UNIVERSITY OF CALIFORNIA SAN DIEGO

Multiomics strategies for a system-wide understanding of microbiomes

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

in

Biology

by

Anupriya Tripathi

Committee in Charge:

Professor Rob Knight, Chair
Professor Joe Pogliano, Co-Chair
Professor Pieter Dorrestein
Professor Gabriel Haddad
Professor Stephen Mayfield
Professor Kit Pogliano

2020

The dissertation of Anupriya Tripathi is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California San Diego
2020

DEDICATION

To my family, friends, mentors and all who love exploring the mysteries of the Universe

EPIGRAPH

Knowledge can be communicated, but not wisdom. One can find it, live it, do wonders through it, but one cannot communicate and teach it.

—Hermann Hesse, *Siddhartha*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I am deeply grateful to my advisors Professors Rob Knight and Pieter Dorrestein for giving me the opportunity and the unparalleled freedom to pursue my research ideas under their supervision. I am very appreciative of Rob and Pieter for creating an exceptionally inclusive work environment and for their belief in me; it has helped me become the scientist I am today. I am extremely thankful to my doctoral committee, Professors Kit Pogliano, Gabriel Haddad, Stephen Mayfield, and Joe Pogliano for their guidance and encouragement over the years, and for their confidence in my work.

I am fortunate to have had invaluable mentorship during my doctoral studies. Amnon Amir and Justine Debelius were instrumental in guiding me in my first steps as a microbiome data scientist; Serene Jiang helped me develop better statistical intuitions for analyzing high-dimensional omics data. I am especially grateful to Yoshiki Vazquez-Baeza for mentoring me when I started developing software; I really appreciate Yoshiki's scientific as well as life advice which made this journey extremely rewarding and enjoyable. I also want to extend gratitude to Professor Bernd Schnabl for all the helpful discussions that advanced my understanding of metabolic diseases as well as for being invested in my success.

I had the opportunity to work closely with amazing scientists and peers: Daniel McDonald, Greg Humphrey, Julia Gauglitz, Alison Vrbanac, Louis-Felix Nothias, Kai Durkopf, Zech Xu, Qiyun Zhu, Michael Meehan, Ming Wang, Madeleine Ernst, Justin van der Hooft, Gail Ackermann, and many others, who taught me the power of teamwork and collaborative research. I especially thank Jerry Kennedy and Bryn Taylor for delightful water-cooler conversations; having such colleagues made me look forward to coming to the lab each morning.

I am immensely thankful to my family, especially my mom and brother, for helping me develop a strong value system, mental tools, and the confidence to move 8148 miles to pursue my doctoral studies; their endless faith in me kept me strong through the highs and lows of graduate school. I am very grateful to my second family – my dearest friends: Shashank Sarbada, Rushil Nagda, Siddharth Sadani, Rishika Sinha, Chhavi Sahni, Shashank Mehta, Shivam Mangla, Chetty Arun, Jayant Jain, Pragya Sidhwani, Akshay Rangesh, Min Zhang, Viraj Deshpande and Janet Johnsson for celebrating my smallest victories and for supporting me in countless ways whenever I needed them.

Finally, I am highly appreciative of UC San Diego leadership for standing by the ideals of diversity and inclusivity, and providing students with an environment that made us feel celebrated as scientists and human beings.

Chapter 1, Introduction part 1, in full, is a reprint of previously published material: Tripathi A, Marotz C, Gonzalez A, Vázquez-Baeza Y, Song SJ, Bouslimani A, McDonald D, Zhu Q, Sanders JG, Smarr L, Dorrestein PC & Knight, R. *Are microbiome studies ready for hypothesis-driven research?* Current opinion in microbiology. 2018 Aug 1;44:61-9. I was one of the primary investigators and authors of this paper. The co-authors listed above supervised or provided support for the research and have given permission for the inclusion of the work in this dissertation.

Chapter 1, Introduction part 2, in full, is a reprint of previously published material: Tripathi, A., Debelius, J., Brenner, D. A., Karin, M., Loomba, R., Schnabl, B., & Knight, R. (2018). The gut-liver axis and the intersection with the microbiome. *Nature Reviews. Gastroenterology & Hepatology*, *15*(7), 397–411. I was the primary investigator and authors of this paper. The co-authors listed above supervised or provided support for the research and have given permission for the inclusion of the work in this dissertation.

Chapter 2, in full, is a reprint of previously published material: Caussy, C., Tripathi, A., Humphrey, G., Bassirian, S., Singh, S., Faulkner, C., Bettencourt, R., Rizo, E., Richards, L., Xu, Z. Z., Downes, M. R., Evans, R. M., Brenner, D. A., Sirlin, C. B., Knight, R., & Loomba, R. (2019). A gut microbiome signature for cirrhosis due to nonalcoholic fatty liver disease. *Nature Communications*, *10*(1), 1406. I was one of the primary investigator and author of this paper. The co-authors listed above supervised or provided support for the research and have given permission for the inclusion of the work in this dissertation.

Chapter 3, in full, is a reprint of previously published material: Tripathi, A., Melnik, A. V., Xue, J., Poulsen, O., Meehan, M. J., Humphrey, G., Jiang, L., Ackermann, G., McDonald, D., Zhou, D., Knight, R., Dorrestein, P. C., & Haddad, G. G. (2018). Intermittent Hypoxia and Hypercapnia, a Hallmark of Obstructive Sleep Apnea, Alters the Gut Microbiome and Metabolome. In *mSystems* (Vol. 3, Issue 3). I was the primary investigator and author of this paper. The co-authors listed above supervised or provided support for the research and have given permission for the inclusion of the work in this dissertation.

Chapter 4, in full, is a reprint of previously published material: Tripathi, A., Xu, Z. Z., Xue, J., Poulsen, O., Gonzalez, A., Humphrey, G., Meehan, M. J., Melnik, A. V., Ackermann, G., Zhou, D., Malhotra, A., Haddad, G. G., Dorrestein, P. C., & Knight, R. (2019). Intermittent Hypoxia and Hypercapnia Reproducibly Change the Gut Microbiome and Metabolome across Rodent Model Systems. *mSystems*, *4*(2). I was one of the primary investigator and author of this paper. The co-authors listed above supervised or provided support for the research and have given permission for the inclusion of the work in this dissertation.

Chapter 5, in full, is a preprint of: Tripathi, A., Vázquez-Baeza, Y., Gauglitz, J. M., Wang, M., Dührkop, K., Nothias-Esposito, M., Acharya, D. D., Ernst, M., van der Hooft, J. J. J., Zhu, Q., McDonald, D., Gonzalez, A., Handelsman, J., Fleischauer, M., Ludwig, M., Böcker, S., Nothias, L.-F., Knight, R., & Dorrestein, P. C. (n.d.). *Chemically-informed Analyses of Metabolomics Mass Spectrometry Data with Qemistree*. I was one of the primary investigators and authors of this paper. The co-authors listed above supervised or provided support for the research and have given permission for the inclusion of the work in this dissertation.

VITA

2011-2015 Bachelor of Technology, Indian Institute of Technology, Roorkee

2015-2020 Doctor of Philosophy, University of California San Diego

PUBLICATIONS

Tripathi A, Vázquez-Baeza Y, Gauglitz JM, Wang M, Dührkop K, Nothias-Esposito M, Acharya DD, Ernst M, van der Hooft JJJ, Zhu Q, McDonald D, Gonzalez A, Handelsman J, Fleischauer M, Ludwig M, Böcker S, Nothias L-F, Knight R, Dorrestein PC. Chemically-informed Analyses of Metabolomics Mass Spectrometry Data with Qemistree.

Gauglitz JM, Morton JT, **Tripathi A**, Hansen S, Gaffney M, Carpenter C, Weldon KC, Shah R, Parampil A, Fidgett AL, Swafford AD, Knight R, Dorrestein PC. 2020. Metabolome-Informed Microbiome Analysis Refines Metadata Classifications and Reveals Unexpected Medication Transfer in Captive Cheetahs. *mSystems* 5.

Gauglitz JM, Aceves CM, Aksenov AA, Aleti G, Almaliti J, Bouslimani A, Brown EA, Campeau A, Caraballo-Rodríguez AM, Chaar R, da Silva RR, Demko AM, Di Ottavio F, Elijah E, Ernst M, Ferguson LP, Holmes X, Jarmusch AK, Jiang L, Kang KB, Koester I, Kwan B, Li J, Li Y, Melnik AV, Molina-Santiago C, Ni B, Oom AL, Panitchpakdi MW, Petras D, Quinn R, Sikora N, Spengler K, Teke B, **Tripathi A**, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vrbanac A, Vu AQ, Wang SC, Weldon K, Wilson K, Wozniak JM, Yoon M, Bandeira N, Dorrestein PC. 2020. Untargeted mass spectrometry-based metabolomics approach unveils molecular changes in raw and processed foods and beverages. *Food Chem* 302:125290.

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

Caussy C, **Tripathi A**, Humphrey G, Bassirian S, Singh S, Faulkner C, Bettencourt R, Rizo E, Richards L, Xu ZZ, Downes MR, Evans RM, Brenner DA, Sirlin CB, Knight R, Loomba R. 2019. A gut microbiome signature for cirrhosis due to nonalcoholic fatty liver disease. *Nat Commun* 10:1406.

Allaband C, Brown SD, **Tripathi A**, Russell BJ, Poulsen O, Meehan M, Ackermann G, Elijah EO, Smith T, Fouts JK, Knight R, Dorrestein P, Haddad GG, Zarrinpar A. 2019. Sa1915 – Altered Circadian Rhythm Cycling of Microbes and Metabolites in a Murine Model of Sleep Apnea. *Gastroenterology*.

Xu ZZ, Amir A, Sanders J, Zhu Q, Morton JT, Bletz MC, **Tripathi A**, Huang S, McDonald D, Jiang L, Knight R. 2019. Calour: an Interactive, Microbe-Centric Analysis Tool. *mSystems* 4.

Tripathi A, Xu ZZ, Xue J, Poulsen O, Gonzalez A, Humphrey G, Meehan MJ, Melnik AV, Ackermann G, Zhou D, Malhotra A, Haddad GG, Dorrestein PC, Knight R. 2019. Intermittent Hypoxia and Hypercapnia Reproducibly Change the Gut Microbiome and Metabolome across Rodent Model Systems. *mSystems* 4.

Allaband C, McDonald D, Vázquez-Baeza Y, Minich JJ, **Tripathi A**, Brenner DA, Loomba R, Smarr L, Sandborn WJ, Schnabl B, Dorrestein P, Zarrinpar A, Knight R. 2019. Microbiome 101: Studying, Analyzing, and Interpreting Gut Microbiome Data for Clinicians. *Clin Gastroenterol Hepatol* 17:218–230.

Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu Y-X, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS 2nd, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, **Tripathi A**, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG. 2019. Author Correction: Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* 37:1091.

Tripathi A, Marotz C, Gonzalez A, Vázquez-Baeza Y, Song SJ, Bouslimani A, McDonald D, Zhu Q, Sanders JG, Smarr L, Dorrestein PC, Knight R. 2018. Are microbiome studies ready for hypothesis-driven research? *Curr Opin Microbiol* 44:61–69.

Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, Gonzalez A, Kosciolek T, McCall L-I, McDonald D, Melnik AV, Morton JT, Navas J, Quinn RA, Sanders JG, Swafford AD, Thompson LR, **Tripathi A**, Xu ZZ, Zaneveld JR, Zhu Q, Caporaso JG, Dorrestein PC. 2018. Best practices for analysing microbiomes. *Nat Rev Microbiol* 16:410–422.

McCall L-I, **Tripathi A**, Vargas F, Knight R, Dorrestein PC, Siqueira-Neto JL. 2018. Experimental Chagas disease-induced perturbations of the fecal microbiome and metabolome. *PLoS Negl Trop Dis* 12:e0006344.

Zhou Y, Xu ZZ, He Y, Yang Y, Liu L, Lin Q, Nie Y, Li M, Zhi F, Liu S, Amir A, González A, **Tripathi A**, Chen M, Wu GD, Knight R, Zhou H, Chen Y. 2018. Gut Microbiota Offers Universal Biomarkers across Ethnicity in Inflammatory Bowel Disease Diagnosis and Infliximab Response Prediction. *mSystems* 3.

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

Tripathi A, Debelius J, Brenner DA, Karin M, Loomba R, Schnabl B, Knight R. 2018. The gut-liver axis and the intersection with the microbiome. *Nat Rev Gastroenterol Hepatol* 15:397–411.

McDonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, Aksenov AA, Behsaz B, Brennan C, Chen Y, DeRight Goldasich L, Dorrestein PC, Dunn RR, Fahimipour AK, Gaffney J, Gilbert JA, Gogul G, Green JL, Hugenholtz P, Humphrey G, Huttenhower C, Jackson MA, Janssen S, Jeste DV, Jiang L, Kelley ST, Knights D, Kosciolek T, Ladau J, Leach J, Marotz C, Meleshko D, Melnik AV, Metcalf JL, Mohimani H, Montassier E, Navas-Molina J, Nguyen TT, Peddada S, Pevzner P, Pollard KS, Rahnavard G, Robbins-Pianka A, Sangwan N, Shorenstein J, Smarr L, Song SJ, Spector T, Swafford AD, Thackray VG, Thompson LR, **Tripathi A**, Vázquez-Baeza Y, Vrbanac A, Wischmeyer P, Wolfe E, Zhu Q, American Gut Consortium, Knight R. 2018. American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems* 3.

Thompson LR, The Earth Microbiome Project Consortium, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, **Tripathi A**, Gibbons SM, Ackermann G, Navas-Molina JA, Janssen S, Kopylova E, Vázquez-Baeza Y, González A, Morton JT, Mirarab S, Xu ZZ, Jiang L, Haroon MF, Kanbar J, Zhu Q, Song SJ, Kosciolek T, Bokulich NA, Lefler J, Brislawn CJ, Humphrey G, Owens SM, Hampton-Marcell J, Berg-Lyons D, McKenzie V, Fierer N, Fuhrman JA, Clauset A, Stevens RL, Shade A, Pollard KS, Goodwin KD, Jansson JK, Gilbert JA, Knight R. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*.

Melnik AV, da Silva RR, Hyde ER, Aksenov AA, Vargas F, Bouslimani A, Protsyuk I, Jarmusch AK, **Tripathi A**, Alexandrov T, Knight R, Dorrestein PC. 2017. Coupling Targeted and Untargeted Mass Spectrometry for Metabolome-Microbiome-Wide Association Studies of Human Fecal Samples. *Anal Chem* 89:7549–7559.

ABSTRACT OF THE DISSERTATION

Multiomics strategies for a system-wide understanding of microbiomes

by

Anupriya Tripathi

Doctor of Philosophy in Biology

University of California San Diego, 2020

Professor Rob Knight, Chair

Professor Joe Pogliano, Co-Chair

The current state of the world – stricken by a Coronavirus pandemic that spread across the globe harming millions of lives in a matter of months – is a humbling reminder that we have barely scraped the surface of understanding complex biological systems.

The central theme of this work is the idea that human health is a function of many complex biological machineries working synergistically, together constituting the human superorganism. The dissertation begins with a description of the computational advancements required to

understand the complexity of the human microbiota and its role in human health; this is followed by a review of the role of our commensal microbiota in nonalcoholic steatohepatitis, alcoholic steatohepatitis and liver cirrhosis, and hepatocellular carcinoma.

Chapters 2, 3, and 4 report original research studies highlighting the potential of leveraging high-throughput molecular assays and supervised algorithms to develop new microbiome-based diagnostic, prognostic, and therapeutic modalities to improve the management of metabolic diseases. These chapters also underscore the importance of multi-omics approaches to probe biological systems for a system-wide understanding.

Chapter 5 introduces Qemistree – a new tool that adapts statistical concepts from microbial ecology for the analysis of high-dimensional metabolomics data. Qemistree underscores the importance of mapping existing analytical solutions across omics domains in order to integrate heterogeneous data layers and comprehensively understand biological systems.

Chapter 1. Introduction

This introduction is divided ino two sections. The first section is a an opinion piece recognizing the current knowledge gaps preventing targeted microbiome perturbation.

The second chapter describes gut–liver communications describing evidence from animal and human studies, compares liver disease spectrum and highlights key points for designing microbiome-based studies for liver disease research. This serves as a transition to my research on finding early-stage microbial biomarkers of liver disease.

1.1. Are microbiome studies ready for hypothesis-driven research?

Hypothesis-driven research has led to many scientific advances, but hypotheses cannot be tested in isolation: rather, they require a framework of aggregated scientific knowledge to allow questions to be posed meaningfully. This framework is largely still lacking in microbiome studies, and the only way to create it is by discovery- and tool-driven research projects. Here we describe the value of several such projects from our own laboratories, including the American Gut Project, the Earth Microbiome Project (which is an umbrella project integrating many smaller hypothesis-driven projects), and the knowledgebase-driven tools GNPS and Qiita. We argue that an investment of community resources in these infrastructure tasks, and in the controls and standards that underpin them, will greatly enhance the investment of hypothesis-driven research programs.

1.1.1 Introduction

Microbiome research is making dramatic progress, with thousands of papers now published each year linking specific microbes and/or host-microbe co-metabolites to specific diseases, physiological properties, or environmental parameters. Much of this research is performed in a traditional, hypothesis-driven way, or at least presented as a rational reconstruction that fits this model, much as Darwin re-wrote much of his discovery-driven work as hypothesis driven to increase its respectability under the influence of contemporary philosophers of science such as William Whewell (1). However, it should be noted that hypothesis-driven science was not always so respectable -- Isaac Newton famously wrote "*Hypotheses non fingo*", or "I feign no hypotheses", in an essay appended to the second edition of the *Principia* (2) -- so the tradition of modifying how science is framed in order to meet respectability criteria dates back at least 300

years. In any case, what can be framed as a singular hypothesis suffers important limitations based on what we can measure, and what we already know.

Ten years ago Chris Anderson, editor of Wired magazine, set off an international debate with his article "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete" (3). The idea was that with enough data, hypotheses will emerge from the data ("Let the data speak for itself") has become widely discussed in the rapidly growing data science profession. A thoughtful review of this topic was written in EMBO Reports in 2015-"Could Big Data be the end of theory in science? A few remarks on the epistemology of data-driven science" (4). As the author points out:

> Francis Bacon, the "father of the scientific method" himself, in his Novum Organum (1620), argued that scientific knowledge should not be based on preconceived notions but on experimental data. Deductive reasoning, he argued, is eventually limited because setting a premise in advance of an experiment would constrain the reasoning so as to match that premise. Instead, he advocated a bottom-up approach: In contrast to deductive reasoning, which has dominated science since Aristotle, inductive reasoning should be based on facts to generalize their meaning, drawing inferences from observations and data.

One constant in microbiome research has been that most factors that we would intuitively suspect to drive differences in the microbiome are of minor importance. For example, although long-term dietary changes have a major effect on the microbiome, short-term changes don't(5,6). Similarly, sex has a very limited impact on microbiomes across the human body (7,8) and has a much weaker effect than many other variables such as age (even within adults) and the time of year the sample was collected (9,10). Perhaps more surprisingly, factors such as temperature and pH have a much smaller impact on environmental microbiomes than salinity (11,12), and even the saline vs. non-saline difference is much smaller than the host-associated vs free-living difference (12,13). Samples from different sites of the same person's body can be more different from one

another in terms of their overall microbial communities than radically different free-living microbial communities, such as soils versus oceans (12). Differences of this magnitude can also occur within the gut of a single person, with sufficiently large perturbation (DOI: 10.1101/277970).

As a consequence, it is easy to incorrectly frame hypotheses, especially when supervised ordination and classification techniques are used in experiments with many confounding variables. For example, suppose that for mouse experiments we don't know that cage effects are important in the microbiome (14), then we profile the microbiomes in each of two cages of each of two different genotypes of mice. Our results are likely to be driven by which pair of cages happens to resemble each other more closely. If the variable of cage is not measured, or not tested in an unsupervised model, we might never know that our results are driven by this important confounding variable! There may be many more important confounding variables that we are not yet aware of, so longitudinal studies with meticulous metadata annotation will be crucial for defining which environmental factors matter. This is especially important in the context of clinical samples, where single data points are often collected and obtaining contextual information in retrospect is exceedingly difficult (15).

Similarly, a frequent practice is to discard unannotated microbes or unannotated molecules, focusing on the subset of microbes or molecules that can be matched to an existing database. Because databases of both microbes and molecules are heavily biased (microbes, by studies of known pathogens which come from only a small number of taxonomic groups, and molecules, by commercially available compounds), the entities that actually best discriminate among classes of samples may be lost in the analysis: often, only 60% of sequences and 2% of molecular features from an untargeted metabolomics experiment can be annotated by existing references (16,17).

However, a rational reconstruction of why the annotatable microbes or molecules are plausible can always be developed by creative scientists looking to respond to their reviewers' criticism that their manuscript is "too descriptive".

1.1.2 The need for maps

An important metaphor in science and information visualization is the idea of the map, whether of real spaces or of abstract spaces. Indeed, as data volumes increase, it is frequent that the field moves from tests of hypotheses among sites, to tests of these hypotheses with replicates at each site, to spatially or temporally explicit sampling, to detailed spatial maps. This progression has already occurred in 16S rRNA amplicon-based microbiome studies over the past decade (12,18), and has increasingly been taking place in mass spectrometry-based metabolome studies over the past four years (19-24).

The value of spatial maps is so self-evident that the results are often cursed by obviousness. For example, the finding that metabolomes cluster by individual, as revealed by principal coordinates analysis (PCoA), is interesting (Figure 1.1.1A). However, the finding that a given molecule such as lauryl sulphate (*m/z* 355.219) covers one individual, but is absent from the other individual is obvious (Figure 1.1.1B), especially when you know that individual subject A uses a stereotypically gendered product such as Nivea for Men, which is the source of the molecule (20). How such personal lifestyle (often hygiene, health or beautification related) influences the microbiome is not known; it is also not known how even some basic parameters such as, skin temperature, skin pH, amount of sebum influences the microbial communities on the skin. Similarly, the finding that samples from four individuals differ to a statistically significant extent in their levels of specific purines and that within an individual, such molecules are also non-

randomly distributed, might well be an intriguing finding prompting more investigation. However, a spatial map with dense sampling of the same individuals (Figure 1.1.1C) makes it obvious that the molecule is something that is touched and consumed, and sometimes spilled, allowing one to guess that it is probably caffeine and that one person likely spends time in the ocean based on the distribution of *Synechococcus* spp. (Figure 1.1.1D) (both of which are in fact the case) (22).

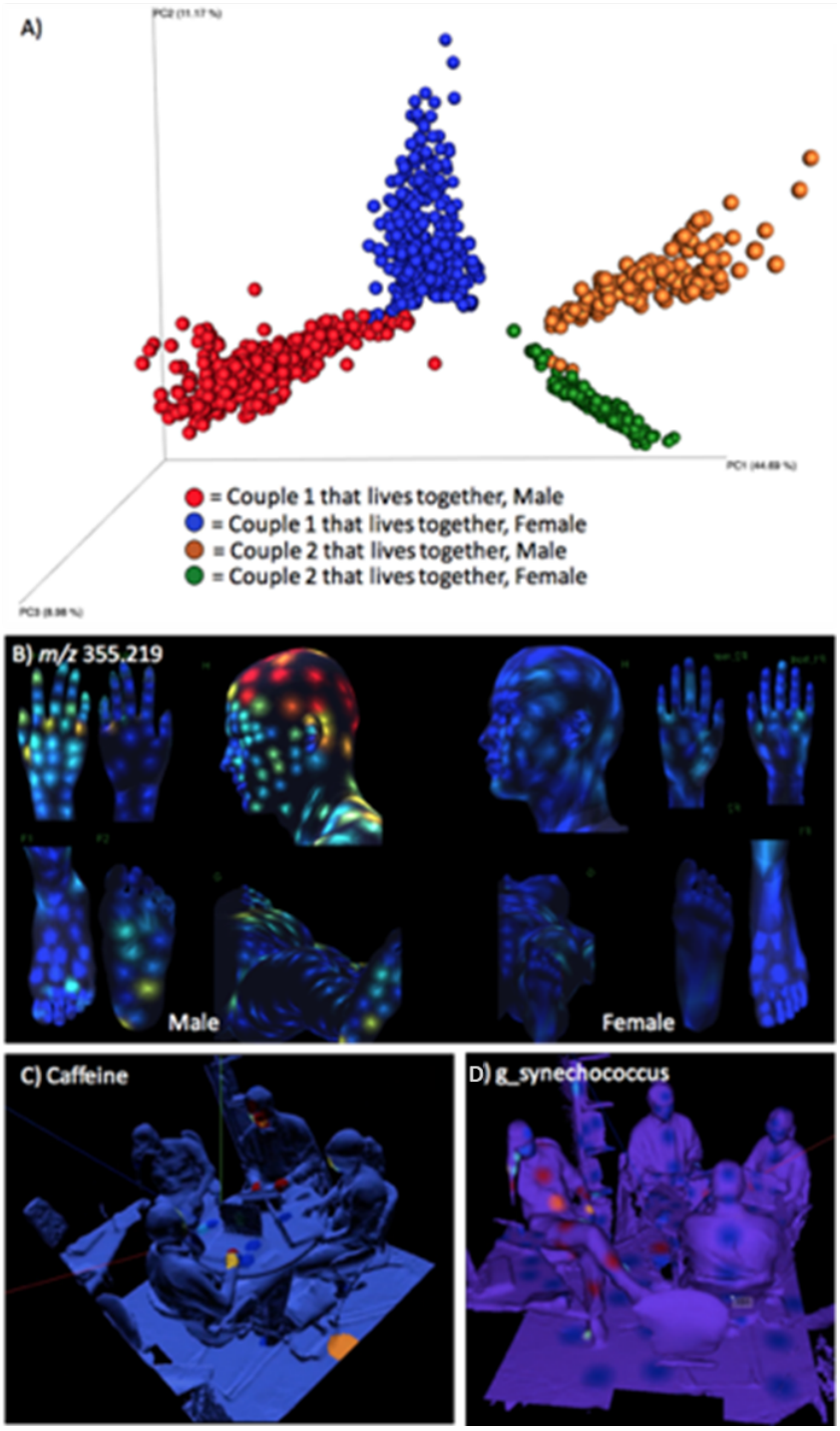

Figure 1.1.1 *Spatial analysis based on metabolomics of skin samples and a human habitat*. **A)** Principal coordinates analysis (Hellinger distance) of metabolomics data of skin swabs obtained from several hundreds locations on the human body of four volunteers. **B)** The detection of lauryl sulfate (*m/z* 355.219) from the shampoo Nivea for Men on a male volunteer. **C)** The distribution of caffeine (m/z 195.088) on four individuals and the office environment. **D)** The distribution of *Synechococcus* spp. on within that same office environment.

However, the fact remains that for most microbes and for most molecules, we have no idea where they are in and on the human body, in natural environments, or in human-impacted environments including built environments. Just as John Snow's map of cholera instantly led to the hypothesis that this disease was water-borne and stemmed from the Broad Street pump, reinforced by the map's revelation that the block that drank alcohol had no incidence of disease (25). The power of maps is shown by the history that this visual display of disease incidences by street became the foundation for the science and practice of epidemiology. In an analogous manner, systematically collected maps of microbes and of molecules across different spatial scales will dramatically improve our ability to make useful inferences from this data. Integration of these maps with other data layers ranging from air pollution to food deserts and neighborhood walkability, together with zoomable user interfaces (consider the utility of Google Maps versus earlier fixed-scale maps on DVD), will fundamentally transform the types of questions that can be asked of microbiome and metabolomics data.

The value of abstract maps, whether ordinations such as principal coordinates analysis (PCoA), non-metric multidimensional scaling (NMDS), t-distributed stochastic neighbor embedding (t-SNE), network diagrams obtained from object similarity (sequence or spectrum), or from co-occurrence across samples, is also considerable. In particular, when the right data frame and metrics are chosen, the key result is often immediately obvious. Consider, for example, the starting and ending time point of a fecal transplantation series (26) (Figure 1.1.2A), where it's obvious that the clusters are statistically significantly different, but it is not obvious what direction this difference is in or what it means. However, when we perform a meta-analysis and put these samples in the context of the Human Microbiome Project data (8), one of the most important abstract maps in human microbiome science, we see immediately that the difference between start

and endpoint is much greater than the difference between healthy and diseased samples, and when we add the intermediate timepoints we see that the transition occurs very rapidly. These types of examples prompt similar data collection and visualization techniques in metabolomics, in order to understand how we can identify a desirable metabolomic state (for example, by comparing healthy and sick individuals), and guide an undesirable state into a desirable one by optimizing the trajectory towards the desired state in a series of perturbations. Only the existence of a map can allow rational hypotheses about what to try, especially in the context of n=1 studies or in cases where response heterogeneity among individuals is extreme.

1.1.3 The need for tools

We have seen, quite literally, the value of maps. But how do we build them? The key to acquiring high-resolution data, whether spatially or temporally resolved, or dense enough in an abstract space, is to make sampling fast, cheap, and sufficiently precise. Unfortunately, the trade-offs among these approaches are typically not well understood.

In DNA sequencing, a common question is whether, given a fixed sequencing budget, it is better to have more sequences per sample, or more samples. In general the answer to this question depends on the hypothesis to be tested. But, as noted above, all too frequently the "hypothesis" is retrofitted to an arbitrarily collected dataset. What guidelines can be provided for aspiring microbial cartographers?

