Drewnowski, A.; Aggarwal, A.; Cook, A.; Stewart, O.; Moudon, A. V. *Preventive Medicine* **2016**, DOI: 10.1016/j.ypmed.2015.11.021.

(25) McDonald, D. et al. *mSystems* **2018**, DOI: 10.1128/msystems.00031-18.

(26) Lozupone, C.; Hamady, M.; Knight, R. *BMC Bioinformatics* **2006**, DOI: 10.1186/1471-2105-7-371.

(27) Morton, J. T.; Marotz, C.; Washburne, A.; Silverman, J.; Zaramela, L. S.; Edlund, A.; Zengler, K.; Knight, R. *Nature Communications* **2019**, DOI: 10.1038/s41467-019-10656-5.

(28) Fedarko, M. W.; Martino, C.; Morton, J. T.; Marotz, C. A.; Minich, J. J.; Allen, E. E.; Knight, R. Qurro., 2019.

(29) Zhu, Q.; Mai, U.; Pfeiffer, W.; Janssen, S.; Asnicar, F.; Sanders, J. G.; Belda-Ferre, P.; Al-Ghalith, G. A.; Kopylova, E.; McDonald, D.; Kosciolek, T.; Yin, J. B.; Huang, S.; Salam, N.; Jiao, J.-Y.; Wu, Z.; Xu, Z. Z.; Cantrell, K.; Yang, Y.; Sayyari, E.; Rabiee, M.; Morton, J. T.; Podell, S.; Knights, D.; Li, W.-J.; Huttenhower, C.; Segata, N.; Smarr, L.; Mirarab, S.; Knight, R. *Nature Communications* **2019**, *10*, 5477.

(30) Shannon, C. E. *Bell System Technical Journal* **1948**, DOI: 10.1002/j.1538-7305.1948.tb01338.x.

(31) Faith, D. P. *Biological Conservation* **1992**, DOI: 10.1016/0006-3207(92)91201-3.

(32) Jiang, L.; Vazquez-Baeza, Y.; Gonzalez, A.; Natarajan, L.; Knight, R.; Thompson, W. K. *JSM Proceedings* **2019**, 1836–1853.

(33) Wang, X.; Xiao, J.; Jia, Y.; Pan, Y.; Wang, Y. *Heliyon* **2018**, DOI: 10.1016/j.heliyon.2018.e00649.

(34) Fröhlich-Wyder, M. T.; Guggisberg, D.; Badertscher, R.; Wechsler, D.; Wittwer, A.; Irmler, S. *International Dairy Journal* **2013**, DOI: 10.1016/j.idairyj.2013.03.004.

(35) Giraffa, G. *Frontiers in Microbiology* **2014**, DOI: 10.3389/fmicb.2014.00338.

(36) McLeod, A.; Zagorec, M.; Champomier-Vergès, M. C.; Naterstad, K.; Axelsson, L. *BMC microbiology* **2010**, DOI: 10.1186/1471-2180-10-120.

(37) Morton, J. T.; Aksenov, A. A.; Nothias, L. F.; Foulds, J. R.; Quinn, R. A.; Badri, M. H.; Swenson, T. L.; Van Goethem, M. W.; Northen, T. R.; Vazquez-Baeza, Y.; Wang, M.; Bokulich, N. A.; Watters, A.; Song, S. J.; Bonneau, R.; Dorrestein, P. C.; Knight, R. *Nature Methods* **2019**, DOI: 10.1038/s41592-019-0616-3.

(38) Devillard, E.; McIntosh, F. M.; Duncan, S. H.; Wallace, R. J. *Journal of Bacteriology* **2007**, DOI: 10.1128/JB.01359-06.

(39) Sieber, R.; Collomb, M.; Aeschlimann, A.; Jelen, P.; Eyer, H. **2004**, DOI: 10.1016/S0958-6946(03)00151-1.

(40) Yang, B.; Gao, H.; Stanton, C.; Ross, R. P.; Zhang, H.; Chen, Y. Q.; Chen, H.; Chen, W. *Progress in Lipid Research* **2017**, DOI: 10.1016/j.plipres.2017.09.002.

(41) Van Nieuwenhove, C. P.; Teran, V.; Gonzalez, S. N. *IntechOpen* **2012**, DOI: 10.5772/50321.

(42) Koba, K.; Yanagita, T. *Obesity Research and Clinical Practice* **2014**, DOI: 10.1016/j.orcp.2013.10.001.