In our experience, for amplicon sequencing, the value of having more samples has always outweighed the value of having more sequences per sample, down to surprisingly low thresholds. For example, Figure 1.1.3 shows the Earth Microbiome Project dataset (12) sampled at 500,000 sequences per sample, 1000 sequences per sample, and just 200 sequences per sample. The overall

pattern, e.g. the host/non-host split and the saline/non-saline split, are much clearer with more samples than with more precision about the location of each sample in PCoA space. Multinomial sampling considerations make it immediately clear why this is true: with 100 sequences per sample, the standard error in inferring the proportion of a taxon at 5% frequency is ~sqrt(100*.95*.05) or 2.18, or about 50% error in proportion; the standard error at a taxon at 1% frequency is about ±1, or about 100% error. Consequently, even low-abundance taxa are sampled with enough accuracy to place a sample in the context of an overall map with surprisingly few sequences. Logically, this must be true, or all ordination diagrams in microbial ecology before the advent of next-generation sequencing would have been useless, yet many revealed biologically interesting principles. The goal for better amplicon maps should therefore be to process vast numbers of additional samples inexpensively, exploiting the power of modern sequencers.

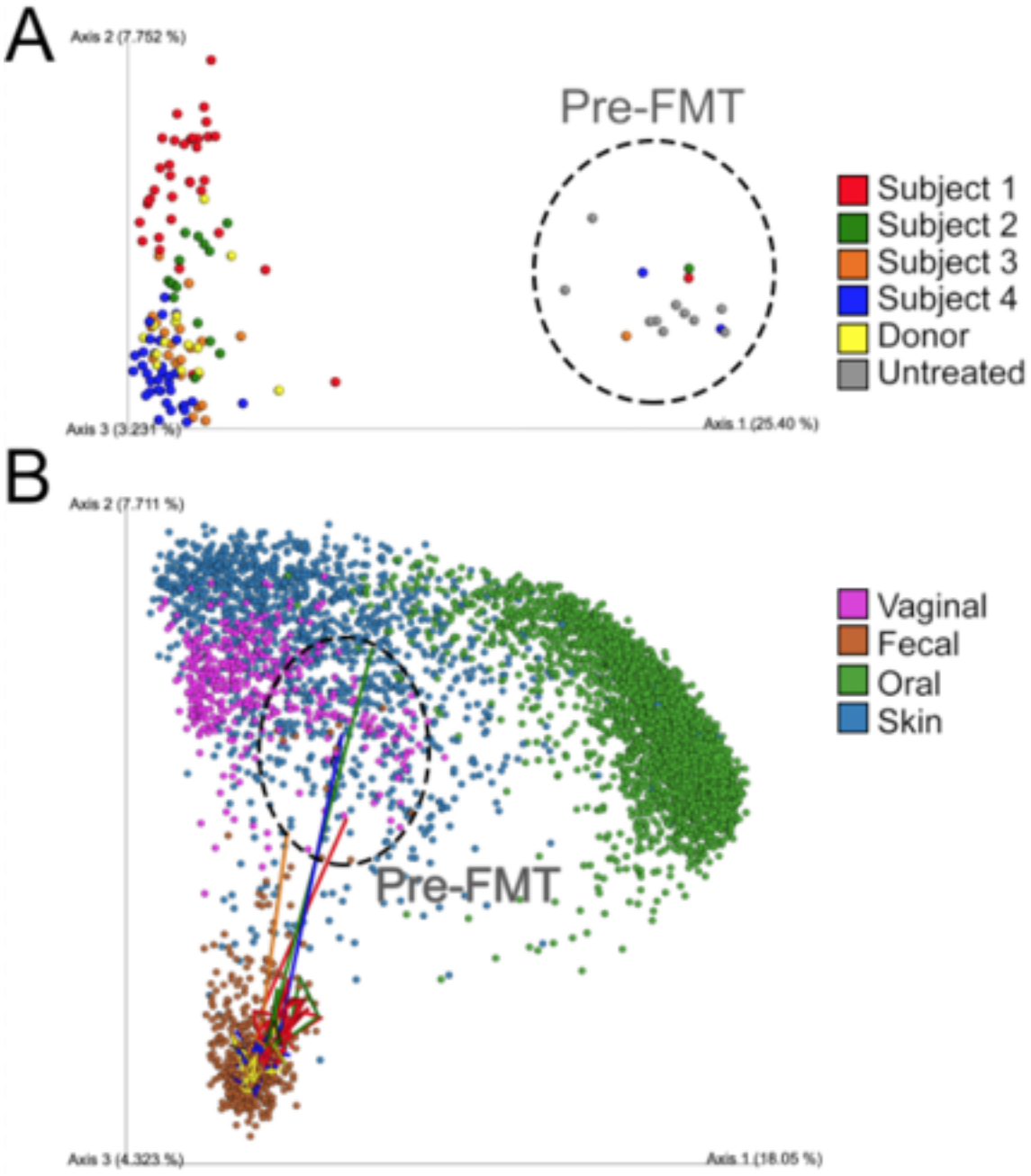

Figure 1.1.2 *Untangling the meaning of complex microbial interactions through meta-analyses*. **(A)** Principal coordinates analysis (unweighted UniFrac) of *Clostridium difficile* Infection subjects, before and after a fecal transplant, along with the fecal donor and 10 untreated subjects (26). **(B)** Principal coordinates analysis (unweighted UniFrac) of the Human Microbiome Project (HMP) (8) combined with the data in panel A, the longitudinal samples for subjects 1-4 are connected as lines displaying the temporal variability and the shift from a disjointed untreated state of the patients vs. the healthy frame of the HMP.

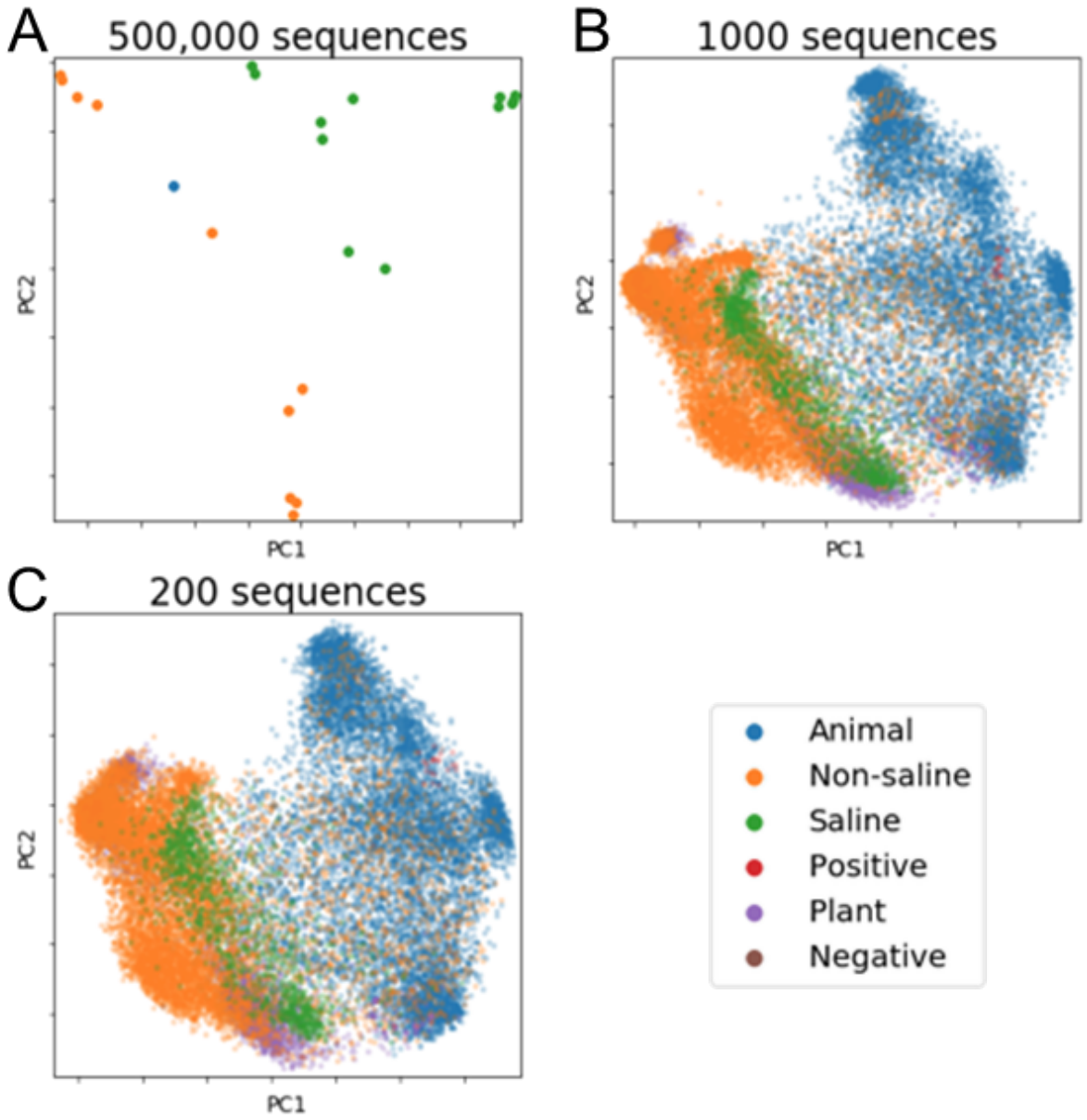

Figure 1.1.3 *Broader sampling improves maps of the microbial world, even with low resolution. All panels show principal coordinates analysis of unweighted UniFrac distances between samples.* **(A)** Samples rarefied to 500,000 sequences, showing only those exceeding this threshold sampling depth. **(B)** Samples rarefied to 1000 sequences. **(C)** Samples rarefied to 200 sequences. Even with few observations per sample, the overall relationships among sample types are preserved; in contrast, the overall pattern is lost with too few samples no matter how exquisitely characterized.

Shotgun metagenomics, however, poses a different challenge, because typically only a small fraction of the sequences can be confidently associated with known taxonomy or function. Further, the goals are often different because of the value of genome assembly in identifying

biosynthetic pathways, allowing taxonomic resolution at the species or strain level, and generating high-resolution single nucleotide polymorphism (SNP) profiles to characterize novel strains and to confirm functional variants (27). As a result, although the same sampling principles as for amplicon data apply if the goal is to provide a high-level taxonomic profile, far more sequences must be collected to have the same level of confidence in the result. Consequently, the most important areas for tool development in shotgun metagenomics are either several additional orders of magnitude drop in sequencing cost, reference databases that are more comprehensive and unbiased, and algorithms that are more efficient and accurate in read alignment, genome assembly and separation. In particular, methods that can identify genetic variation from lower-coverage data, and methods for estimating features of interest from less data or with efficient target capture, are of significant necessity. Another important consideration in shotgun metagenomics requires host DNA depletion, both experimentally and computationally, because total DNA extracts from biological specimens can be dominated by host DNA that is not picked up by standard PCR primers for bacterial/archaeal amplicon sequencing (28).

The challenges in metabolomics are somewhat different (29). Sequencing has reduced in cost by nine orders of magnitude per data volume. In comparison, mass spectrometry, during the same time, has only reduced in cost of data volume collection by two orders of magnitude (29). However the main limitation is the enormous diversity in chemistry. Unlike just four bases one has to "identify" to enable sequencing, there are hundreds to thousands of molecules that need to be identified from a list of millions, if the molecule is known to exist at all. The chemical diversity also impacts the choice of extraction solvents during sample preparation, type of separation methods, type of instruments used and data analysis approaches. Further, because the multiplexing strategies that are successful in both amplicon- and shotgun-based sequencing approaches are not

available in mass spectrometry, instrument time is directly proportional to the number of samples. Consequently, although it is easy to slip a few more samples into a mass spec run, instrument time is limiting for large-scale projects. As was the case with sequencing a decade ago, the vast majority of molecular features that are found in a sample are currently unidentified, and many are likely technical artifacts of various steps in the process, e.g. adducts formed in the gas phase, solvent artifacts (30) and multimers of the same compound (29). Better methods and incentives for aggregating community knowledge (17) (e.g. retention of knowledge of the large number of manual annotations performed by the community) and for automatically assigning unknown mass peaks and fragmentation spectra to molecules and have an estimation of error rates (31), as opposed to heuristics subject to personal interpretation rules (32), are urgently needed. Global Natural Products Social molecular networking (GNPS) (17) offers alternative solutions for computational mass spectrometry infrastructure. Spectral datasets can be publicly deposited with a unique identifier and transformed to "living data" as they will be continuously searched against reference libraries to update users on new identifications. Furthermore, annotations can also be made by the scientific community within GNPS and propagated to all other data sets in the public domain with notifying subscribers on new annotations. This living data concept is crucial way to ensure that collected metabolomics data can still be useful over time. Other examples include automated species metabolome references (33) and the Molecular Explorer (17) for cross-searching annotated MS/MS spectra between datasets. Connections between several datasets, within the same knowledge base or between different spectral repositories such as Metabolights (34) and Metabolomics Workbench (35), can be made to highlight annotated compounds found in several data sets Such analysis is a trivial task in sequencing but still novel in mass spectrometry.

Integration of taxonomic, genomic and metabolomic data remains an important unsolved challenge. Although genome mining is successful for identifying the sources of individual natural products, matching up the overall taxonomic or functional profile to a molecular profile remains challenging because of procedural and analytical differences in data acquisition. In particular, the likelihood of time lags in chemical production or in genomic response to environmental changes, which may appear on different timescales, make integrated analysis of snapshot data extremely challenging (36). In cases where microbial and molecular composition is driven by a dominant effect (e.g. a dataset composed of soil and fecal samples), the molecular and metagenomic datasets will appear concordant by Procrustes analysis (37), which measures the fit of one ordination space to another. It is likely that an integrated systems biology approach that maps all data layers onto common pathways will be needed. This task is complicated at present not only because most genes, pathways, and molecules are unknown (especially those involving biotransformations of environmental or food inputs) but also because, even for the known components of the system, we still lack coherent ontological conventions across databases which may aid in connecting these data layers. Integrating this extended universe of possible molecules and their transformations across space, time, and species in complex ecologies will require fundamentally new approaches, and orders of magnitude more computing power, than are available today.

1.1.4 The need for standards

Another branch of non-hypothesis-driven research, but critically important to framing precise hypotheses, is the development of standards. In microbiome science these broadly take three tracks: procedural standards for sample collection and handling, analytical standards for

determining the accuracy and fidelity of readouts, and annotation standards for integrating results across studies.

The lack of agreed-on standards stems from the origin of much of microbiome science in the discipline of ecology, where the fundamental questions revolved around finding new kinds of organisms to fill out the phylogenetic tree of life, and around finding statistically significant differences in microbial diversity or composition among sets of samples within the context of an individual study. Because the goal was to test whether any difference existed in the microbiome or metabolome as a function of disease, physiological, or environmental state, biases (including missing taxa, or missing classes of molecules) were not terribly important as long as a difference could be discovered.

However, this situation diverges radically from the present situation, where physicians and engineers expect to be able to measure the correct, absolute abundance of all microbes or molecules in a given sample simultaneously. The realities of nucleic acid or organic extraction, detection methods for sequences and molecules, and downstream data processing simply do not support this important goal. However, in general, we don't even know how far we are from it, or what the specific blind spots are. Consequently, without consistent and well-defined measurements underpinned by a mechanistic causal model of change, the state of microbiome-based predictions is much more like astrology than like astronomy.

In order to move from pre-science to science in predicting microbiome changes, we need known reference standards that can be spiked into samples at different stages, from original specimen to DNA or molecule, that are agreed on, widely used in the field, and have an inexhaustible supply. Previous efforts, such as the HMP standards, have been limited by insufficient availability of materials, taxonomic complexity, or both. KatharoSeq in particular (38)

benefits from having different spike-in standards at the level of the primary sample and at the level of DNA, allowing different sources of contamination to be tracked down. Comparable development in mass spectrometry, perhaps with isotope-labeled molecules or molecules otherwise unlikely to occur in biological specimens and that can be introduced at different steps, would be of tremendous value.

Sample collection and storage can introduce biases of varying degrees in specimen readout (39-41), but for most sample types the precise implications of different forms of degradation are unknown. Consequently, the conservative recommendation is always to expensively collect pristine samples (e.g. flash-frozen in liquid nitrogen), even while more practical methods would often suffice. For a few sample types, such as amplicon processing of stool, considerable data is now available on a range of conditions (41-44), and researchers can make more informed decisions about which methods to use. However, we know much less about the implications of sample degradation for most other types of biospecimens, and for the implications for reading out different molecular fractions with mass spectrometry (although see (45)). Understanding these principles would greatly expand accessibility of these techniques to field, clinical, and self-collected specimens (by patients and citizen-scientists), as the American Gut Project is already doing for amplicon collection from stool.

Finally, integrating samples from different studies remains extremely challenging because of differences in annotation (often called "metadata"). For example, different studies may refer to "stool", "feces", "gut", or other synonyms or rely on different units of measurement (e.g., Celsius vs. Fahrenheit). Efforts such as the Genomic Standards Consortium MIxS family of standards (46), the Earth Microbiome Project Ontology (EMPO) (12), and other annotation schemes assist considerably in these tasks, but have been applied to relatively few datasets to date. The potential

for natural language processing (NLP) and/or data-based methods for automatically applying annotations, perhaps semi-supervised by human guidance, is considerable. These types of strategies were successful in Qiita for inferring EMPO annotations for tens of thousands of samples in Qiita primarily based off the researcher reported “sample_type.” Resources like Qiita, which allow researchers to deposit microbiome studies, provide mechanisms to help researchers use standard compliant metadata. However, further development is necessary to enable researchers to “discover” the types of variables and controlled vocabularies that are in common across the resource.

1.1.5 Conclusions

Although hypothesis-driven science has immense value, it depends to a considerable degree on a framework of maps, tools, and standards whose development often does not fit meaningfully into a hypothesis-driven framework and is therefore heavily criticized in settings such as grant review panels. However, without these types of development, hypotheses more explicit than “differences in the microbiome” or “elevation or depletion of specific taxa or molecules” cannot be tested, and completely new ideas about how to read out or control the microbiome will not be developed.

Extraordinary advances in data collection technologies leave us in a world where we regularly make millions of observations of organisms about which we know virtually nothing -- as exemplified by the recent 'discovery' of the most abundant phage in the human gut via metagenome mining (47). The amount of information contained in these observations in principle is enough to allow us to fine-tune more labor-intensive experiments to test critical questions with great efficiency. In practice, though, much of this information remains inaccessible. In order to

bring about a future of precision medicine and precision ecological remediation, where we can specify precise microbiome changes and bring them about through defined interventions, a vast amount of non-hypothesis-driven research, often dismissed as "technical work" or "fishing expeditions", remains to be done.

1.1.6 Acknowledgements

This work was supported in part by the National Institute of Justice Award 2015-DN-BX-K047, by the Alfred P. Sloan Foundation, and by the National Institutes of Health.

1.1.7 Author contributions

Chapter 1, Introduction part 1, in full, is a reprint of previously published material: Tripathi A, Marotz C, Gonzalez A, Vázquez-Baeza Y, Song SJ, Bouslimani A, McDonald D, Zhu Q, Sanders JG, Smarr L, Dorrestein PC & Knight, R. *Are microbiome studies ready for hypothesis-driven research?* Current opinion in microbiology. 2018 Aug 1;44:61-9. The co-authors listed above supervised or provided support for the research and have given permission for the inclusion of the work in this dissertation.

1.1.8 References

1. Ruse M: The Darwinian Revolution: Science Red in Tooth and Claw. Chicago: University of Chicago Press; 1999.

2. Cohen IB: The First English Version of Newton's Hypotheses non fingo. Isis 1962, 53:379-388.

3. Anderson C: THE END OF THEORY: THE DATA DELUGE MAKES THE SCIENTIFIC METHOD OBSOLETE. In Wired. Edited by. San Francisco: Conde Nast; 2008.

4. Mazzocchi F: Could Big Data be the end of theory in science? A few remarks on the epistemology of data-driven science. EMBO Rep 2015, 16:1250-1255.

5. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA, Biddinger SB, Dutton RJ, Turnbaugh PJ. 2014. Diet rapidly and reproducibly alters the human gut microbiome. Nature 505:559–563.

6. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen Y-Y, Keilbaugh SA, Bewtra M, Knights D, Walters WA, Knight R, Sinha R, Gilroy E, Gupta K, Baldassano R, Nessel L, Li H, Bushman FD, Lewis JD. 2011. Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. Science.

7. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R: Bacterial community variation in human body habitats across space and time. Science 2009, 326:1694-1697.

8. Human Microbiome Project C: Structure, function and diversity of the healthy human microbiome. Nature 2012, 486:207-214.

9. Davenport ER, Mizrahi-Man O, Michelini K, Barreiro LB, Ober C, Gilad Y: Seasonal variation in human gut microbiome composition. PLoS One 2014, 9:e90731.

10. Smits SA, Leach J, Sonnenburg ED, Gonzalez CG, Lichtman JS, Reid G, Knight R, Manjurano A, Changalucha J, Elias JE, Dominguez-Bello MG, Sonnenburg JL. 2017. Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. Science 357:802–806.

11. Lozupone CA, Knight R: Global patterns in bacterial diversity. Proc Natl Acad Sci U S A 2007, 104:11436-11440.

12. Thompson LR, The Earth Microbiome Project Consortium, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, Navas-Molina JA, Janssen S, Kopylova E, Vázquez-Baeza Y, González A, Morton JT, Mirarab S, Xu ZZ, Jiang L, Haroon MF, Kanbar J, Zhu Q, Song SJ, Kosciolek T, Bokulich NA, Lefler J, Brislawn CJ, Humphrey G, Owens SM, Hampton-Marcell J, Berg-Lyons D, McKenzie V, Fierer N, Fuhrman JA, Clauset A, Stevens RL, Shade A, Pollard KS, Goodwin KD, Jansson JK, Gilbert JA, Knight R. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature.

13. Ley RE, Lozupone CA, Hamady M, Knight R, Gordon JI: Worlds within worlds: evolution of the vertebrate gut microbiota. Nat Rev Microbiol 2008, 6:776-788.

14. McCafferty J, Muhlbauer M, Gharaibeh RZ, Arthur JC, Perez-Chanona E, Sha W, Jobin C, Fodor AA: Stochastic changes over time and not founder effects drive cage effects in microbial community assembly in a mouse model. ISME J 2013, 7:2116-2125.

15. Debelius J, Song SJ, Vazquez-Baeza Y, Xu ZZ, Gonzalez A, Knight R: Tiny microbes, enormous impacts: what matters in gut microbiome studies? Genome Biol 2016, 17:217.

16. da Silva RR, Dorrestein PC, Quinn RA: Illuminating the dark matter in metabolomics. Proc Natl Acad Sci U S A 2015, 112:12549-12550.

17. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T, Porto C, Bouslimani A, Melnik AV, Meehan MJ, Liu W-T, Crüsemann M, Boudreau PD, Esquenazi E, Sandoval-Calderón M, Kersten RD, Pace LA, Quinn RA, Duncan KR, Hsu C-C, Floros DJ, Gavilan RG, Kleigrewe K, Northen T, Dutton RJ, Parrot D, Carlson EE, Aigle B, Michelsen CF, Jelsbak L, Sohlenkamp C, Pevzner P, Edlund A, McLean J, Piel J, Murphy BT, Gerwick L, Liaw C-C, Yang Y-L, Humpf H-U, Maansson M, Keyzers RA, Sims AC, Johnson AR, Sidebottom AM, Sedio BE, Klitgaard A, Larson CB, P CAB, Torres-Mendoza D, Gonzalez DJ, Silva DB, Marques LM, Demarque DP, Pociute E, O'Neill EC, Briand E, Helfrich EJN, Granatosky EA, Glukhov E, Ryffel F, Houson H, Mohimani H, Kharbush JJ, Zeng Y, Vorholt JA, Kurita KL, Charusanti P, McPhail KL, Nielsen KF, Vuong L, Elfeki M, Traxler MF, Engene N, Koyama N, Vining OB, Baric R, Silva RR, Mascuch SJ, Tomasi S, Jenkins S, Macherla V, Hoffman T, Agarwal V, Williams PG, Dai J, Neupane R, Gurr J, Rodríguez AMC, Lamsa A, Zhang C, Dorrestein K, Duggan BM, Almaliti J, Allard P-M, Phapale P, Nothias L-F, Alexandrov T, Litaudon M, Wolfender J-L, Kyle JE, Metz TO, Peryea T, Nguyen D-T, VanLeer D, Shinn P, Jadhav A, Müller R, Waters KM, Shi W, Liu X, Zhang L, Knight R, Jensen PR, Palsson BO, Pogliano K, Linington RG, Gutiérrez M, Lopes NP, Gerwick WH, Moore BS, Dorrestein PC, Bandeira N. 2016. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol 34:828–837.

18. Gonzalez A, Knight R: Advancing analytical algorithms and pipelines for billions of microbial sequences. Curr Opin Biotechnol 2012, 23:64-71.

19. Bouslimani A, Melnik AV, Xu Z, Amir A, da Silva RR, Wang M, Bandeira N, Alexandrov T, Knight R, Dorrestein PC: Lifestyle chemistries from phones for individual profiling. Proc Natl Acad Sci U S A 2016, 113:E7645-E7654.

20. Bouslimani A, Porto C, Rath CM, Wang M, Guo Y, Gonzalez A, Berg-Lyon D, Ackermann G, Moeller Christensen GJ, Nakatsuji T, Zhang L, Borkowski AW, Meehan MJ, Dorrestein K, Gallo RL, Bandeira N, Knight R, Alexandrov T, Dorrestein PC. 2015. Molecular cartography of the human skin surface in 3D. Proc Natl Acad Sci U S A 112:E2120–9.

21. Garg N, Wang M, Hyde E, da Silva RR, Melnik AV, Protsyuk I, Bouslimani A, Lim YW, Wong R, Humphrey G, Ackermann G, Spivey T, Brouha SS, Bandeira N, Lin GY, Rohwer F, Conrad DJ, Alexandrov T, Knight R, Dorrestein PC. 2017. Three-Dimensional Microbiome and Metabolome Cartography of a Diseased Human Lung. Cell Host Microbe 22:705–716.e4.

22. Kapono CA, Morton JT, Bouslimani A, Melnik AV, Orlinsky K, Knaan TL, Garg N, Vázquez-Baeza Y, Protsyuk I, Janssen S, Zhu Q, Alexandrov T, Smarr L, Knight R, Dorrestein PC. 2018. Creating a 3D microbial and chemical snapshot of a human habitat. Sci Rep 8:3669.

23. Petras D, Nothias L-F, Quinn RA, Alexandrov T, Bandeira N, Bouslimani A, Castro-Falcón G, Chen L, Dang T, Floros DJ, Hook V, Garg N, Hoffner N, Jiang Y, Kapono CA, Koester I, Knight R, Leber CA, Ling T-J, Luzzatto-Knaan T, McCall L-I, McGrath AP, Meehan MJ, Merritt JK, Mills RH, Morton J, Podvin S, Protsyuk I, Purdy T, Satterfield K, Searles S, Shah S, Shires S, Steffen D, White M, Todoric J, Tuttle R, Wojnicz A, Sapp V, Vargas F, Yang J, Zhang C, Dorrestein PC. 2016. Mass Spectrometry-Based Visualization of Molecules Associated with Human Habitats. Anal Chem 88:10775–10784.

24. Protsyuk I, Melnik AV, Nothias LF, Rappez L, Phapale P, Aksenov AA, Bouslimani A, Ryazanov S, Dorrestein PC, Alexandrov T: 3D molecular cartography using LC-MS facilitated by Optimus and 'ili software. Nat Protoc 2018, 13:134-154.

25. Paneth N, Vinten-Johansen P, Brody H, Rip M: A rivalry of foulness: official and unofficial investigations of the London cholera epidemic of 1854. Am J Public Health 1998, 88:1545-1553.

26. Weingarden A, González A, Vázquez-Baeza Y, Weiss S, Humphry G, Berg-Lyons D, Knights D, Unno T, Bobr A, Kang J, Khoruts A, Knight R, Sadowsky MJ. 2015. Dynamic changes in short- and long-term bacterial composition following fecal microbiota transplantation for recurrent Clostridium difficile infection. Microbiome 3:10.

27. Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N: Microbial strain-level population structure and genetic diversity from metagenomes. Genome Res 2017, 27:626-638.

28. Marotz CA, Sanders JG, Zuniga C, Zaramela LS, Knight R, Zengler K: Improving saliva shotgun metagenomics by chemical host DNA depletion. Microbiome 2018, 6:42.

29. Aksenov AA, da Silva R, Knight R, Lopes NP, Dorrestein PC: Global chemical analysis of biology by mass spectrometry. Nat Rev Chem 2017, 1:0054.

30. Sauerschnig C, Doppler M, Bueschl C, Schuhmacher R: Methanol Generates Numerous Artifacts during Sample Extraction and Storage of Extracts in Metabolomics Research. Metabolites 2017, 8.

31. Scheubert K, Hufsky F, Petras D, Wang M, Nothias LF, Duhrkop K, Bandeira N, Dorrestein PC, Bocker S: Significance estimation for large scale metabolomics annotations by spectral matching. Nat Commun 2017, 8:1494.

32. Members: MB, MSI Board Members: 2007. The Metabolomics Standards Initiative. Nature Biotechnology.

33. Salek RM, Conesa P, Cochrane K, Haug K, Williams M, Kale N, Moreno P, Jayaseelan KV, Macias JR, Nainala VC, Hall RD, Reed LK, Viant MR, O'Donovan C, Steinbeck C. 2017. Automated assembly of species metabolomes through data submission into a public repository. Gigascience 6:1–4.

34. Haug K, Salek RM, Conesa P, Hastings J, de Matos P, Rijnbeek M, Mahendraker T, Williams M, Neumann S, Rocca-Serra P, Maguire E, González-Beltrán A, Sansone S-A, Griffin JL, Steinbeck C. 2013. MetaboLights--an open-access general-purpose repository for metabolomics studies and associated meta-data. Nucleic Acids Res 41:D781–6.

35. Sud M, Fahy E, Cotter D, Azam K, Vadivelu I, Burant C, Edison A, Fiehn O, Higashi R, Nair KS, et al.: Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. Nucleic Acids Res 2016, 44:D463-470.

36. Nicholson JK, Lindon JC: Systems biology: Metabonomics. Nature 2008, 455:1054-1056.

37. Muegge BD, Kuczynski J, Knights D, Clemente JC, Gonzalez A, Fontana L, Henrissat B, Knight R, Gordon JI: Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. Science 2011, 332:970-974.

38. Minich JJ, Zhu Q, Janssen S, Hendrickson R, Amir A, Vetter R, Hyde J, Doty MM, Stillwell K, Benardini J, Kim JH, Allen EE, Venkateswaran K, Knight R. 2018. KatharoSeq Enables High-Throughput Microbiome Analysis from Low-Biomass Samples. mSystems 3.

39. Gika HG, Theodoridis GA, Wilson ID: Liquid chromatography and ultra-performance liquid chromatography-mass spectrometry fingerprinting of human urine: sample stability under different handling and storage conditions for metabonomics studies. J Chromatogr A 2008, 1189:314-322.

40. Lou JJ, Mirsadraei L, Sanchez DE, Wilson RW, Shabihkhani M, Lucey GM, Wei B, Singer EJ, Mareninov S, Yong WH: A review of room temperature storage of biospecimen tissue and nucleic acids for anatomic pathology laboratories and biorepositories. Clin Biochem 2014, 47:267-273.

41. Song SJ, Amir A, Metcalf JL, Amato KR, Xu ZZ, Humphrey G, Knight R: Preservation Methods Differ in Fecal Microbiome Stability, Affecting Suitability for Field Studies. mSystems 2016, 1.

42. Choo JM, Leong LE, Rogers GB: Sample storage conditions significantly influence faecal microbiome profiles. Sci Rep 2015, 5:16350.

43. Hale VL, Tan CL, Niu K, Yang Y, Cui D, Zhao H, Knight R, Amato KR: Effects of field conditions on fecal microbiota. J Microbiol Methods 2016, 130:180-188.

44. Vogtmann E, Chen J, Amir A, Shi J, Abnet CC, Nelson H, Knight R, Chia N, Sinha R: Comparison of Collection Methods for Fecal Samples in Microbiome Studies. Am J Epidemiol 2017, 185:115-123.

45. Loftfield E, Vogtmann E, Sampson JN, Moore SC, Nelson H, Knight R, Chia N, Sinha R: Comparison of Collection Methods for Fecal Samples for Discovery Metabolomics in Epidemiologic Studies. Cancer Epidemiol Biomarkers Prev 2016, 25:1483-1490.

46. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA, 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 BW, Blaser MJ, Bonazzi V, Booth T, Bork P, Bushman FD, Buttigieg PL, Chain PSG, Charlson E, Costello EK, Huot-Creasy H, Dawyndt P, DeSantis T, Fierer N, Fuhrman JA, Gallery RE, Gevers D, Gibbs RA, San Gil I, Gonzalez A, Gordon JI, Guralnick R, Hankeln W, Highlander S, Hugenholtz P, Jansson J, Kau AL, Kelley ST, Kennedy J, Knights D, Koren O, Kuczynski J, Kyrpides N, Larsen R, Lauber CL, Legg T, Ley RE, Lozupone CA, Ludwig W, Lyons D, Maguire E, Methé BA, Meyer F, Muegge B, Nakielny S, Nelson KE, Nemergut D, Neufeld JD, Newbold LK, Oliver AE, Pace NR, Palanisamy G, Peplies J, Petrosino J, Proctor L, Pruesse E, Quast C, Raes J, Ratnasingham S, Ravel J, Relman DA, Assunta-Sansone S, Schloss PD, Schriml L, Sinha R, Smith MI, Sodergren E, Spo A, Stombaugh J, Tiedje JM, Ward DV, Weinstock GM, Wendel D, White O, Whiteley A, Wilke A, Wortman JR, Yatsunenko T, Glöckner FO. 2011. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol 29:415–420.

47. Yutin N, Makarova KS, Gussow AB, Krupovic M, Segall A, Edwards RA, Koonin EV: Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. Nat Microbiol 2018, 3:38-46.

1.2. The gut-liver axis and the intersection with the microbiome

In the past decade, an exciting realization has been that diverse liver diseases—ranging from NASH, alcoholic steatohepatitis and cirrhosis, to hepatocellular carcinoma—fall along a spectrum. Work on the biology of the gut-liver axis has assisted in understanding the basic biology of both alcoholic fatty liver disease and NAFLD. Of immense importance is the advancement in understanding of the role of the microbiome, driven by high-throughput DNA sequencing and improved computational techniques that enable the complexity of the microbiome to be interrogated, together with improved experimental designs. Here, we review gut–liver communications in liver disease, explore the molecular, genetic and microbiome relationships, and discuss prospects for exploiting the microbiome to determine liver disease stage and to predict the effects of pharmaceutical, dietary and other interventions at a population and individual level. Although much work remains to be done in understanding the relationship between the microbiome and liver disease, rapid progress towards clinical applications is being made, especially in study designs that complement human intervention studies with mechanistic work in mice that have been humanized in multiple respects, including the genetic, immunological and microbiome characteristics of individual patients. These 'avatar mice' could be especially useful for guiding new microbiome-based or microbiome-informed therapies.

1.2.1 Introduction

The crosstalk between the gut and liver is increasingly recognized, strengthened by the parallel rise in incidence of liver diseases and gastrointestinal and immune disorders (1, 2). The most common type of liver disease, NAFLD, affects >65 million Americans with a cost burden of US$103 billion annually within the USA (3). To manage the socioeconomic burden of

gastrointestinal-associated liver diseases by developing new therapeutic modalities, specific molecular events that facilitate interaction between the gut and the liver must be elucidated. As we begin to appreciate these links, animal models (4–6) and well-designed clinical studies (7–9) are already revealing key components of these interactions.

The present understanding of the etiology of the spectrum of liver diseases (Figure 1.2.1) is underpinned by proinflammatory changes in the host. Intestinal dysbiosis (anomalous or imbalanced gut microbial composition) and increased intestinal permeability leads to translocation of microbes and microbial products including cell wall components (endotoxins from Gram-negative bacteria, β-glucan from fungi) and DNA, together referred to as microbial-(or pathogen-) associated molecular patterns (MAMPs or PAMPs). These patterns are recognized by immune receptors on liver cells (such as Kupffer cells and hepatic stellate cells) and intestinal lamina propria (an immune cell-rich tissue beneath the intestinal epithelium), which initiate and maintain inflammatory cascades that ultimately lead to liver damage in the form of fibrosis (10–13). This damage can progress from cirrhosis (severe fibrosis) to hepatocellular carcinoma (HCC), the most predominant form (>80%) of primary liver cancer (14). Previously demonstrated associations between intestinal health and several different types of neoplasia suggest a potential role of the microbiota in HCC (15, 16). Additionally, the liver and microbiota engage in co-metabolism of xenobiotics including carcinogens which can independently predispose the host to HCC (17, 18).

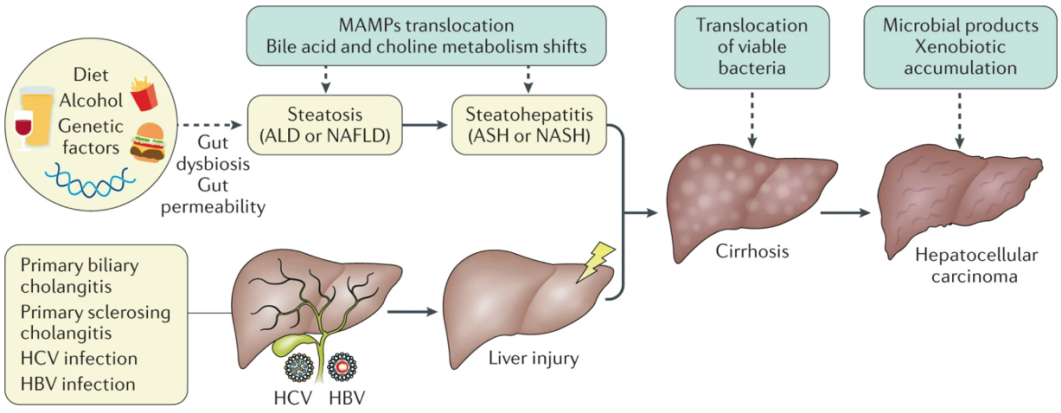

Figure 1.2.1 *Physiological manifestations of liver injury along a spectrum of progression*. Risk factors such as alcohol abuse, unbalanced diet, infection (HBV or HCV) or immune dysfunction (PBC/PSC) can independently lead to liver injury. Individuals who abuse alcohol and individuals with obesity often develop steatosis (fatty liver), which is characterized by increased intestinal permeability and dysbiosis. Subsequently, bile acid and choline homeostasis is disturbed along with increased translocation of MAMPs across the gut-barrier, leading to steatohepatitis, the progressive form of liver damage. Both, steatosis-dependent and steatosis-independent liver damage can progress to cirrhosis (end-stage liver damage), which is marked by translocation of viable bacteria to the liver and severe inflammation. As liver function is progressively compromised, tumor-promoting metabolites and xenobiotics accumulate. These could activate oncogenic pathways causing hepatocellular carcinoma, the most predominant form of primary liver cancers.
(MAMPs: Microbial-associated molecular patterns; ALD: Alcoholic liver disease; NAFLD: Nonalcoholic fatty liver disease; ASH: Alcoholic steatohepatitis; NASH: Nonalcoholic steatohepatitis; HBV: Hepatitis B virus; HCV: Hepatitis C virus; PSC: Primary sclerosing cholangitis; PBC: Primary biliary cholangitis)

The missing links in the complex interaction network between host and microorganisms are being discovered piece by piece using various experimental designs (detailed later). These findings encourage microbiome-oriented therapeutic modalities to treat liver-associated and other metabolic diseases. Here, we review the current understanding of the aetiology of liver diseases and highlight the open research questions (Box 1.2.1) to motivate focused research in this area with special attention to the role of the microbiome.

1.2.2 How do the liver and gut communicate?

The gut and liver communicate via tight bidirectional links through the biliary tract, portal vein and systemic circulation (Figure 1.2.2). The liver communicates with the intestine by

releasing bile acids and many bioactive mediators into the biliary tract and the systemic circulation. In the intestine, host and microbes metabolize endogenous (bile acids, amino acids) as well as exogenous (from diet and environmental exposure) substrates, the products of which translocate to the liver through the portal vein and influence liver functions (19). Some crucial links between the gut and liver are discussed herein.

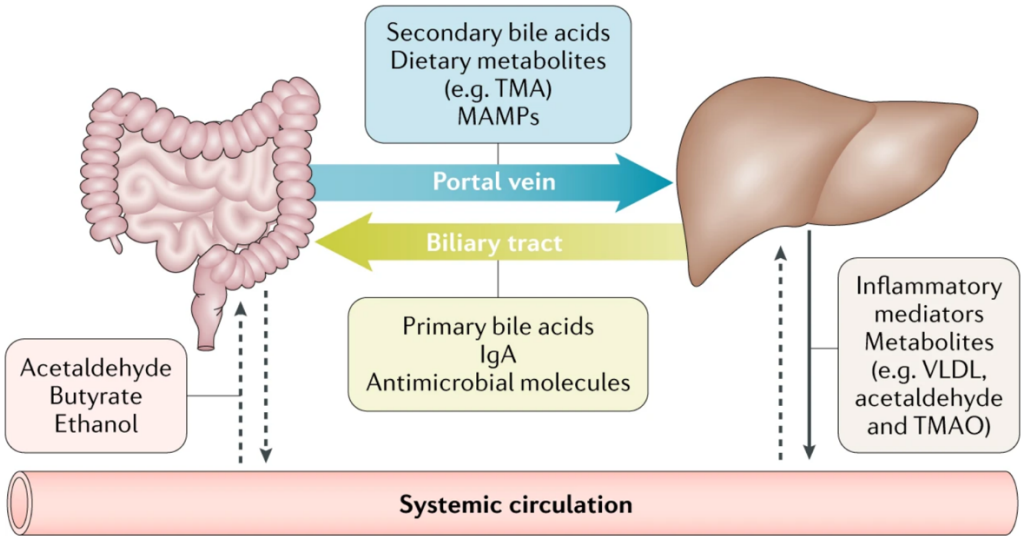

Figure 1.2.2 *Bidirectional communication between gut and liver*. The liver transports bile salts and antimicrobial molecules (IgA, angiogenin 1) to the intestinal lumen through the biliary tract. This process maintains gut eubiosis by controlling unrestricted bacterial overgrowth. Bile salts also act as important signaling molecules via nuclear receptors (such as FXR, TGR5) to modulate hepatic bile acid synthesis, glucose metabolism, lipid metabolism and energy utilization from diet. On the other hand, gut-products such as host and/or microbial metabolites and MAMPs translocate to the liver via the portal vein and influence liver functions. Additionally, systemic circulation extends the gut–liver axis by transporting liver metabolites from dietary, endogenous or xenobiotic substances (for example, FFAs, choline metabolites, ethanol metabolites) to the intestine through the capillary system. Owing to this medium of transport and ease of diffusion of systemic mediators across blood capillaries, these factors could affect the intestinal barrier both, positively (for example, butyrate) or negatively (for example, acetaldehyde)
(TMA: Trimethylamine; TMAO: Trimethylamine N-oxide; MAMPs: Pathogen-associated molecular patterns; VLDL: Very low-density lipoprotein; FXR: Farnesoid X receptor; TGR5: Takeda G-protein coupled receptor 5; FFA: Free fatty acid)

Enterohepatic circulation of bile acids: Bile acids (BAs) are amphipathic molecules synthesized from cholesterol in the pericentral hepatocytes. These primary BAs are conjugated to glycine or taurine and released in the biliary tract. On reaching the small intestine through the

duodenum, BAs, together with other biliary components, facilitate emulsification and absorption of dietary fats, cholesterol and fat-soluble vitamins. About 95% of the BAs are actively reabsorbed in the terminal ileum and transported back to the liver (20, 21). The remaining 5% are deconjugated, dehydrogenated and dehydroxylated by the colonic microbiota to form secondary bile acids, which reach the liver via passive absorption into the portal circulation (22). The liver recycles BAs and secretes them back to the biliary tract completing the so-called enterohepatic circulation, that is, a system of exchange between the gut and the liver.

A carrier-mediated process transports hydrophilic primary BAs across cell membranes for uptake into intestinal epithelial cells. Regulatory effects of BAs have been best studied with respect to farnesoid X receptor (FXR) and takeda G-protein-coupled receptor 5 (TGR5). BAs bind to FXR in the enterocytes and induce transcription of an enterokine, fibroblast growth factor 19 (FGF19; FGF15 in mouse). FGF19 reaches the liver through the portal vein and downregulates *de novo* bile acid synthesis by inhibiting CYP7A1 in hepatocytes, forming a feedback system for modulating BA production (23). FXR activation is known to affect glucose and lipid metabolism (24, 25). Additionally, BAs bind to TGR5 on the plasma membrane and act on tissues beyond enterohepatic circulation. This binding mediates host energy expenditure (26, 27), glucose homeostasis(28) and anti-inflammatory immune responses (29, 30).

BAs and the gut microbiota closely interact and modulate each other; BAs exert direct control on the intestinal microbiota. By binding to FXR, they induce production of antimicrobial peptides such as angogenin1 and RNAse family member 4, which are directly involved in inhibiting gut microbial overgrowth and subsequent gut barrier dysfunction (31, 32). Intestinal dysbiosis shifts the balance between primary and secondary bile acids and their subsequent enterohepatic cycling, the metabolic effects of which are not comprehensively understood.

However, because of differences in the affinity of these two classes of BAs for FXR, these shifts have been associated with changes in hepatic bile acid synthesis and metabolic stress (22, 33–35). An imbalance in BAs and gut bacteria elicits a cascade of host immune responses relevant to the progression of liver diseases.

Intestinal permeability: The central components of the intestinal barrier are enterocytes that are tightly bound to adjacent cells by apical junctional proteins that include claudins, occludins, E-cadherins, desmosomes, and junctional adhesion molecules (36). This barrier restricts movement of microbes and molecules from the gut lumen, while allowing permselective, active transport of nutrients across the tight junctions. The intestinal barrier is further strengthened by several additional lines of defense. Mucins (heavily glycosylated protein aggregates) form a physical barrier between luminal bacteria and the underlying epithelial layer (37), and antibacterial lectins, such as regenerating islet-derived protein III-gamma (REG3G), which are produced by intestinal Paneth cells to target bacteria associated with mucosal lining (38, 39). Moreover, immunoglobulins (specifically secretory IgA) produced by plasma cells and transported into the lumen through the intestinal epithelial cells neutralize microbial pathogens by blockading epithelial receptors (40). Finally, commensal bacteria are closely associated with the gut mucosa, and reinforce barrier integrity by stimulating cell-mediated immunity via Toll-like receptor-mediated signaling (38, 41) or by producing metabolites that directly strengthen tight junctions (short-chain fatty acids or SCFAs) (42–44) and inhibit other microbes (45–47).

Breakdown of one or more of these barrier components compromises gut-barrier integrity. The major drivers of increased permeability include gut inflammation and dysbiosis (48, 49), which have been linked to consumption of a high-fat Western diet (50–52), chronic alcohol consumption (53–55), prolonged antibiotic usage (56) and immune-mediated inflammatory

diseases such as IBD (57). An important association between the gut microbiota, inflammation and gut-barrier integrity is provided by *Akkermansia muciniphila*, a Gram-negative anaerobe that colonizes the intestinal mucus layer. Reduced abundance of *A. muciniphila* has been associated with thinning of mucus layer and increased inflammation, which promotes both, alcoholic and nonalcoholic liver damage (58, 59). When the gut barrier is compromised, microbes and microbe-derived molecules can translocate to the liver through the portal system, causing inflammation and hepatic injury (13). Some translocated intestinal products might also directly interact with host factors and contribute to exacerbation of liver disease (Figure 1.2.3) (60–65).

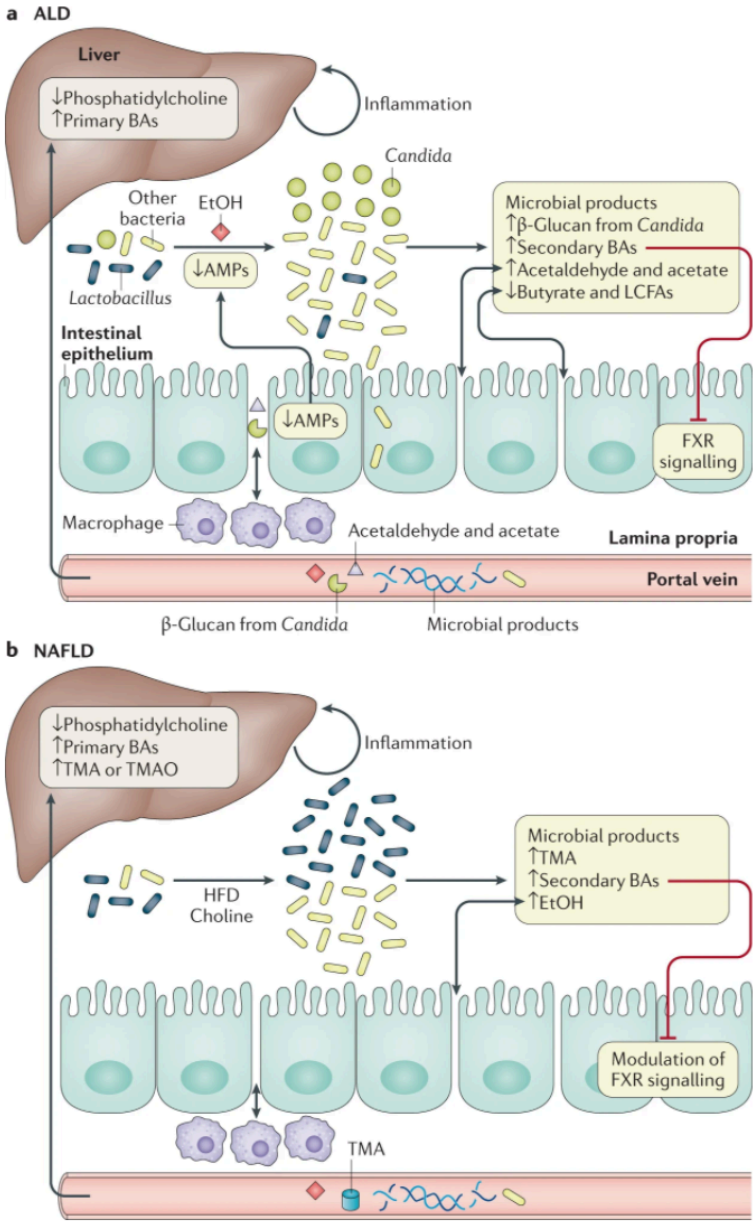

Figure 1.2.3 *Interplay between the liver and gut microbiome in alcoholic liver disease and NAFLD*. Intestinal dysbiosis and bacterial overgrowth is observed in both ALD **(a)** and NAFLD **(b)**. Bacterial overgrowth causes an increase in secondary BAs, which disrupts FXR-mediated modulation of BA levels, leading to an overall increase in hepatic BA synthesis. A reduction in hepatic phosphatidylcholine is also seen in both ALD and NAFLD, which causes triglyceride accumulation in the liver (fatty liver). While ALD-associated dysbiosis is characterized by reduction in *Lactobacillus* and *Candida* overgrowth, patients with NAFLD have higher abundance of *Lactobacillus* (effects on fungal population remain to be investigated). In ALD and NAFLD, increased ethanol and its metabolite acetaldehyde in the intestinal lumen mediates weakening of intestinal tight junctions. Consequently, increased translocation of MAMPs (seen in ALD and NAFLD) and gut metabolites such as acetaldehyde, acetate (seen in ALD) and TMA (seen in NAFLD) elicits intestinal and hepatic inflammatory responses, leading to progressive liver damage.
(AMP: Antimicrobial peptides; BA: Bile acids; EtOH: Ethanol; FXR: Farnesoid X receptor; HFD: High-fat diet; LCFA: Long-chain fatty acids; TMA: Trimethylamine; TMAO: Trimethylamine N-oxide)

Systemic circulation: Bacteria and MAMPs: Intestinal permeability is characterized by compromised tight junctions between enterocytes, and is consistently seen across the spectrum of liver diseases (66, 67). Liver damage is associated with small intestinal bacterial overgrowth (SIBO) and dysbiosis of the lower gastrointestinal tract (68). Together, these processes lead to increased translocation of MAMPs into the portal circulation. On reaching the liver, MAMPs induce localized inflammation through pattern-recognition receptors (PRRs) on Kupffer cells (69) and hepatic stellate cells (70, 71). Endotoxin-mediated activation of Toll-like receptor (TLR)4 (69, 70) along with TLR9 (activated by methylated DNA)(71) and TLR2 (activated by Gram-positive bacteria) (72) are the primary drivers of immune response in liver disease. TLR signaling in Kupffer cells activates a downstream proinflammatory cascade, leading to MyD88-mediated activation of NF-kB (13). Additionally, TLR4 signaling also promotes fibrosis by downregulating Bambi (a decoy receptor for TGF-β) in hepatic stellate cells (13). These steps lead to expression of inflammatory cytokines, oxidative and endoplasmic reticulum (ER) stress, and subsequent liver damage (73).

Choline metabolites: Choline is a macronutrient that is important for liver function, brain development, nerve function, muscle movement and maintaining a healthy metabolism (74); notably, rodents fed a choline-deficient diet have been used to model human NASH (75–77). Choline is processed into phosphatidylcholine (lecithin) by the host, which assists in excretion of very-low-density lipoproteins (VLDL) particles from the liver. This process prevents hepatic accumulation of triglycerides (liver steatosis) (78). Additionally, choline can also be converted to trimethylamine (TMA) by intestinal bacteria; TMA can translocate to the liver through the portal circulation where it is converted to trimethylamine N-oxide (TMAO) (79). The significance of methylamines is increasingly being recognized with respect to liver, cardiometabolic and more

recently, neurological disorders (79, 80). Increased systemic circulation of TMAO is concomitant with reduced levels of host-produced phosphatidylcholine, an imbalance characteristic of intestinal dysbiosis. TMAO has been linked with liver damage due to increased triglyceride accumulation (hepatic steatosis) (9, 79, 81–83) and, consequently, NAFLD (9).

Free fatty acids: Free fatty acids include SCFAs and long-chain fatty acids (LCFAs). Butyrate, propionate (produced by bacterial fermentation) and acetate (produced by both host and bacteria) are the dominant SCFAs in the large intestine. Butyrate is an energy source for the enterocytes and facilitates maintenance of the intestinal barrier (42–44). Alcohol-induced liver injury is suggested to be marked by reduced levels of butyrate and propionate (84, 85) and increased levels of acetate (possibly produced by ethanol metabolism in the lumen, but predominantly derived from ethanol metabolism in the liver). Increased levels of acetaldehyde can weaken gut barrier [86] and induce hepatic stress on translocation of intestinal antigens to the liver (87, 88). Butyrate supplementation in the form of a glycerol ester, tributyrin, reduced ethanol-induced intestinal permeability and subsequent liver injury in mice on a short-term alcohol diet (85). However, how tributyrin mechanistically protects the intestinal barrier remains to be established.

Luminal species of LCFAs include pentadecanoic acid (C15:0), palmitic acid (C16:0), heptadecanoic acid (C17:0), and stearic acid (C18:0). In mice fed alcohol chronically, C15:0 and C17:0, which are only produced by bacterial fermentation, are markedly reduced when compared with control mice on isocaloric diet (84, 89). There is also an overall reduction in total saturated LCFA levels which is associated with decreased luminal abundance of lactobacilli (known metabolizers of saturated LCFAs) (84). To our knowledge, restoring *Lactobacillus spp*. by LCFA supplementation has not been experimentally demonstrated. However, dietary supplementation of

Lactobacillus rhamnosus has been shown to increase luminal LCFA levels (89), suggesting that *Lactobacillus*-induced increase in intestinal FFAs contribute to its probiotic effects (90–96).

Ethanol and acetaldehyde: The mucosa of the gastrointestinal tract absorbs ethanol by simple diffusion. Within the gastrointestinal tract, the majority of ethanol from food and beverages is absorbed by the stomach (~20%) and small intestine (~70%) (97, 98). Although, microbial fermentation contributes to luminal ethanol concentration, the biggest share of alcohol in the large intestine comes from the systemic circulation (13).

Gut microbiota and enterocytes express alcohol-metabolizing enzymes such as alcohol dehydrogenase, which co-metabolizes ethanol into acetaldehyde and, to a lesser-studied extent, acetate. The liver also responds to circulating levels of ethanol by upregulating its ethanol metabolism pathway (87, 88). The importance of microbes for xenobiotics metabolism was underscored by a study that demonstrated an increase in hepatic expression of ethanol metabolizing genes in germ-free mice, and exacerbation in hepatic steatosis (87).

Non-alcoholic and alcoholic liver diseases (Table 1.2.1) are characterized by increased luminal and circulating levels of ethanol and its metabolites, acetaldehyde and acetate (65, 99). These metabolites have independently been associated with liver damage (62–64). Acetaldehyde has been implicated in weakening the intestinal tight junctions, compromising the gut barrier and enabling translocation of microbial products (100–105). It has also been associated with downregulating the expression of antimicrobial peptides (AMPs) in the intestine (106, 107), and eliciting inflammatory and adaptive host immune responses (108–110). Additionally, alcoholic liver disease (ALD) is marked by reduced levels of intestinal butyrate (84, 111, 112) (an energy source for enterocytes), which is linked to weakening of intestinal tight junctions and, hence, permeability (85, 113–115).

1.2.3 Microbiome and specific liver disease

NAFLD: NAFLD refers to a spectrum of liver disease that can be broadly classified into two categories: nonalcoholic fatty liver (NAFL), the non-progressive form of NAFLD, and NASH, the progressive form of NAFLD (116). NASH is generally linked to type 2 diabetes mellitus, cardiovascular risk factors and obesity (117, 118), although NAFLD has also been reported in lean individuals, emphasizing that genetic and environmental factors also contribute to disease development (119–122).

Several studies have stressed the role of the gut microbiota in NAFLD but causality is yet to be established (123). Patients with NAFLD have a higher prevalence of SIBO (66, 124) and microbial dysbiosis (125). Using 16S amplicon sequencing, Boursier et al.(125) found that the bacterial genera, *Bacteroides* and *Ruminococcus* were substantially increased, and *Prevotella* was reduced in patients with NASH (stage 2 fibrosis or higher) compared to those without NASH. Loomba et al. (7) utilized whole-genome metagenomics to characterize the gut microbiota in patients with NAFLD with and without advanced fibrosis (stages 3 and 4) and showed an increased abundance of *Escherichia coli* and *Bacteroides vulgatus* in patients with advanced fibrosis. An enrichment of *Escherichia* (genera) was also seen in paediatric patients with NASH compared with children with obesity but without NASH (65). Consistent with preclinical studies, these studies indicate an association between Gram-negative bacteria and progression of liver fibrosis (126).