(43) Dilzer, A.; Park, Y. *Critical Reviews in Food Science and Nutrition* **2012**, DOI: 10.1080/10408398.2010.501409.

(44) Kishino, S.; Takeuchi, M.; Park, S. B.; Hirata, A.; Kitamura, N.; Kunisawa, J.; Kiyono, H.; Iwamoto, R.; Isobe, Y.; Arita, M.; Arai, H.; Ueda, K.; Shima, J.; Takahashi, S.; Yokozeki, K.; Shimizu, S.; Ogawa, J. *Proceedings of the National Academy of Sciences of the United States of America* **2013**, DOI: 10.1073/pnas.1312937110.

(45) Terán, V.; Pizarro, P. L.; Zacarías, M. F.; Vinderola, G.; Medina, R.; Van Nieuwenhove, C. *Journal of Functional Foods* **2015**, DOI: 10.1016/j.jff.2015.09.046.

(46) Chung, S. H.; Kim, I. H.; Park, H.; Kang, H. S.; Yoon, C. S.; Jeong, H. Y.; Choi, N. J.; Kwon, E. G.; Kim, Y. J. *Journal of Agricultural and Food Chemistry* **2008**.

(47) Coakley, M.; Ross, R. P.; Nordgren, M.; Fitzgerald, G.; Devery, R.; Stanton, C. *Journal of Applied Microbiology* **2003**, DOI: 10.1046/j.1365-2672.2003.01814.x.

(48) Subar, A. F.; Freedman, L. S.; Tooze, J. A.; Kirkpatrick, S. I.; Boushey, C.; Neuhouser, M. L.; Thompson, F. E.; Potischman, N.; Guenther, P. M.; Tarasuk, V.; Reedy, J.; Krebs-Smith, S. M. *The Journal of Nutrition* **2015**, DOI: 10.3945/jn.115.219634.

(49) Guenther, P. M.; Kirkpatrick, S. I.; Reedy, J.; Krebs-Smith, S. M.; Buckman, D. W.; Dodd, K. W.; Casavale, K. O.; Carroll, R. J. *The Journal of Nutrition* **2014**, DOI: 10.3945/jn.113.183079.

(50) Marotz, C.; Amir, A.; Humphrey, G.; Gaffney, J.; Gogul, G.; Knight, R. *BioTechniques* **2017**, DOI: 10.2144/000114559.

(51) Gonzalez, A.; Navas-Molina, J. A.; Kosciolek, T.; McDonald, D.; Vázquez-Baeza, Y.; Ackermann, G.; DeReus, J.; Janssen, S.; Swafford, A. D.; Orchanian, S. B.; Sanders, J. G.; Shorenstein, J.; Holste, H.; Petrus, S.; Robbins-Pianka, A.; Brislawn, C. J.; Wang, M.; Rideout, J. R.; Bolyen, E.; Dillon, M.; Caporaso, J. G.; Dorrestein, P. C.; Knight, R. *Nature Methods* **2018**, DOI: 10.1038/s41592-018-0141-9.

(52) McDonald, D.; Kaehler, B.; Gonzalez, A.; DeReus, J.; Ackermann, G.; Marotz, C.; Huttley, G.; Knight, R. *mSystems* **2019**, DOI: 10.1128/msystems.00215-19.

(53) Amir, A.; McDonald, D.; Navas-Molina, J. A.; Debelius, J.; Morton, J. T.; Hyde, E.; Robbins-Pianka, A.; Knight, R. *mSystems* **2017**, DOI: 10.1128/msystems.00199-16.

(54) 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.; Knight, R. *mSystems* **2017**, DOI: 10.1128/msystems.00191-16.

(55) McDonald, D.; Price, M. N.; Goodrich, J.; Nawrocki, E. P.; Desantis, T. Z.; Probst, A.; Andersen, G. L.; Knight, R.; Hugenholtz, P. *ISME Journal* **2012**, DOI: 10.1038/ismej.2011.139.

(56) *Nature Biotechnology* **2019**, DOI: 10.1038/s41587-019-0209-9.

(57) Weiss, S.; Xu, Z. Z.; Peddada, S.; Amir, A.; Bittinger, K.; Gonzalez, A.; Lozupone, C.; Zaneveld, J. R.; Vázquez-Baeza, Y.; Birmingham, A.; Hyde, E. R.; Knight, R. *Microbiome* **2017**, DOI: 10.1186/s40168-017-0237-y.

(58) Mirarab, S.; Nguyen, N.; Warnow, T. *Pacific Symposium on Biocomputing* **2012**, DOI: 10.1142/9789814366496_0024.