Genetically modified mouse models have been used to study NAFLD-associated gut dysbiosis and permeability for mechanistic insights in liver disease progression. Rahman et al. (127) used JAM-A (junctional adhesion molecule-A protein)-knockout mice to demonstrate that deficiency in this tight junction protein with a diet rich in saturated fats, fructose and cholesterol leads to increased intestinal permeability and liver inflammation. This inflammation could be

alleviated by administering antibiotics, underscoring the importance of microbial translocation in promoting immune response in the liver. Another group used mice deficient in *Muc2* (predominant mucin in the intestinal mucus layer) and found that there was a compensatory increase in intestinal levels of antimicrobial protein-coding genes, *Reg3b* and leading to an overall protective response against NAFLD (107).

The contribution of liver-damaging inflammation in response to translocation of microbes and MAMPs has been elucidated (49). Using inflammasome-deficient mouse models (NLRP3-/- or NLRP6-/-), Henao-Mejia et al. conclude that there is an accumulation of MAMPs in portal circulation, which enhanced the expression of hepatic TNF, thereby promoting liver inflammation and NASH progression. Furthermore, cohousing inflammasome-deficient mice with wild-type controls exacerbated diet-induced hepatic steatosis and obesity in healthy cage mates, suggesting transferability of disease via the microbiota.

Increasing links between NAFLD and the gut microbiome at both the observational and mechanistic levels make the gut microbiota an attractive source of biomarkers for early diagnosis of NAFLD. In a comparison between children with obesity with and without NASH, Zhu and colleagues (65) observed markedly elevated gut microbial production of ethanol in those with NASH. Adults with NAFLD also show increased serum TMAO (9) and hepatic bile acid synthesis (35), and decreased production of phosphatidylcholine (128). Furthermore, Loomba et al. observed differences in carbon and amino acid metabolism in gut microbiome of patients with NAFLD-associated advanced fibrosis (7). This proof-of-concept study provides preliminary evidence to support the utility of a microbiome-derived metagenomics signature to detect advanced fibrosis as well as candidacy for anti-fibrotic treatment trials in NAFLD.

Alcoholic liver disease: The manifestation of ALD in patients who chronically abuse alcohol is a consequence of multifactorial interactions involving genetics, immune system, gut microbiome and environmental factors (100, 129–131). Like NAFLD, the non-progressive form of ALD is characterized by accumulation of fat inside the liver (fatty liver or steatosis), whereas the progressive form is marked by inflammation and liver injury (alcoholic steatohepatitis or ASH).

Our understanding of the compositional and mechanistic contributions of the gut microbiota in ALD is improving. As in NAFLD, SIBO has been demonstrated as an important hallmark of alcohol-associated liver disease in humans (35) and mouse models (106, 131). Intestinal dysbiosis in individuals who abuse alcohol is characterized by marked enrichment of Enterobactericaeae (family) and reduction in abundances of *Bacteroidetes* and *Lactobacillus* (genera) (106, 132–134). It has also been demonstrated that alcohol-induced dysbiosis is only partially reversible by alcohol withdrawal or probiotic (oral supplementation of *Lactobacillus plantarum* 8PA3 and *Bifidobacterium bifidum*) treatment (94, 113). Interestingly, patients dependent on alcohol also displayed reduced fungal diversity and *Candida* overgrowth, presenting the first evidence of the role of gut mycobiome in pathogenesis of liver diseases (8).

Genetically modified murine models have advanced our mechanistic understanding of the contribution of various components of the gut-barrier in the etiology and progression of ALD. Using *Reg3b*$^{-/-}$ or *Reg3g*$^{-/-}$ mice, it was found that REG3 lectins protected against alcoholic steatohepatitis by reducing mucosa-associated microbiota, thereby preventing translocation of viable bacteria (135). Muc2-deficient mice were protected against alcohol-induced liver inflammation (similar to HFD-induced inflammation in NAFLD model) due to a compensatory

increase in Reg3g and Reg3b lectins (107). Furthermore, IgA-knockout in mice led to increased levels of IgM and a net protective effect against ASH progression (136).

In response to ethanol-induced gut-barrier dysfunction and translocation, TLRs and other PRRs activate hepatic Kupffer cells and macrophages, as was demonstrated in male Wistar rats (137). This step initiates inflammatory cascades releasing TNF, IL-1, IL-10, IL-12 and TGF-β (138–140). Using TLR4 chimeric mice, it was shown that endotoxin-induced release of TGF-β is mediated by a MyD88-NF-κB-dependent pathway, providing an explanatory mechanism for endotoxin-induced liver inflammation (69). Furthermore, increased translocation of fungal β-glucan also induced liver inflammation via CLEC7A receptor on hepatic Kupffer cells such that treatment of mice with antifungal agents reduced intestinal fungal overgrowth, decreased β-glucan translocation and ameliorated ethanol-induced liver disease (8).

Alongside immunological responses to barrier dysfunction, ALD is also marked by system-wide changes in many bioactive compounds. Alcohol consumption leads to an increase in hepatic bile acid synthesis in humans and mice (141, 142).This increase could be explained by dysbiosis-associated disruption in FXR activation in enterocytes as FXR-deficient mice were more likely to develop ethanol-induced steatohepatitis (143), and treatment with an FXR agonist (WAY-362450) had protective effects against liver damage (144). Alcohol-associated dysbiosis in mice was further linked to reduced LCFA biosynthesis such that LCFA supplementation restored eubiosis. In fact, a statistically significant correlation between *Lactobacillus spp*. and bacterial LCFA (C15:0 and C17:0) was found in patients with ALD but not in healthy individuals as controls (84). Butyrate (a SCFA) production was also negatively altered following ethanol exposure and administration of butyrate in the form of tributyrin mitigated alcohol-induced liver injury in mice (85).

With increasing evidence of mechanistic links between the gut microbiota and liver disease progression, fecal microbiota transplantation (FMT) is being explored as a therapeutic option for ALD (68, 131). However, larger, carefully designed trials across multiple ethnic groups are needed before FMT can be considered safe in routine clinical practice for managing ALD.

Cirrhosis: Cirrhosis (or end-stage liver disease) is an extreme manifestation of chronic liver injury characterized by loss of liver cells, thick fibrous scar and regenerating nodules; this topic has been extensively reviewed elsewhere so we only provide a brief discussion here (145). NAFLD, ALD, primary biliary cholangitis (PBC; Box 1.2.2), primary sclerosing cholangitis (PSC; Box 1.2.3) or hepatitis can each progress to cirrhosis and constitute its subtypes. ASH and NASH have emerged as the second and third leading causes of cirrhosis in adults in the USA (after chronic hepatitis C infection) and based on the etiology there is a variable risk of developing HCC (146–148).

Alterations in the gut microbiota including dysbiosis and SIBO have been associated with and its complications (149–152). Treatment for portal systemic encephalopathy and decompensated cirrhosis includes treatment with nonsystemic antibiotics such as rifaximin to reduce intestinal microbiota overgrowth (153–155). Gut microbiome alterations were observed in patients with alcohol-associated and hepatitis-associated cirrhosis in a Chinese cohort (156), with an invasion of the lower intestinal tract by microbes associated with the oral cavity such as *Veillonella* and *Steptococcus*. Concordant with these findings, Chen and colleagues also found an over-representation of genera including *Veillonella*, *Megasphaera*, *Dialister*, *Atopobium* and *Prevotella* in the duodenum of patients with cirrhosis. The genera *Neisseria* and *Gemella* were discriminative between HBV-related and PBC-related cirrhosis (152). In 2017, Bajaj and colleagues observed statistically significant fungal dysbiosis in patients with cirrhosis and showed

that *Bacteroidetes* to *Ascomycota* ratio could independently predict hospitalization in these patients (157).

All experimental models of liver fibrosis result in gut microbial dysbiosis and increased intestinal permeability, and treatment of gastrointestinal tract with nonabsorbable antibiotics (such as rifaximin, neomycin) improved survival by immunomodulation, reducing translocation and incidences of infection (158). Mice with genetic ablations of the receptors for bacterial product ligands (TLR2, TLR4, TLR9, and NLP3) are protected from experimental liver fibrosis (158). The current treatment philosophy involves decreasing the bacterial product ligands or blocking their receptors, which results in decreased inflammatory and fibrogenic signaling in the liver, although no antifibrotic drug is currently available for routine clinical practice.

Hepatocellular carcinoma: The etiology of non-viral HCC follows a so-called multiple-hit pathway, whereby liver steatosis, followed by oxidative stress, ER stress together with intestinal dysbiosis and inflammation contribute to the final manifestation of cancer. The gut microbiota dramatically changes in composition in hosts with HCC. *Clostridium* species have been found to be enriched in obesity-induced mouse models of HCC (159, 160), but clinical studies of patients with HCC detected an overgrowth of intestinal *Escherichia coli* (161). Murine models and human studies have reported a migration of *Helicobacter* species to HCC tumor tissues (162–165). Notably, members of this genus are known to promote tumor-development by activating NF-kB and WNT signaling and suppressing anti-tumor immunity, and might have a potential role in HCC development (162, 166).

To get insights into the molecular events explaining the progression of liver disease to HCC, various murine models (using diet, toxin plus diet- and genetics plus diet) have been explored. However, most of these models have proven suboptimal because they either do not

develop all intermediate pathological & metabolic stages, or they manifest HCC incompletely (Febbraio and Karin, unpublished data). We have highlighted some frequently-used rodent models of liver disease, their usage and caveats in Table 1.2.2 to aide future research.

Accumulating evidence suggests that HCC-associated dysbiosis is accompanied by gut-barrier dysfunction, bacterial translocation, systemic circulation of their tumor-promoting metabolites and activation of proinflammatory and oncogenic signaling pathways (167). The intestinal poly-immunoglobulin receptor (PIgR) regulates the transport of IgA into the intestinal lumen and maintains microbial homeostasis (168). PIgR$^{-/-}$ mice modelling NASH-induced HCC had increased levels of systemic and liver IgA, and a concomitant increase in hepatic tumorigenesis due to localized inhibition of liver cytotoxic T cells that prevent HCC development (169). Furthermore, the application of broad spectrum antibiotics (such as ampicillin, amoxicillin) has been shown to attenuate liver inflammation and HCC-development in mice (159, 170), highlighting the role of the intestinal microbiome in liver tumorigenesis. In another mouse model in which HCC was induced by diethylnitrosamine (a carcinogen), activation of TLR4 due to LPS translocation upregulated the hepatic mitogen EREG in hepatic stellate cells and activated NF-kB, resulting in enhanced tumor cell proliferation (170). Additionally, the secondary bile acid deoxycholic acid (increased in dietary or genetic obesity), a metabolic byproduct of gut bacteria, was shown to upregulate proinflammatory genes, such as IL6 and TNF, to provoke a senescence-associated secretory phenotype (SASP) in hepatic stellate cells suggesting that SASP could be playing a key role in at least obesity-linked HCC development (159, 171–173).

In addition to its role in HCC development, the gut microbiome also modulates pro-tumorigenic adaptive immune response via type 17 T helper (Th17) cells, which produce the proinflammatory cytokine IL-17A (174–176). The therapeutic efficacy of the anticancer drug

cyclophosphamide depended on the interplay between Th17 signaling and gut microbiome such that germ-free tumor-bearing mice or mice given non-absorbable antibiotics had reduced Th17 response and a subsequent resistance to therapeutic effects of cyclophosphamide was seen (177).

Increased understanding of the role of the gut microbiota has motivated successful microbiome-based therapeutic modalities for HCC, such as treatment with synthetic bile acids to reduce HCC risk in patients with NAFLD (178), non-selective beta-blockers in the intestinal mucosa which prevent bacterial translocation and liver inflammation (179) and administration of probiotics in rodents models of HCC slowed tumor growth and reduced tumor size (180).

1.2.4 Experimental design of microbiome studies

Given the intense interest in the past decade in links between the microbiome and liver disease, we provide a brief overview of experimental models useful for researchers entering this field.

Association studies and case–control design: Much of our knowledge of the human microbiome comes from association studies that use either a cross-sectional or case–control design. Well-designed case–control studies are critical to demonstrate a potential relationship between microbes and a disease of interest. However, these studies cannot establish causality, and are often subject to confounding variables such as differences in diet or medication between cases and controls. Most studies are conducted at a single time point in a population with the disease, and no long-term follow-up is performed. Consequently, these studies can only identify microbes that differentiate individuals with the disease and the control population. Although these microbes identified might have been causative agents, it is nearly impossible to separate this association from secondary effects associated with the condition. For example, medication plays a major part

in shaping the microbiome; a study of patients with type II diabetes mellitus found that treatment with metformin had a larger effect on the microbiome than the disease (181). Similarly, we hypothesize the physiology of the disease might also contribute to changes in community structure.

Association studies are also often confounded by the selection of poor controls. The microbiome is dynamic (182, 183), and cumulative exposures over an individual's life, shaped by their diet (184), lifestyle (185), medical history (181, 186), genetics (187) and other factors (188) create a unique community. Thus, if cases and controls are not correctly selected, association studies might detect differences due to confounding factors. Matching cases and controls based on age and sex is often not sufficient. In cases in which this matching to control for confounding variables is not possible, it is critically important to collect information about potential confounding factors.

Comparisons across current cross-sectional studies are also challenging due to large effects caused by inter-study differences in technical parameters, including sample collection, storage, primer selection and analysis techniques (188). Differences across studies increase the challenge of meta-analysis and make identifying causative clades more difficult (189). Some of these problems can be ameliorated by using consistent methodology between (188, 189). Efforts like the Microbiome Quality Control Project are exploring sources of technical variation (190), while analysis platforms like Qiita (www.qitta.ucsd.edu) provide a database of consistently annotated studies for comparison.

Twin studies: Twin studies provide a potential antidote to some of the problems with association studies. Twin pairs are naturally controlled for age and some early life exposures.(191) Monozygotic twin pairs also share the same genetic background, further limiting potential confounders (191). Twin studies can be leveraged in two ways. First, identifying differences

between discordant and concordant twin pairs represent more powerful association studies, due to the partial internal control. Although these studies are particularly useful in young children due to shared environment, the approach can also be used with adults (192). Second, twin studies are critical to examine genetic control of the microbiome. A study published in 2016 of the UK Twins cohort suggested strong association of the microbiome and genes, including those associated with dietary preference and serum lipids (193). Twin studies provide a unique opportunity to assess if the familial risk factors are either genetic or environmental in nature. These studies have been applied to study heritability for studying hepatic steatosis and fibrosis now that advanced magnetic resonance imaging based assessment can be used to phenotype individuals (119, 194). However, the sample size requirements for microbiome assessment in twins is large compared to the sample sizes heeded to study heritability may making recruitment for such a study challenging (120, 193). Twin studies may not be appropriate for other rare causes of liver diseases e.g. alpha-1 antitrypsin deficiency, cirrhosis, primary biliary cholangitis, and for such low prevalence diseases a trio family design would be more appropriate and would provide the highest power with the most efficient study design to detect association of a trait such as the role of microbiome on the risk of liver diseases (121).

Longitudinal studies: As the cost of microbiome analysis decreases, longitudinal studies are becoming more common. Understanding temporal fluctuation in the microbiome, and the role of microbes in contributing to disease etiology, will rely on studies over time. Work suggests that community instability might, in and of itself, be a characteristic of an unhealthy ecosystem (195, 196). Prospective studies, such as an investigation examining death from HCC in individuals with NAFLD, have helped identify the role of exposures and etiological factors in contributing to disease outcomes (197). Currently, the appropriate sampling frequency for understanding the

microbiome in prospective studies is unknown, in part due to an overall lack of long term follow up with microbiome studies. Initially, sample collection during standard clinic visits may provide information about the population-scale changes in the population. However, incorporating microbiome samples into these long-term studies will help examine the role of microbial communities—either at a single time point or the community dynamics—as a contributing factor to complex conditions (198).

Animal models: Model animals also have an important role in shaping our understanding of the microbiome in disease (Table 1.2.2). Although rodent microbial communities are distinct from the human microbiota, there are some shared physiological and microbial traits (199). Both rodent and human communities are dominated by the same set of bacterial phyla, although a smaller percentage of genera are shared (199). As such, experimental findings implicating individual organisms or genera in rodents should be taken with caution until they are validated in humans. Instead, rodent models can show phenotypic consequences of microbiome manipulation. This aspect makes rodent models a useful model system to investigate causality, explore interactions and test early interventions.

Both antibiotics and probiotics have been used to study the effect of changing the conventional mouse microbiome on a phenotypic outcome. Broad spectrum antibiotics decrease the total bacterial load, as well as causing major perturbations in the microbial communities (200). In some cases, such as in liver disease models, this approach can demonstrate the role of bacterial products like LPS in modulating inflammation (127). In other cases, like a reported addiction model, it can be used to demonstrate the importance of an intact microbiome in regulating behavior (201). Probiotics can also be used to investigate the effect of a specific bacteria or bacterial cocktail

within a controlled environment. A study of alcoholic fatty liver disease demonstrated an attenuation of the microbiome-mediated inflammation when a probiotic was used (106).

Gnotobiotic or germ-free mice can be used in multiple contexts. Comparisons of specific pathogen-free laboratory mice and germ-free mice can be used to examine the role of the microbiome in modulating an expressed or induced phenotype (202, 203). More importantly, gnotobiotic mice can be humanized with a donor's stool. This approach creates a system in which an individual's microbiota can be tested, either for its ability to modulate a disease phenotype or as a target for intervention (192, 202). For instance, in a small study, mice received their microbiome from a donor with either severe alcoholic hepatitis or no liver disease. Following alcohol treatment, the mice with the microbiome from the patient with alcoholic hepatitis showed greater liver damage than mice that received stool from the healthy donor (204).

Well-designed mouse models that combine our current understanding of liver disease with humanized microbiomes offer some of the greatest potential for preclinical interventions. Avatar, or sometimes called patient-derived xenograft (PDX) mice, are widely used in the cancer community to test the efficacy of chemotherapeutics for individual tumors, including HCC (205, 206). This model better re-capitulates the complexity of a tumor than cell culture. Avatar mice can be further personalized by introducing a human immune system into an immunocompromised mouse, along with the tumor (205). Generating this model in germ-free mice with a humanized microbiome and immune system expands our capacity to understand the role of the microbiome in modulating cancer. For example, this model could be used to study whether the microbiome of a patient with ALD leads to more tumor growth than the microbiome from a healthy individual as control.

1.2.5 Conclusions

An accumulating body of research suggests that the disparate observations in liver-disease-related studies can be unified and explained by the microbiome. It is now widely accepted that liver damage can result from extensive interplay between gut microbiota via specialized molecules (such as TMA, acetaldehyde and LPS) and host-immune system via Kupffer-cell-mediated liver inflammation. However, a comprehensive understanding of the interactions between the microbiome and the liver still evades us. Animal models, particularly rodents, have been instrumental in elucidating many important mechanistic pathways in liver disease etiology. The introduction of the microbiome into these models will provide a more complete view of the cancer ecosystem. Because microbiome research is sensitive to technical variability that often masks underlying biological signals, there is a need for consistency in technical platforms and standardized protocols, so that findings from different laboratories (and model organisms) can be replicated and validated. Additionally, it is critical to use an animal model that mimics human disease as closely as possible in all its physiological and metabolic manifestations.

We are slowly advancing from observation-based studies in humans as research establishes grounds for microbiome-based therapeutic modalities such as FMT and probiotic interventions. However, effectively translating and applying findings accrued through animal models to humans requires well-designed, large-scale clinical trials spanning multiple disease etiologies and patient characteristics. As the role of microbiota in liver disease development, prognosis and treatment is increasingly recognized, we emphasize the need for focused, microbiome-aware efforts to efficiently tackle the socioeconomic burden of this spectrum of liver diseases.

1.2.6 Acknowledgements

The authors would like to thank D. McDonald, T. Kościółek, Z. Xu and A. Plymoth for their helpful discussions. M.K. is supported by the NIH grants R01 AI043477 and R01 CA118165. R.L. is supported in part by the grant R01-DK106419-03. Research reported in this publication was supported in part by the National Institute of Environmental Health Sciences of the National Institutes of Health under Award Number P42ES010337. B.S. is supported by NIH grants R01 AA020703, U01 AA021856, U01AA24726, and by Award Number I01BX002213 from the Biomedical Laboratory Research & Development Service of the VA Office of Research and Development. J.D. is supported by the Robert Wood Johnson Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

1.2.7 Author contributions

Chapter 1, Introduction part 2, in full, is a reprint of previously published material: Tripathi, A., Debelius, J., Brenner, D. A., Karin, M., Loomba, R., Schnabl, B., & Knight, R. (2018). The gut-liver axis and the intersection with the microbiome. *Nature Reviews. Gastroenterology & Hepatology*, *15*(7), 397–411.

A.T., J.D., D.A.B., M.K., R.L., B.S. and R.K. researched data for the article. A.T., J.D., D.A.B., M.K., R.L., B.S. and R.K. made substantial contributions to discussion of content. A.T., D.A.B., M.K., R.L., B.S. and R.K, reviewed/edited the manuscript before submission. A.T., J.D. and R.K. wrote the article. The co-authors listed above supervised or provided support for the research and have given permission for the inclusion of the work in this dissertation.

1.2.8 Competing interests

The authors declare no competing interests.

1.2.9 Supplemental information

Box 1.2.1 Open research questions

Mounting evidence implicates the gut microbiome in the development and progression of different forms of liver disease. However, several questions remain open and must be answered to advance the field.

- Is there a set of microbes (beneficial or harmful) that can read out the current extent, or predict the future extent, of disease progression in patients with alcoholic liver disease and NAFLD?
- Can microbiome research using a consistent set of methodologies, including multi-omics profiling, provide a consistent mechanistic picture that unifies our understanding of the relationships among forms of liver disease?
- Can fecal microbiota transplantation, or collections of probiotic strains isolated from human feces, be expanded as a therapeutic modality for liver disease? Does introducing a humanized microbiome into a hepatocellular carcinoma avatar mouse improve its fidelity in terms of responding to therapeutic options like an individual patient?

Box 1.2.2 Primary biliary cholangitis

Primary biliary cholangitis (PBC) is characterized by inflammation-mediated damage to the small bile ducts inside the liver, which gradually progresses to liver fibrosis and cirrhosis (207). Previously considered a typical autoimmune disorder, the modified etiological understanding of

PBC considers proinflammatory changes in the gut-microbiota, intestinal bile acid disruptions and gut-barrier dysfunction (207–210). Consequently, microbe-associated molecular patterns ascend the biliary duct, perpetuating infection. An immune attack against the biliary epithelial cells is mediated by antibodies that recognize E2 subunit of pyruvate dehydrogenase complex (PDC-E2) due to cross-reactivity with conserved proteins in *Escherichia coli*,(207) *Lactobacillus delbrueckii* (211) and *Novosphingobium aromaticivorans* (212) In fact, genetically susceptible mouse strains developed liver lesions mimicking PBC when infected with *Novosphingobium aromaticivoransm,* which further implicates a role for the microbiome in this disease (213). Ursodeoxycholic acid, a tertiary bile acid produced by *Ruminococcus,* has been approved for PBC treatment (214). Thus, microbiome-based treatment modalities hold promise for managing PBC and should be studied further.

Box 1.2.3 Primary sclerosing cholangitis

Primary sclerosing cholangitis (PSC) is also an immune-mediated disease of the bile ducts (207). However, unlike PBC, PSC can affect bile ducts, both inside and outside of the liver. Gut dysbiosis-mediated bile dysregulation, intestinal permeability and translocation of proinflammatory molecules in the portal vein characterizes PSC (207, 215, 216). The immune reaction in PSC is mediated by autoantibodies, including perinuclear antineutrophil cytoplasmic antibody, that recognize the ubiquitously expressed bacterial antigen FtsZ (217). Furthermore, increase in microbe-associated Toll-like receptor expression and T helper type 17 (Th17) cells has been reported in PSC, which strongly suggests microbiome involvement in disease pathogenesis (218, 219). PSC is closely associated with IBD (220), in particular ulcerative colitis and shares some of its characteristic features (such as increased levels of Th17 cells in the gut). Thus, a

common disease mechanism might be at play, and novel treatment avenues by targeting microbe-associated immune pathways can be explored.

Table1.2.1 Comparison of alcoholic and nonalcoholic liver disease

Factor	Alcoholic liver disease	NAFLD
Small intestinal bacterial overgrowth	Observed (106, 221, 222)	Observed (66)
Gut microbiota	↑ Enterobactericeae (humans) (132, 133) ↓ *Lactobacillus* (133, 222) (humans and mice), Bacteroidetes (humans) (132, 134), *Akkermansia muciniphila* (humans and mice) (58) Gut microbiota protects against alcohol-induced liver injury (87) Reduced fungal diversity; *Candida* overgrowth (8)	↑ Enterobactericeae (humans) (65, 223), *Lactobacillus* (humans) (223, 224), *Bacteroides* (humans and mice) (125, 160), *Ruminococcus* (humans) (125) ↓ *Prevotella* (humans) (125, 223) , *Akkermansia muciniphila* (mice) (59) Gut microbiota mediates high-fat-diet-induced liver steatosis (225, 226) Fungal dysbiosis not demonstrated
Reversibility of gut dysbiosis	Partial reversibility on abstinence (94, 113)	Reversibility not demonstrated
Inflammation	↑ Intestinal TNF (mice) (103) ↑ Systemic inflammatory markers (humans) (48, 227)	↑ Intestinal TNF, IFNγ, IL-6 (humans and mice) (223, 228) ↑ Systemic inflammatory markers (humans) (229)
Transferability via microbiome	FMT from patients alcoholic hepatitis caused severe liver inflammation and injury in mice (204) FMT from ALD-resistant to ALD-susceptible mice prevented liver injury in recipient (131)	Co-housing inflammasome deficient, NASH mice with wild-type mice exacerbated liver steatosis wild-type cage mates(49) FMT from NAFLD-susceptible mice promoted liver injury in recipient(230)
Translocation	↑ PAMPs translocation (endotoxins (48, 231–233), β-glucan (8), viral or bacterial DNA (231, 234) (humans and mice)	↑ PAMPs translocation (endotoxins (232, 235), viral or bacterial DNA (236) (humans and mice)
Bile acids	↑ Total plasma bile acids (humans) (237) ↑ Hepatic bile acid synthesis (humans and mice) (141, 142)	↑ Total serum bile acids (humans) (238) ↑ Hepatic bile acid synthesis (humans), total fecal bile acids, primary to secondary bile acid ratio (35)

Table1.2.1 Comparison of alcoholic and nonalcoholic liver disease (continued)

Factor	Alcoholic liver disease	NAFLD
Choline	↓ Phosphatidylcholine in plasma and liver (rats) (239, 240) (Changes in trimethylamine not demonstrated)	↓ Phosphatidylcholine in plasma (mice) (241) ↑ Intestinal trimethylamine (mice) (241)
Free-fatty acids	↓ Bacterial fatty-acid biosynthesis (mice) (84) LCFA and SCFA supplementation reduced ethanol-induced liver injury (mice) (84, 115)	↑ Free-fatty acids in the liver (242)
Ethanol	↑ Blood ethanol, luminal acetaldehyde (130) ↑ Systemic acetate (84, 241)	↑ Blood ethanol (65, 243, 244)

ALD, alcoholic liver disease; FMT, faecal microbiota transplantation; LCFA, long-chain fatty acid; PAMPs, pathogen-associated molecular patterns; SCFA, short-chain fatty acid.

Table 1.2.2 Experimental mouse models for liver disease

Model	Description	Liver pathology	Microbiome Features
Diet			
High-fat diet	Diet using higher saturated fat, or supplemented with cholesterol, compared with chow	Induces fatty liver and hepatic steatosis Associated with metabolic syndrome phenotype (245)	Common model for inducing dysbiosis; associated with changes in the microbiome
Choline-deficient diet	A high-fat diet with choline and methionine omitted	Induces fatty liver, steatosis and inflammation and fibrosis The model does not contribute to metabolic syndrome (5)	Small study suggests diet-induced changes (246)
Ethanol-supplemented liquid diet	A model of chronic alcohol abuse administered as an isocaloric diet in which ethanol or maltose and dextrose are supplemented Diet can be administered orally (Lieber-DeCarli (247)) or intragastrically (Tsukamoto-French (248))	Oral supplementation leads to inflammation and fatty liver, representing a good model for early ALD Intragastric administration leads to severe steatosis and mild fibrosis (4)	Diet affects the abundance of several taxa and is associated with changes in the microbiome (58)

Table 1.2.2 Experimental mouse models for liver disease (continued)

Model	Description	Liver pathology	Microbiome Features
Genetic manipulations			
Knockout model	A mouse line in which both copies of a gene have been removed	Pathology depends on the targeted gene For example, *Fxr*$^{-/-}$ mice have more fatty liver accumulation on a high-fat diet (32), *Muc2*$^{-/-}$ are protected from diet-induced liver injury (107), *Gsta4*$^{-/-}$, *Ppara*$^{-/-}$ double knockout mice have increased inflammation and fibrosis compared with either single mutant or wild-type (249)	The microbiome of lineage-derived mice is distinct from wild-type mice, which is likely to be an effect of microbiome drift within the colonies, rather than a direct effect of the genotype (250)
Littermate controls	Mice from a heterozygous cross that lead to wild type and knockout littermates		Much of the mouse microbiome is acquired through vertical transmission; littermates are better microbial controls (251)
Cre-Lox localized mutation	A genetic cross that enables tissue-specific knockout of a gene	Pathology is dependent on the gene A Cre/Lox model of liver-specific E-cadherin knockout shows pathology like primary sclerosing cholangitis, and increases susceptibility to cancer (252) The loss of TLR5 in hepatocytes leads to increased inflammation and fibrosis in a high-fat-diet-induced model of NASH (253)	Microbiome considerations depend on the how the controls are selected
Avatar Mice	Mice transplanted with solid state tumors from patients with cancer	Human hepatocellular carcinoma can be transplanted into the mouse (254)	There is no specific effect on the microbiome

Table 1.2.2 Experimental mouse models for liver disease (continued)

Model	Description	Liver pathology	Microbiome Features
Microbiome			
Antibiotic treatment	Treatment with a broad-spectrum antibiotic	No direct effect on liver disease; antibiotics can moderate the effect of other interventions	Antibiotics can have off target effects and substantially alter the microbial community in addition to decreasing the bacterial load (200)
Probiotic manipulation	Microbial supplementation to modify the microbiota	No direct effect on liver disease; probiotics can modulate the effect of other treatments: *Lactobacillus* to ameliorate alcohol-induced liver injury) (106)	Can lead to the over-abundance of a specific organism or correct defects in the community; however, not all probiotics colonize
Germ-Free Mice	Raised without any bacterial community	Germ-free mice have immune defects (255) These mice are also more susceptible to alcohol-induced liver injury (87)	Useful to demonstrate the importance of bacterial communities for a phenotype
Monoculture gnotobiotic mice	Germ free mice that have been colonized with a single bacterium or defined bacterial community	No direct effect; depends on the community transplanted and challenge	Can test whether the defined community can modulate the phenotype
Mouse transplant	Bacterial communities from mice transplanted into germ-free mice	No direct effect; depends on the community transplanted and challenge	Demonstrates whether mouse phenotype is transferable or can be modulated through the microbial community
Humanized mice	Germ-free mice that have been gavaged with the microbiota from a human donor	No direct effect; depends on the community transplanted and challenge	Demonstrates whether human phenotype is transferable or can be through the microbial community

1.2.10 References

1. Schnabl B, Brenner DA. 2014. Interactions between the intestinal microbiome and liver diseases. Gastroenterology 146:1513–1524.

2. Hartmann P, Seebauer CT, Schnabl B. 2015. Alcoholic liver disease: the gut microbiome and liver cross talk. Alcohol Clin Exp Res 39:763–75.

3. Younossi ZM, Blissett D, Blissett R, Henry L, Stepanova M, Younossi Y, Racila A, Hunt S, Beckerman R. 2016. The economic and clinical burden of nonalcoholic fatty liver disease in the United States and Europe. Hepatology 64:1577–1586.

4. Bertola A, Mathews S, Ki SH, Wang H, Gao B. 2013. Mouse model of chronic and binge ethanol feeding (the NIAAA model). Nat Protoc 8:627–37.

5. Pelz S, Stock P, Brückner S, Christ B. 2012. A methionine-choline-deficient diet elicits NASH in the immunodeficient mouse featuring a model for hepatic cell transplantation. Exp Cell Res 318:276–87.

6. Itagaki H, Shimizu K, Morikawa S, Ogawa K, Ezaki T. 2013. Morphological and functional characterization of non-alcoholic fatty liver disease induced by a methionine-choline-deficient diet in C57BL/6 mice. Int J Clin Exp Pathol 6:2683–96.

7. Loomba R, Seguritan V, Li W, Long T, Klitgord N, Bhatt A, Dulai PS, Caussy C, Bettencourt R, Highlander SK, Jones MB, Sirlin CB, Schnabl B, Brinkac L, Schork N, Chen C-H, Brenner DA, Biggs W, Yooseph S, Venter JC, Nelson KE. 2017. Gut Microbiome-Based Metagenomic Signature for Non-invasive Detection of Advanced Fibrosis in Human Nonalcoholic Fatty Liver Disease. Cell Metab 25:1054-1062.e5.

8. Yang A-M, Inamine T, Hochrath K, Chen P, Wang L, Llorente C, Bluemel S, Hartmann P, Xu J, Koyama Y, Kisseleva T, Torralba MG, Moncera K, Beeri K, Chen C-S, Freese K, Hellerbrand C, Lee SM, Hoffman HM, Mehal WZ, Garcia-Tsao G, Mutlu EA, Keshavarzian A, Brown GD, Ho SB, Bataller R, Stärkel P, Fouts DE, Schnabl B. 2017. Intestinal fungi contribute to development of alcoholic liver disease. J Clin Invest 127:2829–2841.

9. Chen Y-M, Liu Y, Zhou R-F, Chen X-L, Wang C, Tan X-Y, Wang L-J, Zheng R-D, Zhang H-W, Ling W-H, Zhu H-L. 2016. Associations of gut-flora-dependent metabolite trimethylamine-N-oxide, betaine and choline with non-alcoholic fatty liver disease in adults. Sci Rep 6.

10. Csak T, Ganz M, Pespisa J, Kodys K, Dolganiuc A, Szabo G. 2011. Fatty acid and endotoxin activate inflammasomes in mouse hepatocytes that release danger signals to stimulate immune cells. Hepatology 54:133–44.

11. Uesugi T, Froh M, Arteel GE, Bradford BU, Thurman RG. 2001. Toll-like receptor 4 is involved in the mechanism of early alcohol-induced liver injury in mice. Hepatology 34:101–8.

12. Anand G, Zarrinpar A, Loomba R. 2016. Targeting Dysbiosis for the Treatment of Liver Disease. Semin Liver Dis 36:37–47.

13. Seki E, Schnabl B. 2012. Role of innate immunity and the microbiota in liver fibrosis: crosstalk between the liver and gut. J Physiol 590:447–58.

14. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. 2011. Global cancer statistics. CA Cancer J Clin 61:69–90.

15. Schwabe RF, Jobin C. 2013. The microbiome and cancer. Nat Rev Cancer 13:800–812.

16. Garrett WS. 2015. Cancer and the microbiota. Science (80-) 348:80–86.

17. Koppel N, Maini Rekdal V, Balskus EP. 2017. Chemical transformation of xenobiotics by the human gut microbiota. Science 356:eaag2770.

18. Tolba R, Kraus T, Liedtke C, Schwarz M, Weiskirchen R. 2015. Diethylnitrosamine (DEN)-induced carcinogenic liver injury in mice. Lab Anim 49:59–69.

19. Stärkel P, Schnabl B. 2016. Bidirectional Communication between Liver and Gut during Alcoholic Liver Disease. Semin Liver Dis 36:331–339.

20. Chiang JYL. 2013. Bile acid metabolism and signaling. Compr Physiol 3:1191–212.

21. Wahlström A, Sayin SI, Marschall HU, Bäckhed F. 2016. Intestinal Crosstalk between Bile Acids and Microbiota and Its Impact on Host Metabolism. Cell Metab.

22. Arab JP, Karpen SJ, Dawson PA, Arrese M, Trauner M. 2017. Bile acids and nonalcoholic fatty liver disease: Molecular insights and therapeutic perspectives. Hepatology 65:350–362.

23. Zarrinpar A, Loomba R. 2012. Review article: the emerging interplay among the gastrointestinal tract, bile acids and incretins in the pathogenesis of diabetes and non-alcoholic fatty liver disease. Aliment Pharmacol Ther 36:909–21.

24. Copple BL, Li T. 2016. Pharmacology of bile acid receptors: Evolution of bile acids from simple detergents to complex signaling molecules. Pharmacol Res 104:9–21.

25. Sinal CJ, Tohkin M, Miyata M, Ward JM, Lambert G, Gonzalez FJ. 2000. Targeted disruption of the nuclear receptor FXR/BAR impairs bile acid and lipid homeostasis. Cell 102:731–44.

26. Pols TWH, Noriega LG, Nomura M, Auwerx J, Schoonjans K. 2011. The bile acid membrane receptor TGR5 as an emerging target in metabolism and inflammation. J Hepatol 54:1263–72.

27. Broeders EPM, Nascimento EBM, Havekes B, Brans B, Roumans KHM, Tailleux A, Schaart G, Kouach M, Charton J, Deprez B, Bouvy ND, Mottaghy F, Staels B, van Marken Lichtenbelt WD, Schrauwen P. 2015. The Bile Acid Chenodeoxycholic Acid Increases Human Brown Adipose Tissue Activity. Cell Metab 22:418–26.

28. Thomas C, Gioiello A, Noriega L, Strehle A, Oury J, Rizzo G, Macchiarulo A, Yamamoto H, Mataki C, Pruzanski M, Pellicciari R, Auwerx J, Schoonjans K. 2009. TGR5-mediated bile acid sensing controls glucose homeostasis. Cell Metab 10:167–77.

29. Perino A, Schoonjans K. 2015. TGR5 and Immunometabolism: Insights from Physiology and Pharmacology. Trends Pharmacol Sci 36:847–57.

30. Schaap FG, Trauner M, Jansen PLM. 2014. Bile acid receptors as targets for drug development. Nat Rev Gastroenterol Hepatol 11:55–67.

31. Inagaki T, Moschetta A, Lee Y-K, Peng L, Zhao G, Downes M, Yu RT, Shelton JM, Richardson JA, Repa JJ, Mangelsdorf DJ, Kliewer SA. 2006. Regulation of antibacterial defense in the small intestine by the nuclear bile acid receptor. Proc Natl Acad Sci U S A 103:3920–5.

32. Parséus A, Sommer N, Sommer F, Caesar R, Molinaro A, Ståhlman M, Greiner TU, Perkins R, Bäckhed F. 2017. Microbiota-induced obesity requires farnesoid X receptor. Gut 66:429–437.

33. Jiang C, Xie C, Li F, Zhang L, Nichols RG, Krausz KW, Cai J, Qi Y, Fang Z-Z, Takahashi S, Tanaka N, Desai D, Amin SG, Albert I, Patterson AD, Gonzalez FJ. 2015. Intestinal farnesoid X receptor signaling promotes nonalcoholic fatty liver disease. J Clin Invest 125:386–402.

34. Ridlon JM, Kang DJ, Hylemon PB, Bajaj JS. 2014. Bile acids and the gut microbiome. Curr Opin Gastroenterol 30:332–8.

35. Mouzaki M, Wang AY, Bandsma R, Comelli EM, Arendt BM, Zhang L, Fung S, Fischer SE, McGilvray IG, Allard JP. 2016. Bile Acids and Dysbiosis in Non-Alcoholic Fatty Liver Disease. PLoS One 11:e0151829.

36. Odenwald MA, Turner JR. 2017. The intestinal epithelial barrier: a therapeutic target? Nat Rev Gastroenterol Hepatol 14:9–21.

37. Turner JR. 2009. Intestinal mucosal barrier function in health and disease. Nat Rev Immunol 9:799–809.

38. Abreu MT. 2010. Toll-like receptor signalling in the intestinal epithelium: how bacterial recognition shapes intestinal function. Nat Rev Immunol 10:131–144.

39. Gallo RL, Hooper L V. 2012. Epithelial antimicrobial defence of the skin and intestine. Nat Rev Immunol 12:503–516.

40. Mantis NJ, Rol N, Corthésy B. 2011. Secretory IgA's complex roles in immunity and mucosal homeostasis in the gut. Mucosal Immunol 4:603–611.

41. Rakoff-Nahoum S, Paglino J, Eslami-Varzaneh F, Edberg S, Medzhitov R. 2004. Recognition of commensal microflora by toll-like receptors is required for intestinal homeostasis. Cell 118:229–41.

42. Yaku K, Enami Y, Kurajyo C, Matsui-Yuasa I, Konishi Y, Kojima-Yuasa A. 2012. The enhancement of phase 2 enzyme activities by sodium butyrate in normal intestinal epithelial cells is associated with Nrf2 and p53. Mol Cell Biochem 370:7–14.

43. Wächtershäuser A, Stein J. 2000. Rationale for the luminal provision of butyrate in intestinal diseases. Eur J Nutr 39:164–71.

44. Ziegler K, Kerimi A, Poquet L, Williamson G. 2016. Butyric acid increases transepithelial transport of ferulic acid through upregulation of the monocarboxylate transporters SLC16A1 (MCT1) and SLC16A3 (MCT4). Arch Biochem Biophys 599:3–12.

45. Lobos O, Barrera A, Padilla C. 2017. Microorganisms of the Intestinal Microbiota of Oncorhynchus Mykiss Produce Antagonistic Substances Against Bacteria Contaminating Food and Causing Disease in Humans. Ital J food Saf 6:6240.

46. Walsh CJ, Guinane CM, O' Toole PW, Cotter PD. 2017. A Profile Hidden Markov Model to investigate the distribution and frequency of LanB-encoding lantibiotic modification genes in the human oral and gut microbiome. PeerJ 5:e3254.

47. Graham CE, Cruz MR, Garsin DA, Lorenz MC. 2017. Enterococcus faecalis bacteriocin EntV inhibits hyphal morphogenesis, biofilm formation, and virulence of Candida albicans. Proc Natl Acad Sci U S A 114:4507–4512.

48. Leclercq S, Cani PD, Neyrinck AM, Stärkel P, Jamar F, Mikolajczak M, Delzenne NM, De Timary P. 2012. Role of intestinal permeability and inflammation in the biological and behavioral control of alcohol-dependent subjects. Brain Behav Immun 26:911–918.

49. Henao-Mejia J, Elinav E, Jin C, Hao L, Mehal WZ, Strowig T, Thaiss CA, Kau AL, Eisenbarth SC, Jurczak MJ, Camporez J-P, Shulman GI, Gordon JI, Hoffman HM, Flavell RA. 2012. Inflammasome-mediated dysbiosis regulates progression of NAFLD and obesity. Nature 482:179–85.

50. Martinez-Medina M, Denizot J, Dreux N, Robin F, Billard E, Bonnet R, Darfeuille-Michaud A, Barnich N. 2014. Western diet induces dysbiosis with increased E coli in CEABAC10 mice, alters host barrier function favouring AIEC colonisation. Gut 63:116–24.

51. Serino M, Luche E, Gres S, Baylac A, Bergé M, Cenac C, Waget A, Klopp P, Iacovoni J, Klopp C, Mariette J, Bouchez O, Lluch J, Ouarné F, Monsan P, Valet P, Roques C, Amar J, Bouloumié A, Théodorou V, Burcelin R. 2012. Metabolic adaptation to a high-fat diet is associated with a change in the gut microbiota. Gut 61:543–53.

52. Pendyala S, Walker JM, Holt PR. 2012. A high-fat diet is associated with endotoxemia that originates from the gut. Gastroenterology 142:1100-1101.e2.

53. Wang Y, Tong J, Chang B, Wang BB, Zhang D, Wang BB. 2014. Effects of alcohol on intestinal epithelial barrier permeability and expression of tight junction-associated proteins. 9:2352–6.

54. Fukui H, Brauner B, Bode JC, Bode C. 1991. Plasma endotoxin concentrations in patients with alcoholic and non-alcoholic liver disease: reevaluation with an improved chromogenic assay. J Hepatol 12:162–9.

55. Schäfer C, Parlesak A, Schütt C, Bode JC, Bode C. 2002. Concentrations of lipopolysaccharide-binding protein, bactericidal/permeability-increasing protein, soluble CD14 and plasma lipids in relation to endotoxaemia in patients with alcoholic liver disease. Alcohol Alcohol 37:81–6.

56. Tulstrup MV-L, Christensen EG, Carvalho V, Linninge C, Ahrné S, Højberg O, Licht TR, Bahl MI. 2015. Antibiotic Treatment Affects Intestinal Permeability and Gut Microbial Composition in Wistar Rats Dependent on Antibiotic Class. PLoS One 10:e0144854.

57. Forbes JD, Van Domselaar G, Bernstein CN. 2016. The Gut Microbiota in Immune-Mediated Inflammatory Diseases. Front Microbiol 7:1081.

58. Grander C, Adolph TE, Wieser V, Lowe P, Wrzosek L, Gyongyosi B, Ward D V, Grabherr F, Gerner RR, Pfister A, Enrich B, Ciocan D, Macheiner S, Mayr L, Drach M, Moser P, Moschen AR, Perlemuter G, Szabo G, Cassard AM, Tilg H. 2017. Recovery of ethanol-induced Akkermansia muciniphila depletion ameliorates alcoholic liver disease. Gut gutjnl-2016-313432.

59. Everard A, Belzer C, Geurts L, Ouwerkerk JP, Druart C, Bindels LB, Guiot Y. 2013. Cross-talk between Akkermansia muciniphila and intestinal epithelium controls diet-induced obesity. Proc Nati Acad Sci USA 110:9066–9071.

60. Elamin EE, Masclee AA, Dekker J, Jonkers DM. 2013. Ethanol metabolism and its effects on the intestinal epithelial barrier. Nutr Rev 71.

61. Filliol A, Piquet-Pellorce C, Raguénès-Nicol C, Dion S, Farooq M, Lucas-Clerc C, Vandenabeele P, Bertrand MJM, Le Seyec J, Samson M. 2017. RIPK1 protects hepatocytes from Kupffer cells-mediated TNF-induced apoptosis in mouse models of PAMP-induced hepatitis. J Hepatol 66:1205–1213.

62. Ni YH, Huo LJ, Li TT. 2017. [Effect of interleukin-22 on proliferation and activation of hepatic stellate cells induced by acetaldehyde and related mechanism]. Zhonghua Gan Zang Bing Za Zhi 25:9–14.

63. Wu X, Wang Y, Wang S, Xu R, Lv X. 2017. Purinergic P2X7 receptor mediates acetaldehyde-induced hepatic stellate cells activation via PKC-dependent GSK3β pathway 43:164–171.

64. López-Lázaro M. 2016. A local mechanism by which alcohol consumption causes cancer. Oral Oncol 62:149–152.

65. Zhu L, Baker SS, Gill C, Liu W, Alkhouri R, Baker RD, Gill SR. 2013. Characterization of gut microbiomes in nonalcoholic steatohepatitis (NASH) patients: A connection between endogenous alcohol and NASH. Hepatology 57:601–609.

66. Miele L, Valenza V, La Torre G, Montalto M, Cammarota G, Ricci R, Mascianà R, Forgione A, Gabrieli ML, Perotti G, Vecchio FM, Rapaccini G, Gasbarrini G, Day CP, Grieco A. 2009. Increased intestinal permeability and tight junction alterations in nonalcoholic fatty liver disease. Hepatology 49:1877–1887.

67. Pascual S, Such J, Esteban A, Zapater P, Casellas JA, Aparicio JR, Girona E, Gutiérrez A, Carnices F, Palazón JM, Sola-Vera J, Pérez-Mateo M. 2003. Intestinal permeability is increased in patients with advanced cirrhosis. Hepatogastroenterology 50:1482–6.

68. Philips CA, Pande A, Shasthry SM, Jamwal KD, Khillan V, Chandel SS, Kumar G, Sharma MK, Maiwall R, Jindal A, Choudhary A, Hussain MS, Sharma S, Sarin SK. 2017. Healthy Donor Fecal Microbiota Transplantation in Steroid-Ineligible Severe Alcoholic Hepatitis: A Pilot Study. Clin Gastroenterol Hepatol 15:600–602.

69. Seki E, De Minicis S, Österreicher CH, Kluwe J, Osawa Y, Brenner DA, Schwabe RF. 2007. TLR4 enhances TGF-β signaling and hepatic fibrosis. Nat Med 13:1324–1332.

70. sayama F, Hines IN, Kremer M, Milton RJ, Byrd CL, Perry AW, McKim SE, Parsons C, Rippe RA, Wheeler MD. 2006. LPS signaling enhances hepatic fibrogenesis caused by experimental cholestasis in mice. Am J Physiol Gastrointest Liver Physiol 290:G1318-28.

71. Gäbele E, Mühlbauer M, Dorn C, Weiss TS, Froh M, Schnabl B, Wiest R, Schölmerich J, Obermeier F, Hellerbrand C. 2008. Role of TLR9 in hepatic stellate cells and experimental liver fibrosis. Biochem Biophys Res Commun 376:271–276.

72. Hartmann P, Haimerl M, Mazagova M, Brenner DA, Schnabl B. 2012. Toll-Like Receptor 2–Mediated Intestinal Injury and Enteric Tumor Necrosis Factor Receptor I Contribute to Liver Fibrosis in Mice. Gastroenterology 143:1330-1340.e1.

73. Lebeaupin C, Proics E, de Bieville CHD, Rousseau D, Bonnafous S, Patouraux S, Adam G, Lavallard VJ, Rovere C, Le Thuc O, Saint-Paul MC, Anty R, Schneck AS, Iannelli A, Gugenheim J, Tran A, Gual P, Bailly-Maitre B. 2015. ER stress induces NLRP3 inflammasome activation and hepatocyte death. Cell Death Dis 6:e1879.

74. Zeisel SH, da Costa K-A. 2009. Choline: an essential nutrient for public health. Nutr Rev 67.

75. Han J, Dzierlenga AL, Lu Z, Billheimer DD, Torabzadeh E, Lake AD, Li H, Novak P, Shipkova P, Aranibar N, Robertson D, Reily MD, Lehman-McKeeman LD, Cherrington NJ. 2017. Metabolomic profiling distinction of human nonalcoholic fatty liver disease progression from a common rat model. Obesity 25:1069–1076.

76. Muraki Y, Makita Y, Yamasaki M, Amano Y, Matsuo T. 2017. Elevation of liver endoplasmic reticulum stress in a modified choline-deficient l -amino acid-defined diet-fed non-alcoholic steatohepatitis mouse model. Biochem Biophys Res Commun 486:632–638.

77. RUTENBURG AM, SONNENBLICK E, KOVEN I, APRAHAMIAN HA, REINER L, FINE J. 1957. The role of intestinal bacteria in the development of dietary cirrhosis in rats. J Exp Med 106:1–14.

78. Mehedint MG, Zeisel SH. 2013. Choline's role in maintaining liver function: new evidence for epigenetic mechanisms. Curr Opin Clin Nutr Metab Care 16:339–45.

79. Velasquez M, Ramezani A, Manal A, Raj D. 2016. Trimethylamine N-Oxide: The Good, the Bad and the Unknown. Toxins (Basel) 8:326.

80. Del Rio D, Zimetti F, Caffarra P, Tassotti M, Bernini F, Brighenti F, Zini A, Zanotti I. 2017. The Gut Microbial Metabolite Trimethylamine-N-Oxide Is Present in Human Cerebrospinal Fluid. Nutrients 9:1053.

81. Spencer MD, Hamp TJ, Reid RW, Fischer LM, Zeisel SH, Fodor AA. 2011. Association Between Composition of the Human Gastrointestinal Microbiome and Development of Fatty Liver With Choline Deficiency. Gastroenterology 140:976–986.

82. Gogiashvili M, Edlund K, Gianmoena K, Marchan R, Brik A, Andersson JT, Lambert J, Madjar K, Hellwig B, Rahnenführer J, Hengstler JG, Hergenröder R, Cadenas C. 2017. Metabolic profiling of ob/ob mouse fatty liver using HR-MAS 1H-NMR combined with gene expression analysis reveals alterations in betaine metabolism and the transsulfuration pathway. Anal Bioanal Chem 409:1591–1606.

83. Sherriff JL, OSullivan TA, Properzi C, Oddo J-L, Adams LA. 2016. Choline, Its Potential Role in Nonalcoholic Fatty Liver Disease, and the Case for Human and Bacterial Genes. Adv Nutr An Int Rev J 7:5–13.

84. Chen P, Torralba M, Tan J, Embree M, Zengler K, Stärkel P, van Pijkeren J-P, DePew J, Loomba R, Ho SB, Bajaj JS, Mutlu EA, Keshavarzian A, Tsukamoto H, Nelson KE, Fouts DE, Schnabl B. 2015. Supplementation of Saturated Long-Chain Fatty Acids Maintains Intestinal Eubiosis and Reduces Ethanol-induced Liver Injury in Mice. Gastroenterology 148:203-214.e16.

85. Cresci GA, Glueck B, McMullen MR, Xin W, Allende D, Nagy LE. 2017. Prophylactic tributyrin treatment mitigates chronic-binge alcohol-induced intestinal barrier and liver injury. J Gastroenterol Hepatol.

86. Hamarneh SR, Kim B-M, Kaliannan K, Morrison SA, Tantillo TJ, Tao Q, Mohamed MMR, Ramirez JM, Karas A, Liu W, Hu D, Teshager A, Gul SS, Economopoulos KP, Bhan AK, Malo MS, Choi MY, Hodin RA. 2017. Intestinal Alkaline Phosphatase Attenuates Alcohol-Induced Hepatosteatosis in Mice. Dig Dis Sci 62:2021–2034.

87. Chen P, Miyamoto Y, Mazagova M, Lee K-C, Eckmann L, Schnabl B. 2015. Microbiota Protects Mice Against Acute Alcohol-Induced Liver Injury. Alcohol Clin Exp Res 39:2313–2323.

88. Ansari R, Husain K, Rizvi S. 2016. Role of Transcription Factors in Steatohepatitis and Hypertension after Ethanol: The Epicenter of Metabolism. Biomolecules 6:29.

89. Shi X, Wei X, Yin X, Wang Y, Zhang M, Zhao C, Zhao H, McClain CJ, Feng W, Zhang X. 2015. Hepatic and Fecal Metabolomic Analysis of the Effects of Lactobacillus rhamnosus GG on Alcoholic Fatty Liver Disease in Mice. J Proteome Res 14:1174–1182.

90. Kim D-H, Jeong D, Kang I-B, Kim H, Song K-Y, Seo K-H. 2017. Dual function of Lactobacillus kefiri DH5 in preventing high-fat-diet-induced obesity: direct reduction of cholesterol and upregulation of PPARα in adipose tissue. Mol Nutr Food Res 1700252.

91. Nanji AA, Khettry U, Sadrzadeh SM. 1994. Lactobacillus feeding reduces endotoxemia and severity of experimental alcoholic liver (disease). Proc Soc Exp Biol Med 205:243–7.

92. Forsyth CB, Farhadi A, Jakate SM, Tang Y, Shaikh M, Keshavarzian A. 2009. Lactobacillus GG treatment ameliorates alcohol-induced intestinal oxidative stress, gut leakiness, and liver injury in a rat model of alcoholic steatohepatitis. Alcohol 43:163–172.

93. Loguercio C, Federico A, Tuccillo C, Terracciano F, D'Auria MV, De Simone C, Del Vecchio Blanco C. 2005. Beneficial effects of a probiotic VSL#3 on parameters of liver dysfunction in chronic liver diseases. J Clin Gastroenterol 39:540–3.

94. Kirpich IA, Solovieva N V., Leikhter SN, Shidakova NA, Lebedeva O V., Sidorov PI, Bazhukova TA, Soloviev AG, Barve SS, McClain CJ, Cave M. 2008. Probiotics restore bowel flora and improve liver enzymes in human alcohol-induced liver injury: a pilot study. Alcohol 42:675–682.

95. Stadlbauer V, Mookerjee RP, Hodges S, Wright GAK, Davies NA, Jalan R. 2008. Effect of probiotic treatment on deranged neutrophil function and cytokine responses in patients with compensated alcoholic cirrhosis. J Hepatol 48:945–951.

96. Chen R-C, Xu L-M, Du S-J, Huang S-S, Wu H, Dong J-J, Huang J-R, Wang X-D, Feng W-K, Chen Y-P. 2016. Lactobacillus rhamnosus GG supernatant promotes intestinal barrier function, balances T reg and T H 17 cells and ameliorates hepatic injury in a mouse model of chronic-binge alcohol feeding. Toxicol Lett 241:103–110.

97. Levitt MD, Li R, DeMaster EG, Elson M, Furne J, Levitt DG. 1997. Use of measurements of ethanol absorption from stomach and intestine to assess human ethanol metabolism. Am J Physiol 273:G951-7.

98. Norberg A, Jones AW, Hahn RG, Gabrielsson JL. 2003. Role of Variability in Explaining Ethanol Pharmacokinetics. Clin Pharmacokinet 42:1–31.

99. Setshedi M, Wands JR, de la Monte SM. 2010. Acetaldehyde Adducts in Alcoholic Liver Disease. Oxid Med Cell Longev 3:178–185.

100. Rao RK. 2008. Acetaldehyde-induced Barrier Disruption and Paracellular Permeability in Caco-2 Cell Monolayer, p. 171–183. In Methods in molecular biology (Clifton, N.J.).

101. Mir H, Meena AS, Chaudhry KK, Shukla PK, Gangwar R, Manda B, Padala MK, Shen L, Turner JR, Dietrich P, Dragatsis I, Rao R. 2016. Occludin deficiency promotes ethanol-induced disruption of colonic epithelial junctions, gut barrier dysfunction and liver damage in mice. Biochim Biophys Acta - Gen Subj 1860:765–774.

102. Chaudhry KK, Shukla PK, Mir H, Manda B, Gangwar R, Yadav N, McMullen M, Nagy LE, Rao R. 2016. Glutamine supplementation attenuates ethanol-induced disruption of apical junctional complexes in colonic epithelium and ameliorates gut barrier dysfunction and fatty liver in mice. J Nutr Biochem 27:16–26.

103. Chen P, Stärkel P, Turner JR, Ho SB, Schnabl B. 2015. Dysbiosis-induced intestinal inflammation activates tumor necrosis factor receptor I and mediates alcoholic liver disease in mice. Hepatology 61:883–894.

104. Forsyth CB, Voigt RM, Burgess HJ, Swanson GR, Keshavarzian A. 2015. Circadian rhythms, alcohol and gut interactions. Alcohol 49:389–98.

105. Yan AW, Schnabl B. 2012. Bacterial translocation and changes in the intestinal microbiome associated with alcoholic liver disease. World J Hepatol 4:110.

106. Yan AW, E. Fouts D, Brandl J, Stärkel P, Torralba M, Schott E, Tsukamoto H, E. Nelson K, A. Brenner D, Schnabl B. 2011. Enteric dysbiosis associated with a mouse model of alcoholic liver disease. Hepatology 53:96–105.

107. Hartmann P, Chen P, Wang HJ, Wang L, Mccole DF, Brandl K, Stärkel P, Belzer C, Hellerbrand C, Tsukamoto H, Ho SB, Schnabl B. 2013. Deficiency of intestinal mucin-2 ameliorates experimental alcoholic liver disease in mice. Hepatology 58:108–119.