(59) Janssen, S.; McDonald, D.; Gonzalez, A.; Navas-Molina, J. A.; Jiang, L.; Xu, Z. Z.; Winker, K.; Kado, D. M.; Orwoll, E.; Manary, M.; Mirarab, S.; Knight, R. *mSystems* **2018**, DOI: 10.1128/msystems.00021-18.

(60) Lozupone, C.; Knight, R. *Applied and Environmental Microbiology* **2005**, DOI: 10.1128/AEM.71.12.8228-8235.2005.

(61) Anderson, M. J. *Austral Ecology* **2005**, DOI: 10.1139/cjfas-58-3-626.

(62) Wilcoxon, F. *Biometrics Bulletin* **1945**, DOI: 10.2307/3001968.

(63) Dührkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A. A.; Melnik, A. V.; Meusel, M.; Dorrestein, P. C.; Rousu, J.; Böcker, S. *Nature Methods* **2019**, DOI: 10.1038/s41592-019-0344-8.

(64) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orei, M. *BMC Bioinformatics* **2010**, DOI: 10.1186/1471-2105-11-395.

(65) Xia, J.; Sinelnikov, I. V.; Han, B.; Wishart, D. S. *Nucleic Acids Research* **2015**, DOI: 10.1093/nar/gkv380.

(66) Chong, J.; Soufan, O.; Li, C.; Caraus, I.; Li, S.; Bourque, G.; Wishart, D. S.; Xia, J. *Nucleic Acids Research* **2018**, DOI: 10.1093/nar/gky310.

(67) Wang, M. et al. *Nature Biotechnology* **2016**, DOI: 10.1038/nbt.3597.

(68) Sumner, L. W.; Amberg, A.; Barrett, D.; Beale, M. H.; Beger, R.; Daykin, C. A.; Fan, T. W.; Fiehn, O.; Goodacre, R.; Griffin, J. L.; Hankemeier, T.; Hardy, N.; Harnly, J.; Higashi, R.; Kopka, J.; Lane, A. N.; Lindon, J. C.; Marriott, P.; Nicholls, A. W.; Reily, M. D.; Thaden, J. J.; Viant, M. R. *Metabolomics* **2007**, DOI: 10.1007/s11306-007-0082-2.

(69) Nothias, L. F.; Petras, D.; Schmid, R.; Dührkop, K.; Rainer, J.; Sarvepalli, A.; Protsyuk, I.; Ernst, M.; Tsugawa, H. *biorXiv* **2019**.

(70) Mago, T.; Salzberg, S. L. *Bioinformatics* **2011**, DOI: 10.1093/bioinformatics/btr507.

(71) Hillmann, B.; Al-Ghalith, G. A.; Shields-Cutler, R. R.; Zhu, Q.; Gohl, D. M.; Beckman, K. B.; Knight, R.; Knights, D. *mSystems* **2018**, DOI: 10.1128/msystems.00069-18.

(72) Langmead, B.; Salzberg, S. L. *Nature Methods* **2012**, DOI: 10.1038/nmeth.1923.

(73) O'Leary, N. A.; Wright, M. W.; Brister, J. R.; Ciufo, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D.; Astashyn, A.; Badretdin, A.; Bao, Y.; Blinkova, O.; Brover, V.; Chetvernin, V.; Choi, J.; Cox, E.; Ermolaeva, O.; Farrell, C. M.; Goldfarb, T.; Gupta, T.; Haft, D.; Hatcher, E.; Hlavina, W.; Joardar, V. S.; Kodali, V. K.; Li, W.; Maglott, D.; Masterson, P.; McGarvey, K. M.; Murphy, M. R.; O'Neill, K.; Pujar, S.; Rangwala, S. H.; Rausch, D.; Riddick, L. D.; Schoch, C.; Shkeda, A.; Storz, S. S.; Sun, H.; Thibaud-Nissen, F.; Tolstoy, I.; Tully, R. E.; Vatsan, A. R.; Wallin, C.; Webb, D.; Wu, W.; Landrum, M. J.; Kimchi, A.; Tatusova, T.; DiCuccio, M.; Kitts, P.; Murphy, T. D.; Pruitt, K. D. *Nucleic Acids Research* **2016**, DOI: 10.1093/nar/gkv1189.

(74) Vázquez-Baeza, Y.; Pirrung, M.; Gonzalez, A.; Knight, R. *GigaScience* **2013**, DOI: 10.1186/2047-217X-2-16.