108. Park B, Lee H-R, Lee Y-J. 2016. Alcoholic liver disease: focus on prodromal gut health. J Dig Dis 17:493–500.

109. Wang H, Lafdil F, Kong X, Gao B. 2011. Signal transducer and activator of transcription 3 in liver diseases: a novel therapeutic target. Int J Biol Sci 7:536–50.

110. Mottaran E, Stewart SF, Rolla R, Vay D, Cipriani V, Moretti M, Vidali M, Sartori M, Rigamonti C, Day CP, Albano E. 2002. Lipid peroxidation contributes to immune reactions associated with alcoholic liver disease. Free Radic Biol Med 32:38–45.

111. Xie G, Zhong W, Zheng X, Li Q, Qiu Y, Li H, Chen H, Zhou Z, Jia W. 2013. Chronic Ethanol Consumption Alters Mammalian Gastrointestinal Content Metabolites. J Proteome Res 12:3297–3306.

112. Couch RD, Dailey A, Zaidi F, Navarro K, Forsyth CB, Mutlu E, Engen PA, Keshavarzian A. 2015. Alcohol Induced Alterations to the Human Fecal VOC Metabolome. PLoS One 10:e0119362.

113. Leclercq S, Matamoros S, Cani PD, Neyrinck AM, Jamar F, Stärkel P, Windey K, Tremaroli V, Bäckhed F, Verbeke K, de Timary P, Delzenne NM. 2014. Intestinal permeability, gut-bacterial dysbiosis, and behavioral markers of alcohol-dependence severity. Proc Natl Acad Sci 111:E4485–E4493.

114. Arroyo V, Moreau R, Kamath PS, Jalan R, Ginès P, Nevens F, Fernández J, To U, García-Tsao G, Schnabl B. 2016. Acute-on-chronic liver failure in cirrhosis. Nat Rev Dis Prim 2:16041.

115. Cresci GA, Bush K, Nagy LE. 2014. Tributyrin supplementation protects mice from acute ethanol-induced gut injury. Alcohol Clin Exp Res 38:1489–501.

116. Spengler EK, Loomba R. 2015. Recommendations for Diagnosis, Referral for Liver Biopsy, and Treatment of Nonalcoholic Fatty Liver Disease and Nonalcoholic Steatohepatitis. Mayo Clin Proc 90:1233–1246.

117. Loomba R, Abraham M, Unalp A. 2012. Association between diabetes, family history of diabetes and risk of nonalcoholic steatohepatitis and fibrosis. Hepatology 56:943–951.

118. Doycheva I, Cui J, Nguyen P, Costa EA, Hooker J, Hofflich H, Bettencourt R, Brouha S, Sirlin CB, Loomba R. 2016. Non-invasive screening of diabetics in primary care for NAFLD and advanced fibrosis by MRI and MRE. Aliment Pharmacol Ther 43:83–95.

119. Loomba R, Schork N, Chen C-H, Bettencourt R, Bhatt A, Ang B, Nguyen P, Hernandez C, Richards L, Salotti J, Lin S, Seki E, Nelson KE, Sirlin CB, Brenner D, Genetics of NAFLD in Twins Consortium. 2015. Heritability of Hepatic Fibrosis and Steatosis Based on a Prospective Twin Study. Gastroenterology 149:1784–1793.

120. Cui J, Chen C-H, Lo M-T, Schork N, Bettencourt R, Gonzalez MP, Bhatt A, Hooker J, Shaffer K, Nelson KE, Long MT, Brenner DA, Sirlin CB, Loomba R, for the Genetics of NAFLD in Twins. 2016. Shared genetic effects between hepatic steatosis and fibrosis: A prospective twin study. Hepatology 64:1547–1558.

121. Caussy C, Soni M, Cui J, Bettencourt R, Schork N, Chen C-H, Ikhwan M Al, Bassirian S, Cepin S, Gonzalez MP, Mendler M, Kono Y, Vodkin I, Mekeel K, Haldorson J, Hemming A, Andrews B, Salotti J, Richards L, Brenner DA, Sirlin CB, Loomba R, Familial NAFLD Cirrhosis Research Consortium. 2017. Nonalcoholic fatty liver disease with cirrhosis increases familial risk for advanced fibrosis. J Clin Invest 127:2697–2704.

122. Gao B, Bataller R. 2011. Alcoholic Liver Disease: Pathogenesis and New Therapeutic Targets. Gastroenterology 141:1572–1585.

123. Wieland A, Frank DN, Harnke B, Bambha K. 2015. Systematic review: microbial dysbiosis and nonalcoholic fatty liver disease. Aliment Pharmacol Ther 42:1051–1063.

124. Kapil S, Duseja A, Sharma BK, Singla B, Chakraborti A, Das A, Ray P, Dhiman RK, Chawla Y. 2016. Small intestinal bacterial overgrowth and toll-like receptor signaling in patients with non-alcoholic fatty liver disease. J Gastroenterol Hepatol 31:213–21.

125. Boursier J, Mueller O, Barret M, Machado M, Fizanne L, Araujo-Perez F, Guy CD, Seed PC, Rawls JF, David LA, Hunault G, Oberti F, Calès P, Diehl AM. 2016. The severity of nonalcoholic fatty liver disease is associated with gut dysbiosis and shift in the metabolic function of the gut microbiota. Hepatology 63:764–775.

126. Bajaj JS, Betrapally NS, Hylemon PB, Heuman DM, Daita K, White MB, Unser A, Thacker LR, Sanyal AJ, Kang DJ, Sikaroodi M, Gillevet PM. 2015. Salivary microbiota reflects changes in gut microbiota in cirrhosis with hepatic encephalopathy. Hepatology 62:1260–1271.

127. Rahman K, Desai C, Iyer SS, Thorn NE, Kumar P, Liu Y, Smith T, Neish AS, Li H, Tan S, Wu P, Liu X, Yu Y, Farris AB, Nusrat A, Parkos CA, Anania FA. 2016. Loss of Junctional Adhesion Molecule A Promotes Severe Steatohepatitis in Mice on a Diet High in Saturated Fat, Fructose, and Cholesterol. Gastroenterology 151:733-746.e12.

128. Arendt BM, Ma DW, Simons B, Noureldin SA, Therapondos G, Guindi M, Sherman M, Allard JP. 2013. Nonalcoholic fatty liver disease is associated with lower hepatic and erythrocyte ratios of phosphatidylcholine to phosphatidylethanolamine. Appl Physiol Nutr Metab 38:334–340.

129. Rao RK, Seth A, Sheth P. 2004. Recent Advances in Alcoholic Liver Disease I. Role of intestinal permeability and endotoxemia in alcoholic liver disease. AJP Gastrointest Liver Physiol 286:G881–G884.

130. Ferrier L, Bérard F, Debrauwer L, Chabo C, Langella P, Buéno L, Fioramonti J. 2006. Impairment of the Intestinal Barrier by Ethanol Involves Enteric Microflora and Mast Cell Activation in Rodents. Am J Pathol 168:1148–1154.

131. Ferrere G, Wrzosek L, Cailleux F, Turpin W, Puchois V, Spatz M, Ciocan D, Rainteau D, Humbert L, Hugot C, Gaudin F, Noordine M-L, Robert V, Berrebi D, Thomas M, Naveau S, Perlemuter G, Cassard A-M. 2017. Fecal microbiota manipulation prevents dysbiosis and alcohol-induced liver injury in mice. J Hepatol 66:806–815.

132. Mutlu EA, Gillevet PM, Rangwala H, Sikaroodi M, Naqvi A, Engen PA, Kwasny M, Lau CK, Keshavarzian A. 2012. Colonic microbiome is altered in alcoholism. AJP Gastrointest Liver Physiol 302:G966–G978.

133. Tuomisto S, Pessi T, Collin P, Vuento R, Aittoniemi J, Karhunen PJ. 2014. Changes in gut bacterial populations and their translocation into liver and ascites in alcoholic liver cirrhotics. BMC Gastroenterol 14:40.

134. Chen Y, Yang F, Lu H, Wang B, Chen Y, Lei D, Wang Y, Zhu B, Li L. 2011. Characterization of fecal microbial communities in patients with liver cirrhosis. Hepatology 54:562–72.

135. Wang L, Fouts DE, Stärkel P, Hartmann P, Chen P, Llorente C, DePew J, Moncera K, Ho SB, Brenner DA, Hooper LV, Schnabl B. 2016. Intestinal REG3 Lectins Protect against Alcoholic Steatohepatitis by Reducing Mucosa-Associated Microbiota and Preventing Bacterial Translocation. Cell Host Microbe 19:227–239.

136. Inamine T, Yang A-M, Wang L, Lee K-C, Llorente C, Schnabl B. 2016. Genetic Loss of Immunoglobulin A Does Not Influence Development of Alcoholic Steatohepatitis in Mice. Alcohol Clin Exp Res 40:2604–2613.

137. Adachi Y, Bradford BU, Gao W, Bojes HK, Thurman RG. 1994. Inactivation of Kupffer cells prevents early alcohol-induced liver injury. Hepatology 20:453–60.

138. Seo W, Jeong W Il. 2016. Hepatic non-parenchymal cells: Master regulators of alcoholic liver disease? World J Gastroenterol 22:1348–1356.

139. Ju C, Mandrekar P. 2015. Macrophages and Alcohol-Related Liver Inflammation. Alcohol Res 37:251–62.

140. Tilg H, Moschen AR, Szabo G. 2016. Interleukin-1 and inflammasomes in alcoholic liver disease/acute alcoholic hepatitis and nonalcoholic fatty liver disease/nonalcoholic steatohepatitis. Hepatology 64:955–965.

141. Axelson M, Mörk B, Sjövall J. 1991. Ethanol has an acute effect on bile acid biosynthesis in man. FEBS Lett 281:155–9.

142. Xie G, Zhong W, Li H, Li Q, Qiu Y, Zheng X, Chen H, Zhao X, Zhang S, Zhou Z, Zeisel SH, Jia W. 2013. Alteration of bile acid metabolism in the rat induced by chronic ethanol consumption. FASEB J 27:3583–3593.

143. Wu W-B, Chen Y-Y, Zhu B, Peng X-M, Zhang S-W, Zhou M-L. 2015. Excessive bile acid activated NF-kappa B and promoted the development of alcoholic steatohepatitis in farnesoid X receptor deficient mice. Biochimie 115:86–92.

144. Wu W-B, Xu Y-Y, Cheng W-W, Wang Y-X, Liu Y, Huang D, Zhang H-J. 2015. Agonist of farnesoid X receptor protects against bile acid induced damage and oxidative stress in mouse placenta – A study on maternal cholestasis model. Placenta 36:545–551.

145. Bhat M, Arendt BM, Bhat V, Renner EL, Humar A, Allard JP. 2016. Implication of the intestinal microbiome in complications of cirrhosis. World J Hepatol 8:1128–1136.

146. Mells GF, Floyd JAB, Morley KI, Cordell HJ, Franklin CS, Shin S-Y, Heneghan MA, Neuberger JM, Donaldson PT, Day DB, Ducker SJ, Muriithi AW, Wheater EF, Hammond CJ, Dawwas MF, Jones DE, Peltonen L, Alexander GJ, Sandford RN, Anderson CA, Sandford RN, Anderson CA. 2011. Genome-wide association study identifies 12 new susceptibility loci for primary biliary cirrhosis. Nat Genet 43:329–332.

147. Charlton MR, Burns JM, Pedersen RA, Watt KD, Heimbach JK, Dierkhising RA. 2011. Frequency and Outcomes of Liver Transplantation for Nonalcoholic Steatohepatitis in the United States. Gastroenterology 141:1249–1253.

148. Yang JD, Mohamed HA, Cvinar JL, Gores GJ, Roberts LR, Kim WR. 2016. Diabetes Mellitus Heightens the Risk of Hepatocellular Carcinoma Except in Patients With Hepatitis C Cirrhosis. Am J Gastroenterol 111:1573–1580.

149. Bajaj JS, Betrapally NS, Hylemon PB, Thacker LR, Daita K, Kang DJ, White MB, Unser AB, Fagan A, Gavis EA, Sikaroodi M, Dalmet S, Heuman DM, Gillevet PM. 2016. Gut Microbiota Alterations can predict Hospitalizations in Cirrhosis Independent of Diabetes Mellitus. Sci Rep 5:18559.

150. Jun DW, Kim KT, Lee OY, Chae JD, Son BK, Kim SH, Jo YJ, Park YS. 2010. Association Between Small Intestinal Bacterial Overgrowth and Peripheral Bacterial DNA in Cirrhotic Patients. Dig Dis Sci 55:1465–1471.

151. Yao J, Chang L, Yuan L, Duan Z. 2016. Nutrition status and small intestinal bacterial overgrowth in patients with virus-related cirrhosis. Asia Pac J Clin Nutr 25:283–91.

152. Chen Y, Ji F, Guo J, Shi D, Fang D, Li L. 2016. Dysbiosis of small intestinal microbiota in liver cirrhosis and its association with etiology. Sci Rep 6:34055.

153. Mas A, Rodés J, Sunyer L, Rodrigo L, Planas R, Vargas V, Castells L, Rodríguez-Martínez D, Fernández-Rodríguez C, Coll I, Pardo A, Spanish Association for the Study of the Liver Hepatic Encephalopathy Cooperative Group. 2003. Comparison of rifaximin and lactitol in the treatment of acute hepatic encephalopathy: results of a randomized, double-blind, double-dummy, controlled clinical trial. J Hepatol 38:51–8.

154. Bajaj JS, Heuman DM, Wade JB, Gibson DP, Saeian K, Wegelin JA, Hafeezullah M, Bell DE, Sterling RK, Stravitz RT, Fuchs M, Luketic V, Sanyal AJ. 2011. Rifaximin Improves Driving Simulator Performance in a Randomized Trial of Patients With Minimal Hepatic Encephalopathy. Gastroenterology 140:478-487.e1.

155. Vlachogiannakos J, Viazis N, Vasianopoulou P, Vafiadis I, Karamanolis DG, Ladas SD. 2013. Long-term administration of rifaximin improves the prognosis of patients with decompensated alcoholic cirrhosis. J Gastroenterol Hepatol 28:450–455.

156. Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, Guo J, Le Chatelier E, Yao J, Wu L, Zhou J, Ni S, Liu L, Pons N, Batto J-M, Kennedy S, Leonard P, Yuan C, Ding W, Chen Y, Hu X, Zheng B, Qian G, Xu W, Ehrlich S, Zheng S, Li L. 2014. Alterations of the human gut microbiome in liver cirrhosis. Nature 513:59–64.

157. Bajaj JS, Liu EJ, Kheradman R, Fagan A, Heuman DM, White M, Gavis EA, Hylemon P, Sikaroodi M, Gillevet PM. 2017. Fungal dysbiosis in cirrhosis. Gut gutjnl-2016-313170.

158. Fouts DE, Torralba M, Nelson KE, Brenner DA, Schnabl B. 2012. Bacterial translocation and changes in the intestinal microbiome in mouse models of liver disease. J Hepatol 56:1283–1292.

159. Yoshimoto S, Loo TM, Atarashi K, Kanda H, Sato S, Oyadomari S, Iwakura Y, Oshima K, Morita H, Hattori M, Honda K, Ishikawa Y, Hara E, Ohtani N, Ohtani N. 2013. Obesity-induced gut microbial metabolite promotes liver cancer through senescence secretome. Nature 499:97–101.

160. Xie G, Wang X, Liu P, Wei R, Chen W, Rajani C, Hernandez BY, Alegado R, Dong B, Li D, Jia W. 2016. Distinctly altered gut microbiota in the progression of liver disease. Oncotarget 7:19355–19366.

161. Grąt M, Krasnodębski M, Patkowski W, Wronka KM, Masior Ł, Stypułkowski J, Grąt K, Krawczyk M. 2016. Relevance of Pre-Transplant α-fetoprotein Dynamics in Liver Transplantation for Hepatocellular Cancer. Ann Transplant 21:115–24.

162. Fox JG, Feng Y, Theve EJ, Raczynski AR, Fiala JLA, Doernte AL, Williams M, McFaline JL, Essigmann JM, Schauer DB, Tannenbaum SR, Dedon PC, Weinman SA, Lemon SM, Fry RC, Rogers AB. 2010. Gut microbes define liver cancer risk in mice exposed to chemical and viral transgenic hepatocarcinogens. Gut 59:88–97.

163. Rogers AB. 2011. Distance burning. Gut Microbes 2:52–57.

164. Huang Y, Fan X-G, Wang Z-M, Zhou J-H, Tian X-F, Li N. 2004. Identification of helicobacter species in human liver samples from patients with primary hepatocellular carcinoma. J Clin Pathol 57:1273–1277.

165. Krüttgen A, Horz H-P, Weber-Heynemann J, Vucur M, Trautwein C, Haase G, Luedde T, Roderburg C. 2012. Study on the association of helicobacter species with viral hepatitis-induced hepatocellular carcinoma. Gut Microbes 3:228–233.

166. Ward JM, Fox JG, Anver MR, Haines DC, George C V, Collins MJ, Gorelick PL, Nagashima K, Gonda MA, Gilden R V. 1994. Chronic active hepatitis and associated liver tumors in mice caused by a persistent bacterial infection with a novel Helicobacter species. J Natl Cancer Inst 86:1222–7.

167. Mima K, Nakagawa S, Sawayama H, Ishimoto T, Imai K, Iwatsuki M, Hashimoto D, Baba Y, Yamashita Y, Yoshida N, Chikamoto A, Baba H. 2017. The microbiome and hepatobiliary-pancreatic cancers. Cancer Lett 402:9–15.

168. Brandtzaeg P. 2013. Secretory IgA: Designed for Anti-Microbial Defense. Front Immunol 4:222.

169. Shalapour S, Lin X-J, Bastian IN, Brain J, Burt AD, Aksenov AA, Vrbanac AF, Li W, Perkins A, Matsutani T, Zhong Z, Dhar D, Navas-Molina JA, Xu J, Loomba R, Downes M, Yu RT, Evans RM, Dorrestein PC, Knight R, Benner C, Anstee QM, Karin M. 2017. Inflammation-induced IgA+ cells dismantle anti-liver cancer immunity. Nature 551:340–345.

170. Dapito DH, Mencin A, Gwak G-Y, Pradere J-P, Jang M-K, Mederacke I, Caviglia JM, Khiabanian H, Adeyemi A, Bataller R, Lefkowitch JH, Bower M, Friedman R, Sartor RB, Rabadan R, Schwabe RF. 2012. Promotion of Hepatocellular Carcinoma by the Intestinal Microbiota and TLR4. Cancer Cell 21:504–516.

171. Xie G, Wang X, Huang F, Zhao A, Chen W, Yan J, Zhang Y, Lei S, Ge K, Zheng X, Liu J, Su M, Liu P, Jia W. 2016. Dysregulated hepatic bile acids collaboratively promote liver carcinogenesis. Int J Cancer 139:1764–1775.

172. Ruhland MK, Loza AJ, Capietto A-H, Luo X, Knolhoff BL, Flanagan KC, Belt BA, Alspach E, Leahy K, Luo J, Schaffer A, Edwards JR, Longmore G, Faccio R, DeNardo DG, Stewart SA. 2016. Stromal senescence establishes an immunosuppressive microenvironment that drives tumorigenesis. Nat Commun 7:11762.

173. Demaria M, O'Leary MN, Chang J, Shao L, Liu S, Alimirah F, Koenig K, Le C, Mitin N, Deal AM, Alston S, Academia EC, Kilmarx S, Valdovinos A, Wang B, de Bruin A, Kennedy BK, Melov S, Zhou D, Sharpless NE, Muss H, Campisi J. 2017. Cellular Senescence Promotes Adverse Effects of Chemotherapy and Cancer Relapse. Cancer Discov 7:165–176.

174. Gomes AL, Teijeiro A, Burén S, Tummala KS, Yilmaz M, Waisman A, Theurillat J-P, Perna C, Djouder N. 2016. Metabolic Inflammation-Associated IL-17A Causes Non-alcoholic Steatohepatitis and Hepatocellular Carcinoma. Cancer Cell 30:161–175.

175. Li J, Lau GK-K, Chen L, Dong S, Lan H-Y, Huang X-R, Li Y, Luk JM, Yuan Y-F, Guan X. 2011. Interleukin 17A Promotes Hepatocellular Carcinoma Metastasis via NF-kB Induced Matrix Metalloproteinases 2 and 9 Expression. PLoS One 6:e21816.

176. Hammerich L, Heymann F, Tacke F. 2011. Role of IL-17 and Th17 Cells in Liver Diseases. Clin Dev Immunol 2011:1–12.

177. Viaud S, Saccheri F, Mignot G, Yamazaki T, Daillere R, Hannani D, Enot DP, Pfirschke C, Engblom C, Pittet MJ, Schlitzer A, Ginhoux F, Apetoh L, Chachaty E, Woerther P-L, Eberl G, Berard M, Ecobichon C, Clermont D, Bizet C, Gaboriau-Routhiau V, Cerf-Bensussan N, Opolon P, Yessaad N, Vivier E, Ryffel B, Elson CO, Dore J, Kroemer G, Lepage P, Boneca IG, Ghiringhelli F, Zitvogel L. 2013. The Intestinal Microbiota Modulates the Anticancer Immune Effects of Cyclophosphamide. Science (80-) 342:971–976.

178. Neuschwander-Tetri BA, Loomba R, Sanyal AJ, Lavine JE, Van Natta ML, Abdelmalek MF, Chalasani N, Dasarathy S, Diehl AM, Hameed B, Kowdley K V, McCullough A, Terrault N, Clark JM, Tonascia J, Brunt EM, Kleiner DE, Doo E, NASH Clinical Research Network. 2015. Farnesoid X nuclear receptor ligand obeticholic acid for non-cirrhotic, non-alcoholic steatohepatitis (FLINT): a multicentre, randomised, placebo-controlled trial. Lancet 385:956–965.

179. Thiele M, Wiest R, Gluud LL, Albillos A, Krag A. 2013. Can non-selective beta-blockers prevent hepatocellular carcinoma in patients with cirrhosis? Med Hypotheses 81:871–874.

180. Li J, Sung CYJ, Lee N, Ni Y, Pihlajamäki J, Panagiotou G, El-Nezami H. 2016. Probiotics modulated gut microbiota suppresses hepatocellular carcinoma growth in mice. Proc Natl Acad Sci 113:E1306–E1315.

181. 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. 2015. Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. Nature 528:262–6.

182. Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J, Knights D, Gajer P, Ravel J, Fierer N, Gordon JI, Knight R. 2011. Moving pictures of the human microbiome. Genome Biol 12:R50.

183. Halfvarson J, Brislawn CJ, Lamendella R, Vázquez-Baeza Y, Walters WA, Bramer LM, D'Amato M, Bonfiglio F, McDonald D, Gonzalez A, McClure EE, Dunklebarger MF, Knight R, Jansson JK. 2017. Dynamics of the human gut microbiome in inflammatory bowel disease. Nat Microbiol 2:17004.

184. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen Y-Y, Keilbaugh SA, Bewtra M, Knights D, Walters WA, Knight R, Sinha R, Gilroy E, Gupta K, Baldassano R, Nessel L, Li H, Bushman FD, Lewis JD. 2011. Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. Science (80-) 334:105–108.

185. 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 ML, Bork P, Sönnerborg A, Clotet B, Paredes R. 2016. Gut Microbiota Linked to Sexual Preference and HIV Infection. EBioMedicine 5:135–146.

186. Jackson MA, Goodrich JK, Maxan M-E, Freedberg DE, Abrams JA, Poole AC, Sutter JL, Welter D, Ley RE, Bell JT, Spector TD, Steves CJ. 2016. Proton pump inhibitors alter the composition of the gut microbiota. Gut 65:749–756.

187. Goodrich JK, Waters JL, Poole AC, Sutter JL, Koren O, Blekhman R, Beaumont M, Van Treuren W, Knight R, Bell JT, Spector TD, Clark AG, Ley RE. 2014. Human Genetics Shape the Gut Microbiome. Cell 159:789–799.

188. Debelius J, Song SJ, Vazquez-Baeza Y, Xu ZZ, Gonzalez A, Knight R. 2016. Tiny microbes, enormous impacts: what matters in gut microbiome studies? Genome Biol 17:217.

189. Lozupone CA, Stombaugh J, Gonzalez A, Ackermann G, Wendel D, Vazquez-Baeza Y, Jansson JK, Gordon JI, Knight R. 2013. Meta-analyses of studies of the human microbiota. Genome Res 23:1704–1714.

190. Sinha R, Abnet CC, White O, Knight R, Huttenhower C. 2015. The microbiome quality control project: baseline study design and future directions. Genome Biol 16:276.

191. van Dongen J, Slagboom PE, Draisma HHM, Martin NG, Boomsma DI. 2012. The continuing value of twin studies in the omics era. Nat Rev Genet 13:640–653.

192. Smith MI, Yatsunenko T, Manary MJ, Trehan I, Cheng J, Kau AL, Rich SS, Concannon P, Josyf C, Liu J, Houpt E, Li J V, Holmes E, Nicholson J, Mkakosya R, Cheng J, Kau AL, Rich SS, Concannon P, Mychaleckyj JC, Liu J, Houpt E, Li J V, Holmes E, Nicholson J, Knights D, Ursell LK, Knight R, Gordon JI. 2013. Gut microbes of Malawian twin pairs discordant for kwashiorkor. Science (80-) 339:548–554.

193. Goodrich JK, Davenport ER, Beaumont M, Jackson MA, Knight R, Ober C, Spector TD, Bell JT, Clark AG, Ley RE. 2016. Genetic Determinants of the Gut Microbiome in UK Twins. Cell Host Microbe 19:731–743.

194. Sookoian S, Pirola CJ. 2017. Genetic predisposition in nonalcoholic fatty liver disease. Clin Mol Hepatol 23:1–12.

195. Weingarden A, González A, Vázquez-Baeza Y, Weiss S, Humphry G, Berg-Lyons D, Knights D, Unno T, Bobr A, Kang J, Khoruts A, Knight R, Sadowsky MJ. 2015. Dynamic changes in short- and long-term bacterial composition following fecal microbiota transplantation for recurrent Clostridium difficile infection. Microbiome 3:10.

196. Zaneveld JR, McMinds R, Vega Thurber R. 2017. Stress and stability: applying the Anna Karenina principle to animal microbiomes. Nat Microbiol 2:17121.

197. Piscaglia F, Svegliati-Baroni G, Barchetti A, Pecorelli A, Marinelli S, Tiribelli C, Bellentani S, HCC-NAFLD Italian Study Group. 2016. Clinical patterns of hepatocellular carcinoma in nonalcoholic fatty liver disease: A multicenter prospective study. Hepatology 63:827–838.

198. Kostic AD, Gevers D, Siljander H, Vatanen T, Hyötyläinen T, Hämäläinen A-M, Peet A, Tillmann V, Pöhö P, Mattila I, Lähdesmäki H, Franzosa EA, Vaarala O, de Goffau M, Harmsen H, Ilonen J, Virtanen SM, Clish CB, Orešič M, Huttenhower C, Knip M, Xavier RJ, Xavier RJ. 2015. The Dynamics of the Human Infant Gut Microbiome in Development and in Progression toward Type 1 Diabetes. Cell Host Microbe 17:260–273.

199. Nguyen TLA, Vieira-Silva S, Liston A, Raes J. 2015. How informative is the mouse for human gut microbiota research? Dis Model Mech 8:1–16.

200. Manichanh C, Reeder J, Gibert P, Varela E, Llopis M, Antolin M, Guigo R, Knight R, Guarner F. 2010. Reshaping the gut microbiome with bacterial transplantation and antibiotic intake. Genome Res 20:1411–9.

201. Kiraly DD, Walker DM, Calipari ES, Labonte B, Issler O, Pena CJ, Ribeiro EA, Russo SJ, Nestler EJ. 2016. Alterations of the Host Microbiome Affect Behavioral Responses to Cocaine. Sci Rep 6:35455.

202. Sampson TR, Debelius JW, Thron T, Janssen S, Shastri GG, Ilhan ZE, Challis C, Schretter CE, Rocha S, Gradinaru V, Chesselet M-F, Keshavarzian A, Shannon KM,

Krajmalnik-Brown R, Wittung-Stafshede P, Knight R, Mazmanian SK. 2016. Gut Microbiota Regulate Motor Deficits and Neuroinflammation in a Model of Parkinson's Disease. Cell 167:1469-1480.e12.

203. Markle JGM, Frank DN, Mortin-Toth S, Robertson CE, Feazel LM, Rolle-Kampczyk U, von Bergen M, McCoy KD, Macpherson AJ, Danska JS. 2013. Sex differences in the gut microbiome drive hormone-dependent regulation of autoimmunity. Science 339:1084–8.

204. Llopis M, Cassard AM, Wrzosek L, Boschat L, Bruneau A, Ferrere G, Puchois V, Martin JC, Lepage P, Le Roy T, Lefèvre L, Langelier B, Cailleux F, González-Castro AM, Rabot S, Gaudin F, Agostini H, Prévot S, Berrebi D, Ciocan D, Jousse C, Naveau S, Gérard P, Perlemuter G. 2016. Intestinal microbiota contributes to individual susceptibility to alcoholic liver disease. Gut 65:830–839.

205. Byrne AT, Alférez DG, Amant F, Annibali D, Arribas J, Biankin A V., Bruna A, Budinská E, Caldas C, Chang DK, Clarke RB, Clevers H, Coukos G, Dangles-Marie V, Eckhardt SG, Gonzalez-Suarez E, Hermans E, Hidalgo M, Jarzabek MA, de Jong S, Jonkers J, Kemper K, Lanfrancone L, Mælandsmo GM, Marangoni E, Marine J-C, Medico E, Norum JH, Palmer HG, Peeper DS, Pelicci PG, Piris-Gimenez A, Roman-Roman S, Rueda OM, Seoane J, Serra V, Soucek L, Vanhecke D, Villanueva A, Vinolo E, Bertotti A, Trusolino L. 2017. Interrogating open issues in cancer precision medicine with patient-derived xenografts. Nat Rev Cancer 17:254–268.

206. Hidalgo M, Amant F, Biankin A V, Budinská E, Byrne AT, Caldas C, Clarke RB, de Jong S, Jonkers J, Mælandsmo GM, Roman-Roman S, Seoane J, Trusolino L, Villanueva A. 2014. Patient-derived xenograft models: an emerging platform for translational cancer research. Cancer Discov 4:998–1013.

207. Mattner J. 2016. Impact of Microbes on the Pathogenesis of Primary Biliary Cirrhosis (PBC) and Primary Sclerosing Cholangitis (PSC). Int J Mol Sci 17:1864.

208. Verdier J, Luedde T, Sellge G. 2015. Biliary Mucosal Barrier and Microbiome. Viszeralmedizin 31:156–161.

209. Miyake Y, Yamamoto K. 2013. Role of gut microbiota in liver diseases. Hepatol Res 43:139–146.

210. Pflughoeft KJ, Versalovic J. 2012. Human Microbiome in Health and Disease. Annu Rev Pathol Mech Dis 7:99–122.

211. Bogdanos D-P, Baum H, Okamoto M, Montalto P, Sharma UC, Rigopoulou EI, Vlachogiannakos J, Ma Y, Burroughs AK, Vergani D. 2005. Primary biliary cirrhosis is characterized by IgG3 antibodies cross-reactive with the major mitochondrial autoepitope and its Lactobacillus mimic. Hepatology 42:458–465.

212. PADGETT K, SELMI C, KENNY T, LEUNG P, BALKWILL D, ANSARI A, COPPEL R, GERSHWIN M. 2005. Phylogenetic and immunological definition of four lipoylated proteins from , implications for primary biliary cirrhosis. J Autoimmun 24:209–219.

213. Mohammed JP, Fusakio ME, Rainbow DB, Moule C, Fraser HI, Clark J, Todd JA, Peterson LB, Savage PB, Wills-Karp M, Ridgway WM, Wicker LS, Mattner J. 2011. Identification of Cd101 as a Susceptibility Gene for Novosphingobium aromaticivorans-Induced Liver Autoimmunity. J Immunol 187:337–349.

214. Lee J-Y, Arai H, Nakamura Y, Fukiya S, Wada M, Yokota A. 2013. Contribution of the 7β-hydroxysteroid dehydrogenase from Ruminococcus gnavus N53 to ursodeoxycholic acid formation in the human colon. J Lipid Res 54:3062–3069.

215. Olsson R, Björnsson E, Bäckman L, Friman S, Höckerstedt K, Kaijser B, Olausson M. 1998. Bile duct bacterial isolates in primary sclerosing cholangitis: a study of explanted livers. J Hepatol 28:426–32.

216. Pollheimer MJ, Halilbasic E, Fickert P, Trauner M. 2011. Pathogenesis of primary sclerosing cholangitis. Best Pract Res Clin Gastroenterol 25:727–739.

217. Toyoki Y, Sasaki M, Narumi S, Yoshihara S, Morita T, Konn M. 1998. Semiquantitative evaluation of hepatic fibrosis by measuring tissue hydroxyproline. Hepatogastroenterology 45:2261–4.

218. Karrar A, Broomé U, Södergren T, Jaksch M, Bergquist A, Björnstedt M, Sumitran–Holgersson S. 2007. Biliary Epithelial Cell Antibodies Link Adaptive and Innate Immune Responses in Primary Sclerosing Cholangitis. Gastroenterology 132:1504–1514.

219. Katt J, Schwinge D, Schoknecht T, Quaas A, Sobottka I, Burandt E, Becker C, Neurath MF, Lohse AW, Herkel J, Schramm C. 2013. Increased T helper type 17 response to pathogen stimulation in patients with primary sclerosing cholangitis. Hepatology 58:1084–1093.

220. Loftus E V, Sandborn WJ, Lindor KD, Larusso NF. 1997. Interactions between chronic liver disease and inflammatory bowel disease. Inflamm Bowel Dis 3:288–302.

221. Bode JC, Bode C, Heidelbach R, Dürr HK, Martini GA. 1984. Jejunal microflora in patients with chronic alcohol abuse. Hepatogastroenterology 31:30–4.

222. Bull-Otterson L, Feng W, Kirpich I, Wang Y, Qin X, Liu Y, Gobejishvili L, Joshi-Barve S, Ayvaz T, Petrosino J, Kong M, Barker D, McClain C, Barve S. 2013. Metagenomic analyses of alcohol induced pathogenic alterations in the intestinal microbiome and the effect of Lactobacillus rhamnosus GG treatment. PLoS One 8:e53028.

223. Jiang W, Wu N, Wang X, Chi Y, Zhang Y, Qiu X, Hu Y, Li J, Liu Y. 2015. Dysbiosis gut microbiota associated with inflammation and impaired mucosal immune function in intestine of humans with non-alcoholic fatty liver disease. Sci Rep 5:8096.

224. Raman M, Ahmed I, Gillevet PM, Probert CS, Ratcliffe NM, Smith S, Greenwood R, Sikaroodi M, Lam V, Crotty P, Bailey J, Myers RP, Rioux KP. 2013. Fecal microbiome and volatile organic compound metabolome in obese humans with nonalcoholic fatty liver disease. Clin Gastroenterol Hepatol 11:868-75.e1–3.

225. Bäckhed F, Ding H, Wang T, Hooper L V, Koh GY, Nagy A, Semenkovich CF, Gordon JI. 2004. The gut microbiota as an environmental factor that regulates fat storage. Proc Natl Acad Sci U S A 101:15718–23.

226. Bäckhed F, Manchester JK, Semenkovich CF, Gordon JI. 2007. Mechanisms underlying the resistance to diet-induced obesity in germ-free mice. Proc Natl Acad Sci U S A 104:979–84.

227. Achur RN, Freeman WM, Vrana KE. 2010. Circulating cytokines as biomarkers of alcohol abuse and alcoholism. J Neuroimmune Pharmacol 5:83–91.

228. Luck H, Tsai S, Chung J, Clemente-Casares X, Ghazarian M, Revelo XS, Lei H, Luk CT, Shi SY, Surendra A, Copeland JK, Ahn J, Prescott D, Rasmussen BA, Chng MHY, Engleman EG, Girardin SE, Lam TKT, Croitoru K, Dunn S, Philpott DJ, Guttman DS, Woo M, Winer S, Winer DA. 2015. Regulation of obesity-related insulin resistance with gut anti-inflammatory agents. Cell Metab 21:527–42.

229. Luther J, Garber JJ, Khalili H, Dave M, Bale SS, Jindal R, Motola DL, Luther S, Bohr S, Jeoung SW, Deshpande V, Singh G, Turner JR, Yarmush ML, Chung RT, Patel SJ. 2015. Hepatic Injury in Nonalcoholic Steatohepatitis Contributes to Altered Intestinal Permeability. Cell Mol Gastroenterol Hepatol 1:222–232.

230. Le Roy T, Llopis M, Lepage P, Bruneau A, Rabot S, Bevilacqua C, Martin P, Philippe C, Walker F, Bado A, Perlemuter G, Cassard-Doulcier A-M, Gérard P. 2013. Intestinal microbiota determines development of non-alcoholic fatty liver disease in mice. Gut 62:1787–94.

231. Bala S, Marcos M, Gattu A, Catalano D, Szabo G. 2014. Acute binge drinking increases serum endotoxin and bacterial DNA levels in healthy individuals. PLoS One 9:e96864.

232. Bode C, Kugler V, Bode JC. 1987. Endotoxemia in patients with alcoholic and non-alcoholic cirrhosis and in subjects with no evidence of chronic liver disease following acute alcohol excess. J Hepatol 4:8–14.

233. Parlesak A, Schäfer C, Schütz T, Bode JC, Bode C. 2000. Increased intestinal permeability to macromolecules and endotoxemia in patients with chronic alcohol abuse in different stages of alcohol-induced liver disease. J Hepatol 32:742–7.

234. Roh YS, Zhang B, Loomba R, Seki E. 2015. TLR2 and TLR9 contribute to alcohol-mediated liver injury through induction of CXCL1 and neutrophil infiltration. Am J Physiol Gastrointest Liver Physiol 309:G30-41.

235. Jin R, Willment A, Patel SS, Sun X, Song M, Mannery YO, Kosters A, McClain CJ, Vos MB. 2014. Fructose induced endotoxemia in pediatric nonalcoholic Fatty liver disease. Int J Hepatol 2014:560620.

236. Mridha AR, Haczeyni F, Yeh MM, Haigh WG, Ioannou GN, Barn V, Ajamieh H, Adams L, Hamdorf JM, Teoh NC, Farrell GC. 2017. TLR9 is up-regulated in human and murine NASH: pivotal role in inflammatory recruitment and cell survival. Clin Sci (Lond) 131:2145–2159.

237. Alm R, Carlson J, Eriksson S. 1982. Fasting serum bile acids in liver disease. A comparison with histological features. Scand J Gastroenterol 17:213–8.

238. Ferslew BC, Xie G, Johnston CK, Su M, Stewart PW, Jia W, Brouwer KLR, Barritt AS. 2015. Altered Bile Acid Metabolome in Patients with Nonalcoholic Steatohepatitis. Dig Dis Sci 60:3318–28.

239. Fernando H, Bhopale KK, Kondraganti S, Kaphalia BS, Shakeel Ansari GA. 2011. Lipidomic changes in rat liver after long-term exposure to ethanol. Toxicol Appl Pharmacol 255:127–37.

240. Fernando H, Kondraganti S, Bhopale KK, Volk DE, Neerathilingam M, Kaphalia BS, Luxon BA, Boor PJ, Shakeel Ansari GA. 2010. 1H and 31P NMR lipidome of ethanol-induced fatty liver. Alcohol Clin Exp Res 34:1937–47.

241. Dumas M-E, Barton RH, Toye A, Cloarec O, Blancher C, Rothwell A, Fearnside J, Tatoud R, Blanc V, Lindon JC, Mitchell SC, Holmes E, McCarthy MI, Scott J, Gauguier D, Nicholson JK. 2006. Metabolic profiling reveals a contribution of gut microbiota to fatty liver phenotype in insulin-resistant mice. Proc Natl Acad Sci U S A 103:12511–6.

242. Liu J, Han L, Zhu L, Yu Y. 2016. Free fatty acids, not triglycerides, are associated with non-alcoholic liver injury progression in high fat diet induced obese rats. Lipids Health Dis 15:27.

243. Volynets V, Küper MA, Strahl S, Maier IB, Spruss A, Wagnerberger S, Königsrainer A, Bischoff SC, Bergheim I. 2012. Nutrition, intestinal permeability, and blood ethanol levels are altered in patients with nonalcoholic fatty liver disease (NAFLD). Dig Dis Sci 57:1932–41.

244. Engstler AJ, Aumiller T, Degen C, Dürr M, Weiss E, Maier IB, Schattenberg JM, Jin CJ, Sellmann C, Bergheim I. 2016. Insulin resistance alters hepatic ethanol metabolism: studies in mice and children with non-alcoholic fatty liver disease. Gut 65:1564–71.

245. Nakamura A, Terauchi Y. 2013. Lessons from Mouse Models of High-Fat Diet-Induced NAFLD. Int J Mol Sci 14:21240–21257.

246. Ishioka M, Miura K, Minami S, Shimura Y, Ohnishi H. 2017. Altered Gut Microbiota Composition and Immune Response in Experimental Steatohepatitis Mouse Models. Dig Dis Sci 62:396–406.

247. Lieber CS, DeCarli LM. 1982. The feeding of alcohol in liquid diets: two decades of applications and 1982 update. Alcohol Clin Exp Res 6:523–31.

248. Tsukamoto H, Reidelberger RD, French SW, Largman C. 1984. Long-term cannulation model for blood sampling and intragastric infusion in the rat. Am J Physiol 247:R595-9.

249. Ronis MJJ, Mercer KE, Gannon B, Engi B, Zimniak P, Shearn CT, Orlicky DJ, Albano E, Badger TM, Petersen DR. 2015. Increased 4-hydroxynonenal protein adducts in male GSTA4-4/PPAR-α double knockout mice enhance injury during early stages of alcoholic liver disease. Am J Physiol Gastrointest Liver Physiol 308:G403-15.

250. Ericsson AC, Davis JW, Spollen W, Bivens N, Givan S, Hagan CE, McIntosh M, Franklin CL. 2015. Effects of vendor and genetic background on the composition of the fecal microbiota of inbred mice. PLoS One 10:e0116704.

251. Stappenbeck TS, Virgin HW. 2016. Accounting for reciprocal host-microbiome interactions in experimental science. Nature 534:191–9.

252. Nakagawa H, Hikiba Y, Hirata Y, Font-Burgada J, Sakamoto K, Hayakawa Y, Taniguchi K, Umemura A, Kinoshita H, Sakitani K, Nishikawa Y, Hirano K, Ikenoue T, Ijichi H, Dhar D, Shibata W, Akanuma M, Koike K, Karin M, Maeda S. 2014. Loss of liver E-cadherin induces sclerosing cholangitis and promotes carcinogenesis. Proc Natl Acad Sci 111:1090–1095.

253. Etienne-Mesmin L, Vijay-Kumar M, Gewirtz AT, Chassaing B. 2016. Hepatocyte Toll-Like Receptor 5 Promotes Bacterial Clearance and Protects Mice Against High-Fat Diet-Induced Liver Disease. Cell Mol Gastroenterol Hepatol 2:584–604.

254. Wu T, Heuillard E, Lindner V, Bou About G, Ignat M, Dillenseger J-P, Anton N, Dalimier E, Gossé F, Fouré G, Blindauer F, Giraudeau C, El-Saghire H, Bouhadjar M, Calligaro C, Sorg T, Choquet P, Vandamme T, Ferrand C, Marescaux J, Baumert TF, Diana M, Pessaux P, Robinet E. 2016. Multimodal imaging of a humanized orthotopic model of hepatocellular carcinoma in immunodeficient mice. Sci Rep 6:35230.

255. Round JL, Mazmanian SK. 2009. The gut microbiota shapes intestinal immune responses during health and disease. Nat Rev Immunol 9:313–23.

Chapter 2. A gut microbiome signature for cirrhosis due to nonalcoholic fatty liver disease

The presence of cirrhosis in nonalcoholic-fatty-liver-disease (NAFLD) is the most important predictor of liver-related mortality. Limited data exist concerning the diagnostic accuracy of gut-microbiome-derived signatures for detecting NAFLD-cirrhosis. Here we report 16S gut-microbiome compositions of 203 uniquely well-characterized participants from a prospective twin and family cohort, including 98 probands encompassing the entire spectrum of NAFLD and 105 of their first-degree relatives , assessed by advanced magnetic-resonance-imaging . We show strong familial correlation of gut-microbiome profiles, driven by shared housing. We report a panel of 30 features, including 27 bacterial features with discriminatory ability to detect NAFLD-cirrhosis using a Random Forest classifier model. In a derivation cohort of probands, the model has a robust diagnostic accuracy (AUROC of 0.92) for detecting NAFLD-cirrhosis, confirmed in a validation cohort of relatives of proband with NAFLD-cirrhosis (AUROC of 0.87). This study provides evidence for a fecal-microbiome-derived signature to detect NAFLD-cirrhosis.

2.1 Introduction

NAFLD is the most prevalent cause of chronic liver disease worldwide (1, 2). NAFLD-cirrhosis represents the most severe stage of the disease, carries a significant risk of hepatocellular carcinoma (HCC), and is consistently identified as the most important predictor of liver-related morbidity-mortality in NAFLD (3, 4). However, non-invasive, accurate and easy-to-perform modalities for early detection of NAFLD-cirrhosis remain a major unmet need in the field. Over

the last decade, the gut-liver axis has emerged as a pivotal component of NAFLD (5–10) and represents a potential source of non-invasive biomarkers for the detection and stage of liver disease (11, 12). Limited data are available regarding the diagnostic accuracy of a stool microbiome-derived signature for detecting NAFLD-cirrhosis.

We previously demonstrated that first-degree relatives of probands with NAFLD-cirrhosis have a high risk of AF (13). However, factors associated with progression towards NAFLD-cirrhosis among families remain obscure. Although earlier studies reported familial aggregation of NAFLD and NAFLD-related cirrhosis (14–18), and demonstrated that both liver steatosis and fibrosis are heritable (19, 20), known genetic risk only accounts for ~10-30% of the variance observed in NAFLD (21–24) . This suggests an additional role for environmental factors, which predominate over genetic factors in shaping the human gut-microbiome (25–27). Heritability of gut-microbiome features has been reported in twins studies, but limited data exist regarding the similarity of gut-microbiome composition among family members, and whether similar microbiomes associate with disease traits especially in the entire spectrum of NAFLD including NAFLD-cirrhosis.

2.2 Results

Using a uniquely phenotyped twin and family study design including well-characterized participants with and without NAFLD, assessed using MRI-proton-density-fat-fraction (MRI-PDFF) for quantifying hepatic steatosis (28) and MR-elastography (MRE) for quantifying liver fibrosis (29–32), we examined familial correlation of gut-microbiome composition and tested whether a non-invasive stool-microbiome-derived signature accurately detects NAFLD-cirrhosis. This study leverages from a prospectively recruited case-control study design. This cross-sectional

analysis included 203 well-characterized, prospectively recruited participants, encompassing the entire spectrum of NAFLD divided into three groups (NAFLD-cirrhosis, NAFLD without advanced fibrosis, non-NAFLD controls) paired with their first-degree relatives. Subjects included 26 probands with NAFLD-cirrhosis and 37 of their first-degree relatives, 18 probands with NAFLD (MRI-PDDF ≥ 5%) without advanced fibrosis (AF) (MRE < 3.63 kPa) and 17 of their first-degree relatives, and 54 non-NAFLD normal controls (MRI-PDFF < 5%) and 44 of their first-degree relatives. The detailed derivation of the study cohort is shown in Figure 2.S1. The detailed demographic, biochemical, imaging data of the probands and first-degree relatives stratified by the metabolic and liver phenotype of the probands are provided in Supplementary Table 1 and Supplementary Table 2[1], respectively.

[1] https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-019-09455-9/MediaObjects/41467_2019_9455_MOESM1_ESM.pdf

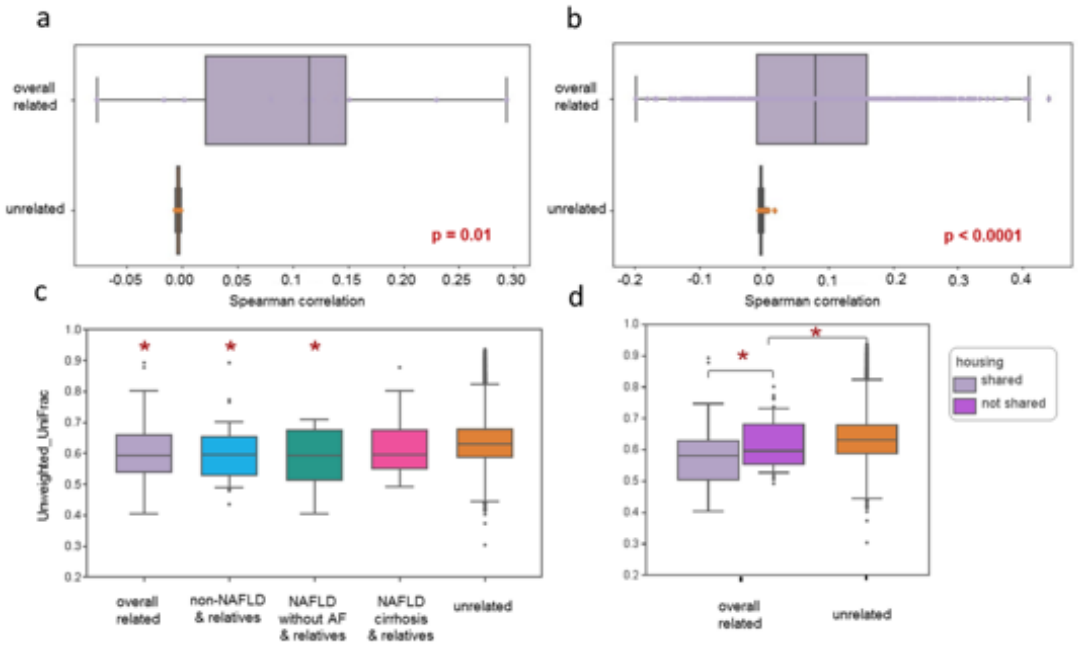

Figure 2.1 *Familial association and shared microbiome among relatives is driven by shared housing*. Distribution of spearman correlation coefficients between relatives (n=86) and unrelated pairs (n=18232) in the familial cohort, plotted for each 16S tag sequence **(a)** and phyla **(b)**. The box plots show the quartiles and whiskers show the rest of the distribution (1.5 inter-quartile range). This analysis was done after filtering rare 16S sequences to avoid spurious correlations due to sparsity (total abundance < 10E-6 across all samples in each disease group). The correlation among related individuals was significantly higher at both 16S tag sequences (p=5E-67) and phylum (p=0.023) levels. Similar plot showing the distribution of unweighted UniFrac distances between related and unrelated pairs stratified by disease status **(c)**. The beta-diversity was significantly lower among related individuals (p=3.22E-05), non-NAFLD controls and relative (n=38 pairs) (p=0.0011) and probands with NAFLD without AF and relatives (n=15) (p=0.0156) when compared to the same among unrelated pairs, while the difference between NALFD-cirrhosis patients and relatives (n=33) and unrelated pairs was not statistically significant (p>0.1). When stratified by shared housing **(d)**, beta-diversity was significantly lower among related individuals sharing a house (n=35 pairs) (p=0.0455). Additionally, related individuals not sharing a house (n=51 pairs) had significantly lower beta-diversity compared to unrelated pairs (p=0.028). Two sided p-values were determined by Kruskal-Wallis test.

We identified a significant familial correlation of the gut-microbiome composition involving shared housing. The gut-microbiome profile showed significant correlation within related pairs compared to random-unrelated pairs at the level of the phyla (p=0.023) Figure 2.1a and at the level of 16S tag sequences (p=5E-67) Figure 2.1b. In our analyses at the phylum level, this familial correlation was mainly driven by significant correlation of Bacteroidetes (r=0.22, p=0.01) and Actinobacteria (0.29, p=0.002) between related individuals. Furthermore,

phylogenetic dissimilarity assessed by unweighted UniFrac distances among related pairs was significantly lower than in random-unrelated pairs (p=3.0E-05). When stratified by the liver phenotype of the proband, the phylogenetic dissimilarity remained significantly lower among non-NAFLD controls and relatives (p=0.001) and probands with NAFLD without AF and relatives (p=0.015) compared to unrelated pairs, while no significant difference was observed among probands with NAFLD-cirrhosis and relatives (p=0.107) Figure 2.1c. These results suggest that familial gut-microbiome similarities are independent of mild/moderate liver phenotype but are impacted by severe stages of liver disease. Finally, related individuals with shared-housing had a lower phylogenetic dissimilarity than those who did not share housing (p=0.045) Figure 2.1d. These results confirm a strong impact of the environment in the familial similarity of the gut-microbiome (25–27) and demonstrate that shared-housing is a major determinant that should be controlled for in study designs assessing the microbiome in liver disease.

The gut-microbiome profile of NAFLD-cirrhosis was first assessed in a *derivation* cohort including the 3 groups of probands encompassing the entire spectrum of NAFLD. As shown in previous studies (11, 12), α-diversity as measured by Faith's phylogenetic diversity decreased with increase in liver damage severity (Figure 2.2a). The β-diversity (unweighted UniFrac distances) was lower among individuals with moderate liver damage (NAFLD without AF) compared to non-NAFLD controls (p=1.1 E-18), whereas it was higher among individuals with severe liver damage (NAFLD-cirrhosis) compared to probands with moderate liver damage (NAFLD without AF) (p=3.3E-15) Figure 2.2b. This suggests an hourglass signature of disease severity in the gut-microbiome, with an initial decrease in phylogenetic diversity associated with a moderate stage of the disease that progress towards a phylogenetic dispersion in individuals with severe stages of disease such as NAFLD-cirrhosis. Further investigations with larger sample sizes are needed to

determine whether this phylogenetic dispersion reflects a distinct profile among NAFLD-cirrhotic patients, and whether it is associated with specific NAFLD-cirrhosis related outcomes.

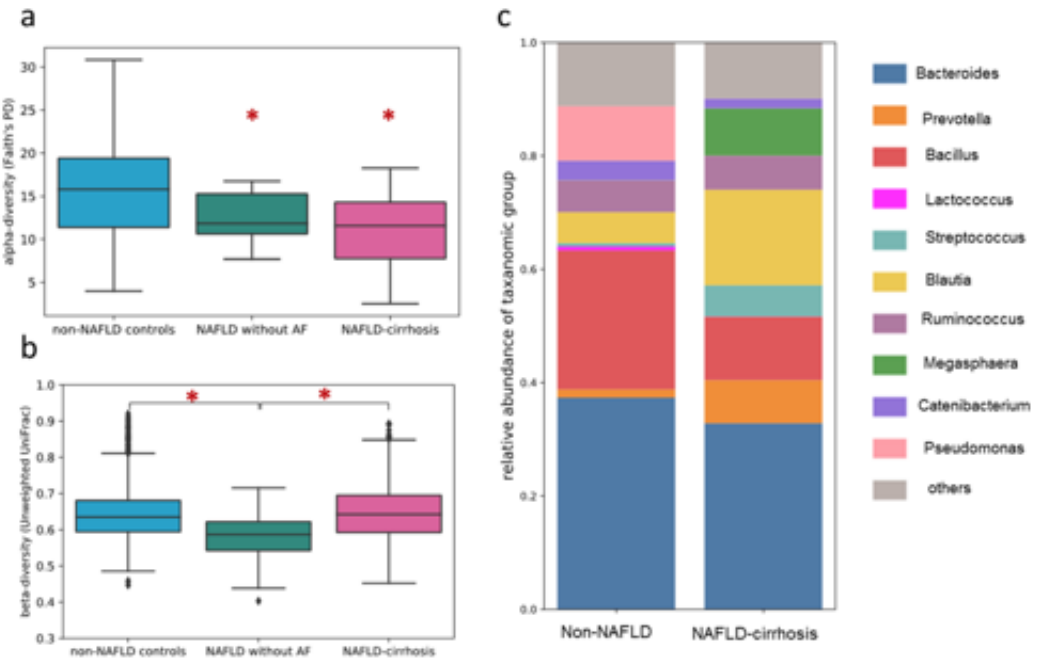

Figure 2.2 *Gut microbiome alteration in NAFLD-cirrhosis*. Comparison between non-NAFLD controls (n=51), NAFLD without advanced fibrosis (n=17), and NAFLD-cirrhosis probands (n=25) with respect to **(a)** alpha-diversity using Faith's Phylogenetic Diversity. Non-NAFLD controls have significantly higher alpha-diversity compared to probands with NAFLD without AF (p=0.0163) and NAFLD-cirrhosis (p=0.0020) **(b)** Similar plot for beta-diversity using unweighted UniFrac distance metric. The beta-diversity among probands with NAFLD without AF was significantly lower than that among non-NAFLD controls (p=1.14E-18) and probands with NAFLD-cirrhosis (p=3.32E-15). The box plots show the quartiles and whiskers show the rest of the distribution (1.5 interquartile range). **(c)** Gut microbiome composition of non-NAFLD controls and NALFD-cirrhosis probands shows differences at bacterial genus level. Two sided p-value were determined by Kruskal-Wallis test.

Several taxa were differentially abundant in NAFLD-cirrhosis compared to non-NAFLD controls. At the genus level, the NAFLD-cirrhosis group was enriched with *Streptococcus* and *Megasphaera,* whereas *Bacillus* and *Lactococcus* were enriched in the non-NAFLD controls Figure 2.2c. Source data are provided as a Source Data file[2]. Species belonging to the family Enterobacteriaceae and the genera *Streptococcus* and *Gallibacterium* were the most enriched in

[2]https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-019-09455-9/MediaObjects/41467_2019_9455_MOESM4_ESM.pdf

NAFLD-cirrhosis, while *Faecalibacterium prausnitzii* and species belonging to the genus, *Catenibacterium* and the families Rikenellaceae, Mogibacterium, Peptostreptococcaceae were enriched in non-NAFLD controls. These results are consistent with the study performed by Ponziani and colleagues in an Italian cohort showing higher Enterobacteriaceae and *Streptococcus* in NAFLD-cirrhosis with and without HCC. In addition, it confirms a shift towards more Gram-negative microbes in advanced fibrosis stages, as previously reported in NAFLD (5, 7, 9).

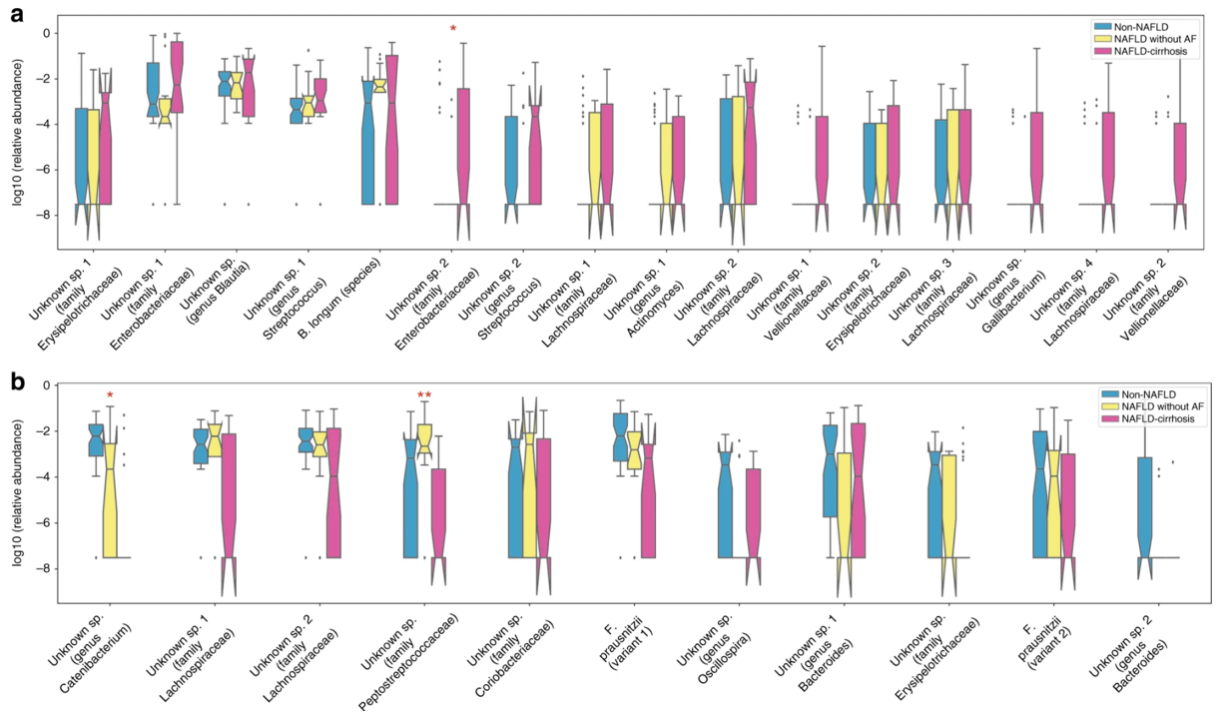

Figure 2.3 *Relative abundance of predictive microbial features in NAFLD-cirrhosis and non-NAFLD controls*. The bacterial features most predictive of NAFLD-cirrhosis (n=25) versus non-NAFLD controls (n=51) sorted by decreasing importance score in the Random Forest classification model. Features increased **(a)** and decreased **(b)** in NAFLD-cirrhosis probands are shown. The feature table is normalized to a total abundance of 1 per sample and relative abundances are plotted on a log10 scale. Each bacterial feature is a unique 16S tag sequence labeled to the highest possible taxonomic rank assigned using QIIME. The box plots show the quartiles and whiskers show the rest of the distribution (1.5 interquartile range). The notches show a 95% confidence interval. Features that are differentially abundant in addition to being important predictors are marked by asterisk (*). Differential abundance was tested using permutation-based, ranked mean test, comparing mean difference between the two groups (33). FDR (<0.1) was controlled using DS-FDR method (34).

A stool-microbiome signature accurately detects NAFLD-cirrhosis. A Random Forest model comprised of 30 features (including 27 bacterial features and age, sex and body mass index

(BMI)) identifies probands with NAFLD-cirrhosis. The bacterial features most important for predicting NAFLD-cirrhosis are shown in Figure 2.3. In a *derivation* cohort of probands, the model had a robust diagnostic accuracy, with an AUROC of 0.92 (±0.05) after cross-validation for detecting NAFLD-cirrhosis Figure 2.4a. The diagnostic accuracy of the model was then confirmed in a *validation* cohort of first-degree relatives of proband with NAFLD-cirrhosis with good diagnostic accuracy, with an AUROC of 0.87 for the detection of advanced fibrosis with a high negative predictive value of 91.6% Figure 2.4b. . In addition, we performed sensitivity analyses in another validation group enriched with mild to moderate stage of NAFLD including probands with NAFLD without AF. The diagnostic accuracy of the model was confirmed and yielded a very good diagnostic accuracy with an AUROC of 86% Figure 2.S2.

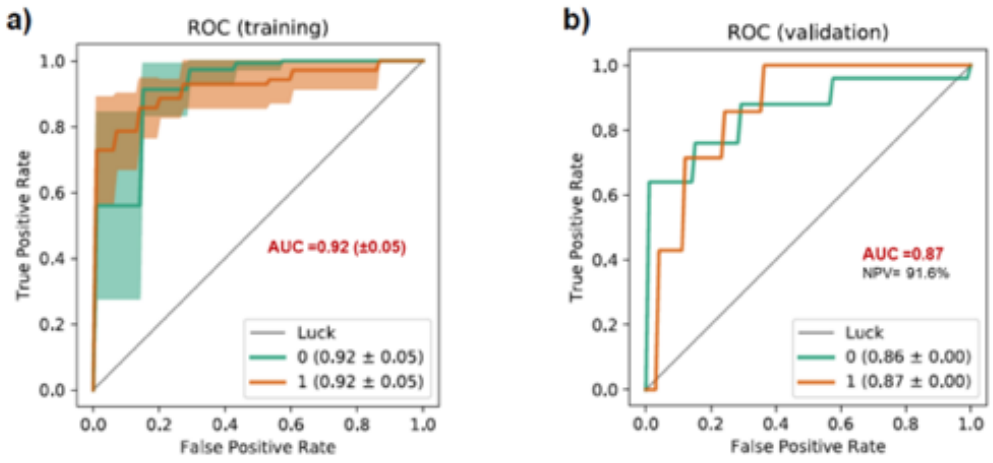

Figure 2.4 *High diagnostic accuracy of a gut-microbiome signature for the detection of NAFLD-cirrhosis*. Receiver operating characteristic (ROC) curves evaluating ability to predict advanced Fibrosis using Random Forest classification. Each curve represents the sensitivity and specificity to distinguish subjects with advanced fibrosis (1) from non-NALFD controls (0). **(a)** Mean ROC curve from cross-validation within training data consisting of NAFLD-cirrhosis probands (n=24) and non-NAFLD controls (n=47). Cross-validation was performed by iteratively (10 times) training the Random Forest model with 70:30 train/test split on this training data. **(b)** ROC curve representing diagnostic accuracy of Random Forest classification model tested on first-degree relatives of NAFLD-cirrhosis probands (n=32). The negative predictive value (NPPV) of the model was 91.6% and the positive predictive value was 62.5%.

In conclusion, using a unique twin and family study design including well-characterized participants with and without NAFLD, we identified a specific stool-microbiome-derived

signature of NAFLD-cirrhosis that yielded a robust diagnostic accuracy for the detection of NAFLD-cirrhosis. Hence, this conveniently assessed microbial biomarker presents an adjunct tool to current invasive approaches to determine the stage of liver disease.

We previously demonstrated that a microbial biomarker can detect AF in biopsy-proven NAFLD (5). The fundamental difference between the previous study and the present study is the clinical context of use of the gut-microbiome signature. In the present study, the clinical question is to accurately differentiate using a non-invasive gut-microbiome signature who among the first-degree relatives have advanced form of NAFLD and who are unaffected in a general setting as opposed to a liver clinic setting. The context of use is critical for biomarker development as suggested by the BEST Guidelines by FDA. In order to address this clinically important question, this study leverages 2 distinct levels of innovations. 1. Study design: This innovative study leverages from a unique prospectively recruited case-control study design. This cross-sectional analysis included 203 well-characterized participants, encompassing the entire spectrum of NAFLD divided into three groups (NAFLD-cirrhosis, NAFLD without advanced fibrosis, non-NAFLD controls) paired with their first-degree relatives. 2. Discovery of shared housing effect: the familial cohort design enabled us to discover the effect of shared housing on the gut-microbiome signature related to NAFLD-cirrhosis. This unique familial cohort study design led us to a new discovery that shared housing had a dominant effect on microbiome. This novel effect would not be apparent from previous study in NAFLD among unrelated individuals. Hence, this study is novel due to its clinical context of use, study design, and co-housing effect on gut microbiome in families with NAFLD-cirrhosis.

We acknowledge the following limitations of this study. This is a single-center study performed in a center with expertise in clinical investigation of NAFLD with advanced MRI-based

phenotyping, and the generalizability of the findings in other clinical settings remains to be established. 16S rRNA sequencing may not have captured additional insights associated with the disease status available at the species or strain level. Finally, the association does not suggest causality, and additional studies are needed to assess whether these microbial species impact gut permeability and/or induce NAFLD progression through cross-talk between serum metabolites and the liver (35). However, the strengths of the study include a prospective study design, detailed phenotyping of participants using the most accurate non-invasive imaging modalities available, and assessment of accuracy using AUROC in both a *derivation* and *validation* cohort. Further multi-center studies including a larger number of individuals are needed to validate the clinical utility of the proposed microbiome-derived signature to detect NAFLD-cirrhosis.

2.3 Materials and methods

Study design

This is a cross-sectional analysis of a prospective family cohort study of participants from the Familial Cirrhosis cohort and Twins and Family cohort who were participating in a biobank initiative and prospectively recruited at the University of California at San Diego (UCSD) NAFLD Research Center between December 2011 and December 2017. All participants underwent a standardized exhaustive clinical research visit including detailed medical history, physical examination, and testing to rule out other causes of chronic liver diseases (see inclusion and exclusion criteria for further details), fasting laboratory tests at the University of California at San Diego (UCSD) NAFLD Research Center (36). Participants also underwent an advanced magnetic resonance examination including magnetic resonance imaging proton-density-fat-fraction (MRI-PDFF) and magnetic resonance elastography (MRE) at the UCSD MR3T Research Laboratory for

the screening of NAFLD and advanced fibrosis (29). Participants from the Familial Cirrhosis cohort also underwent an ultrasound-based vibration controlled transient elastography (VCTE) assessment using a FibroScan. At the time of each research visit, patients provided stool samples. These were collected and immediately stored in a -80*C freezer. Written informed consent was obtained from every participant.

Study participants

Probands with NAFLD-cirrhosis and first-degree relatives: This study included 26 probands with NAFLD-cirrhosis and 37 of their first-degree relatives from the Familial Cirrhosis cohort prospectively recruited at the UCSD NAFLD Research Center[13]. Probands with NAFLD-cirrhosis had documented evidence of NAFLD with either biopsy-proven or meeting imaging criteria for cirrhosis. Definition for NAFLD was based upon American Association for the Study of Liver Study (AASLD) Practice Guidelines (37). The study was approved by the UCSD Institutional Review Board, protocol number 140084.

Inclusion and exclusion criteria of the Familial cirrhosis cohort: Probands and first-degree relatives had to be at least 18 years old. Probands were required to have a documented diagnosis of NAFLD-cirrhosis either by liver biopsy or by documented imaging evidence by a protocol-specified criterion. First-degree relatives (sibling, child, or parent) with written informed consent who did not meet any exclusion criteria were included in the study.

Exclusion criteria included: regular and excessive alcohol consumption within 2 years of recruitment (≥14 drinks/week for men or ≥7 drinks/week for women); use of hepatotoxic drugs or drugs known to cause hepatic steatosis; evidence of liver diseases other than NAFLD, including viral hepatitis (detected with positive serum hepatitis B surface antigen or hepatitis C viral RNA),

Wilson's disease, hemochromatosis, alpha-1 antitrypsin deficiency, autoimmune hepatitis, and cholestatic or vascular liver disease; clinical or laboratory evidence of chronic illnesses associated with hepatic steatosis, including human immunodeficiency virus infection (HIV), celiac disease, cystic fibrosis, lipodystrophy, dysbetalipoproteinemia, and glycogen storage diseases; evidence of active substance abuse, significant systemic illnesses, contraindication(s) to MRI, pregnant or trying to become pregnant, or any other condition which, in the investigator's opinion, may affect the patient's competence or compliance in completing the study.

Proband with NAFLD without advanced fibrosis and non-NAFLD control and first-degree relatives: The study included 140 participants from the Twin and Family study corresponding to 100 twins (50 twin-pairs; 30 monozygotic twin-pairs, 20 dizygotic twin-pairs) and 40 siblings or parents-offspring. The non-NAFLD controls included 54 probands and 44 first-degree relatives and the group with NAFLD without AF included 18 probands and 17 first-degree relatives of community-dwelling controls either twin, sib-sib or parent-offspring pairs(13, 19, 36). These twin, sib-sib, and parent-offspring pairs were prospectively recruited and they reside in southern California. Twins without evidence of NAFLD (MRI-PDFF<5%) and advanced fibrosis (MRE <3.63 kPa) were considered as non-NAFLD control and twins with evidence of NAFLD (MRI-PDFF ≥ 5%) without evidence of advanced fibrosis (MRE <3.63 kPa) and their twin pair were randomly assigned as proband or first-degree relatives. The study was approved by the UCSD Institutional Review Board number 111282.

Inclusion and exclusion criteria for Twin and Family cohort: Patients were included if they were twins, siblings or parent-offspring at least 18 years old, willing and able to complete all research procedures and observations. For each twin pair, a detailed assessment of twinship status (ie, monozygotic (MZ) or dizygotic (DZ)) was obtained. The majority of twin-pairs (34) were

diagnosed by their physician as either MZ or DZ by genetic testing. Furthermore, twin-ship status was confirmed by using a previously published questionnaire (19, 20).

Participants were excluded from the study if they met any of the following criteria: significant alcohol intake (>10 g/day in females or >20 g/day in males) for at least 3 consecutive months over the previous 12 months or if the quantity of alcohol consumed could not be reliably ascertained; clinical or biochemical evidence of liver diseases other than NAFLD (eg, viral hepatitis, HIV, coeliac disease, cystic fibrosis, autoimmune hepatitis); metabolic and/or genetic liver disease (eg, Wilson's disease, haemochromatosis, polycystic liver disease, alpha-1-antitrypsin deficiency, dysbetalipoproteinaemia); clinical or laboratory evidence of systemic infection or any other clinical evidence of liver disease associated with hepatic steatosis; use of drugs known to cause hepatic steatosis (eg amiodarone, glucocorticoids, methotrexate, L-asparaginase and valproic acid) for at least 3 months in the last past 6 months; history of bariatric surgery; presence of systemic infectious illnesses; females who were pregnant or nursing at the time of the study; contraindications to MRI (eg metal implants, severe claustrophobia, body circumference greater than the imaging chamber); any other condition(s) which, based on the principal investigator's opinion, may significantly affect the participant's compliance, competence, or ability to complete the study.

Clinical assessments and laboratory test

All participants underwent a standardized clinical research visit at the UCSD NAFLD Research Center. A detailed history was obtained from all participants. A physical exam, which included vital signs, height, weight, and anthropometric measurements, was performed by a trained clinical investigator. Body mass index was defined as the body weight (in kilograms) divided by height (in meters) squared. Alcohol consumption was documented outside clinical visits and confirmed in the research clinic using the Alcohol Use Disorders Identifications Test and the Skinner questionnaire. A detailed history of medications was obtained and no patient took medications known or suspected to cause steatosis or steatohepatitis. Other causes of liver disease and secondary causes of hepatic steatosis were systemically ruled out using detailed history and laboratory data. After completion of the earlier described elements of the history and physical examination, participants had a comprehensive fasting laboratory including metabolic and liver assessment previously described in references (19, 20, 35, 36).

MRI assessment

MRI was performed at the UCSD MR3T Research Laboratory using the 3T research scanner (GE Signa EXCITE HDxt; GE Healthcare, Waukesha, WI) with all participants in the supine position. MRI-PDFF was used to measure hepatic fat content and MRE was used to measure liver fibrosis. The details of the MRI protocol have been previously described in references (38, 39). The image analysts were blinded to all clinical and biochemical data including the study group of the participants.

Ultrasound-based VCTE assessment

VCTE was performed by a trained technician, using the FibroScan® 502 Touch model (M Probe; XL Probe; Echosens, Paris, France). VCTE measurement was obtained in the supine position with the right arm fully adducted by scanning the area of abdomen at the location of the right liver lobe during a 10 seconds breath hold. Participants were asked to fast at least 3 hours prior to the exam. The details of VCTE assessment have been previously described in references (30, 40). The threshold used for the classification of cirrhosis (stage 4) was VCTE > 11.8 kPa as previously determined in reference (30). Among the first-degree relatives of proband with NAFLD-cirrhosis, 11 did not have an MRE assessment due to contraindication and the presence of advanced fibrosis was determined using a VCTE threshold> 11.8 kPa as previously determined in reference (30).

Liver biopsy was not used for hepatic fat content and fibrosis assessment of controls and first-degree relatives as they were asymptomatic with no suspected liver disease and therefore performing a liver biopsy would have been unethical. A non-invasive, accurate quantitative imaging method was used to estimate liver fat and fibrosis. We have previously shown that MRI-PDFF is a highly precise, accurate, and reproducible non-invasive biomarker for the quantification of liver fat content (41, 42). In addition, MRE is the most accurate, currently available, non-invasive quantitative biomarker of liver fibrosis (30, 43). MRE has been shown to be have excellent diagnostic accuracy in differentiating between normal liver and mild fibrosis (stage 0–2) and between non-advanced fibrosis and advanced fibrosis (stage 3–4) (44, 45).

Definition of NAFLD

Participants were considered to have NAFLD if they had hepatic steatosis (MRI-PDFF ≥5%) and no secondary causes of hepatic steatosis due to factors including the use of steatogenic medications, other liver diseases, and significant alcohol intake (see Exclusion Criteria above for details).

Definition of cirrhosis and advanced fibrosis

Participants were considered to have NAFLD-related cirrhosis if they had NAFLD according to the definition above, and have biopsy proven cirrhosis (histology fibrosis stage 4). We have previously validated that a liver stiffness cut point of >3.63 kPa on MRE provides an accuracy of 0.92 for the detection of advanced fibrosis, and it is the most accurate non-invasive test for the diagnosis of advanced fibrosis (46–48). Advanced fibrosis among first-degree relatives was determined by either imaging evidence of nodularity and presence of intraabdominal varices or other evidence imaging evidence of portal hypertension or liver stiffness assessment with MRE threshold ≥ 3.63 kPa or if MRE were not performed using transient elastography assessment with VCTE threshold ≥ 11.8 kPa.

Microbiome composition by 16S rRNA gene amplicon analysis

DNA extraction and 16S rRNA amplicon sequencing were done using Earth Microbiome Project (EMP) standard protocols (http://www.earthmicrobiome.org/protocols-and-standards/16s) and previously described in references (49, 50). In brief, DNA was extracted using the Qiagen MagAttract PowerSoil DNA kit as previously described) (51). 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 QIIME 1.9.1 script split_libraries_fastq.py, with a Phred quality threshold of 3 and default parameters to generate per-study FASTA sequence files (52).

Statistical analysis

Development of a model utilizing stool derived 16S gut-microbiome profiles to predict NAFLD-cirrhosis. To build a model capable of distinguishing samples belonging to NAFLD-cirrhosis from those of non-NAFLD-controls, we developed a custom machine learning process that employed Random Forest (RF) analysis (53). The set of input features for model building consisted of 16S sequences and patient metadata features. Features from stool microbiome data consisted of the number of 16S sequences (~5700 features) and the patient metadata consisted of age, gender and BMI. The first step in building an RF model consisted of training RF and then selecting features with the most important score > 0.005 (27 features) in a second step. The final random forest model included the 27 bacterial features and important patient metadata (age, sex, and BMI) for a total of 30 predictive features.

Patients' demographic, anthropometric, clinical, and biochemical characteristics were summarized. Categorical variables were shown as counts and percentages, and associations were tested using a chi-squared test or Fisher's exact test. Normally distributed continuous variables were shown as mean (± standard deviation), and differences between groups were analyzed using a two-independent sample t- test or Wilcoxon-Mann-Whitney test. Statistical analysis of cohort

characteristics were performed using SPSS 25.0 (IBM, Chicago, IL). A two-sided p-value <0.05 was considered statistically significant.

Sample size estimation: Based upon our previous study including 16 individuals with NASH-cirrhosis/advanced fibrosis and 33 controls, we could identify significant differences compared to 33 controls (5). The patient data and species abundance had an AUROC of 0.88. Therefore, the study including 26 participants with NAFLD-cirrhosis and 72 controls would be adequate to detect clinically meaningful differences between the sub-groups with a power of at least 80% with a two-tailed p-value of less than 0.01.

2.4 Acknowledgments

RL is supported in part by the American Gastroenterological Association (AGA) Foundation – Sucampo – ASP Designated Research Award in Geriatric Gastroenterology and by a T. Franklin Williams Scholarship Award; Funding provided by: Atlantic Philanthropies, Inc, the John A. Hartford Foundation, OM, the Association of Specialty Professors, and the American Gastroenterological Association and grant K23-DK090303. The project described and RL was partially supported by the National Institutes of Health, Grant UL1TR001442 of CTSA funding beginning August 13, 2015 and beyond. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The gut-microbiome sequencing was funded by Janssen.

2.5 Author contributions

Chapter 2, in full, is a reprint of previously published material: Caussy, C., Tripathi, A., Humphrey, G., Bassirian, S., Singh, S., Faulkner, C., Bettencourt, R., Rizo, E., Richards, L., Xu, Z. Z., Downes, M. R., Evans, R. M., Brenner, D. A., Sirlin, C. B., Knight, R., & Loomba, R.

(2019). A gut microbiome signature for cirrhosis due to nonalcoholic fatty liver disease. *Nature Communications*, *10*(1), 1406.

Cyrielle Caussy: study concept and design, analysis and interpretation of data, drafting of the manuscript, critical revision of the manuscript, approved final submission. Anupriya Tripathi: data analysis and figure generation, interpretation of data, drafting of the manuscript, critical revision of the manuscript, approved final submission. Gregory Humphrey: microbiome sequencing data generation, approved final submission Shirin Bassirian: patient visits, data collection, critical revision of the manuscript, approved final submission. Seema Singh: patient visits, data collection, critical revision of the manuscript, approved final submission. Claire Faulkner: patient visits, data collection, critical revision of the manuscript, approved final submission. Emily Rizo: patient visits, data collection, critical revision of the manuscript, approved final submission. Lisa Richards: patient visits, critical revision of the manuscript, approved final submission. Michael R. Downes: analysis and interpretation of data, drafting of the manuscript, critical revision of the manuscript, approved final submission. Ronald M. Evans: analysis and interpretation of data, drafting of the manuscript, critical revision of the manuscript, approved final submission. David A. Brenner: study concept and design, analysis and interpretation of data, drafting of the manuscript, critical revision of the manuscript, obtained funding, study supervision, approved final submission. Claude B. Sirlin: study concept and design, analysis and interpretation of data, drafting of the manuscript, critical revision of the manuscript, obtained funding, study supervision, approved final submission. Zhenjiang Zech Xu: data interpretation, critical revision of the manuscript, approved final submission. Rob Knight: directed microbiome sequencing and data analysis, critical revision of the manuscript, approved final submission. Rohit Loomba: study concept and design, analysis and interpretation of data, drafting

of the manuscript, critical revision of the manuscript, obtained funding, study supervision, approved final submission. The co-authors listed above supervised or provided support for the research and have given permission for the inclusion of the work in this dissertation.

2.6 Competing interests

The authors declare no competing interests.

2.7 Data and code availability

The 16S sequence data will be deposited at the European Bioinformatics Institute (EMBL-EBI). The analysis was done using QIIME and in-house scripts. All analyses are documented in Jupyter notebooks available at https://github.com/knightlab-analyses/familial-cirrhosis.

2.8 Supplemental figures

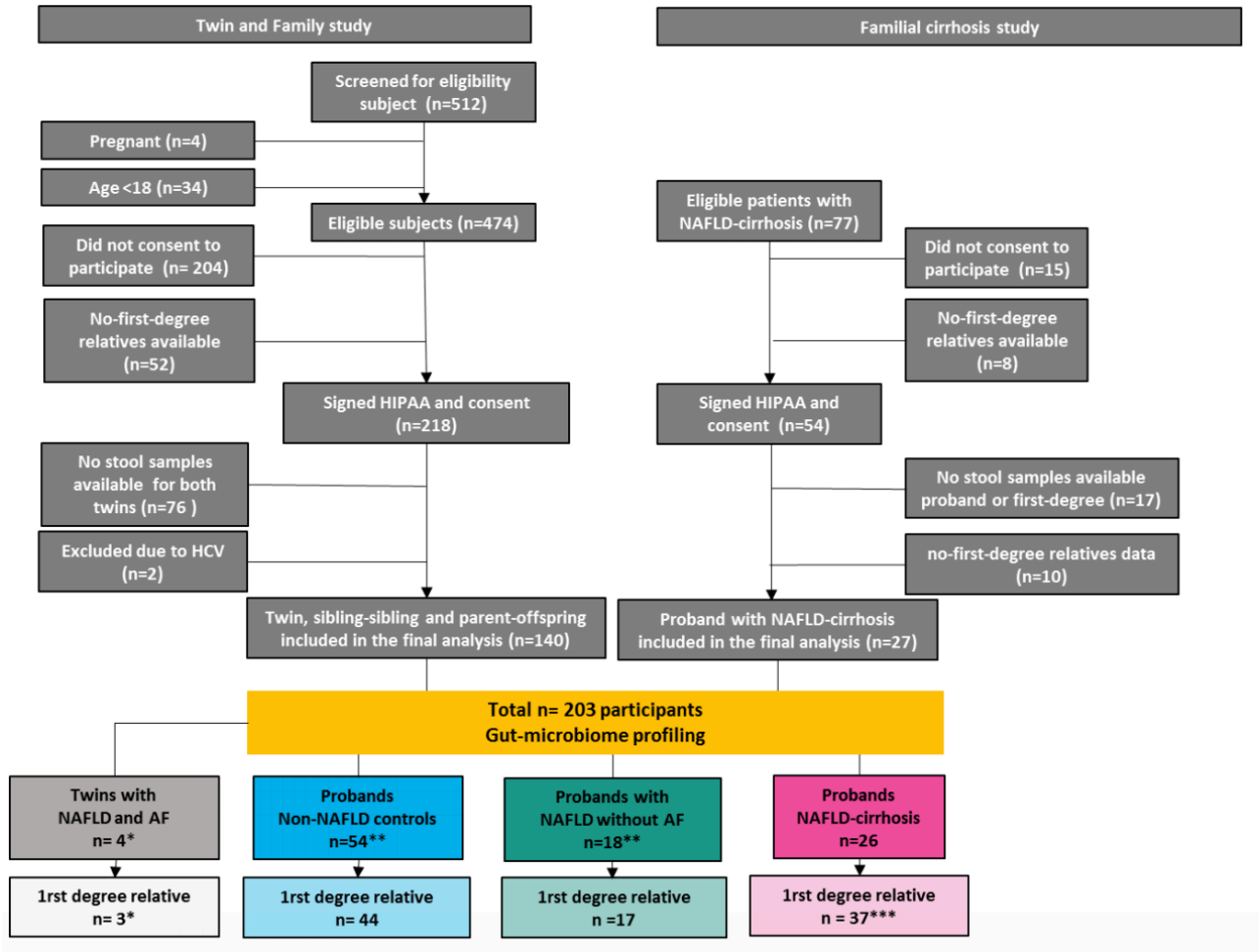

Figure 2.S1 *Study flow-chart*. A total of 203 participants from the Twin and Family study (n=140) and Familial cirrhosis study (n=63) with 16S gut-microbiome profiling were included in the study. *3 twin pairs were concordant for advanced fibrosis and 1 twin had NAFLD- cirrhosis were not assigned in a control group but was included in familial correlation analyses. ** stool samples were not available for the first-degree relative of 10 non-NAFLD controls and 1 proband with NAFLD without advanced fibrosis, the single probands were included in the gut-microbiome signature analysis. ***2 first degree relative did not have liver stiffness assessment (MRE of VCTE but were included in the familial correlation analysis), 11 first-degree relatives did not had an MRE assessment due to CI and were assessed using VCTE using a threshold of 11.8 kPa for the detection of cirrhosis.

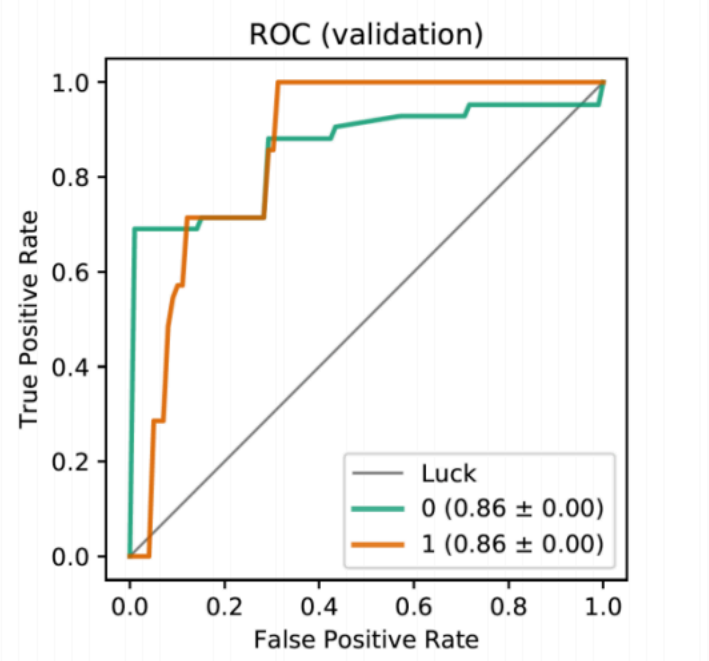

Figure 2.S2 *Sensitivity analyses of the diagnostic accuracy of the gut-microbiome signature for the detection of advanced fibrosis*. Receiver operating characteristic (ROC) curves evaluating ability to predict advanced Fibrosis using Random Forest classification. This curve represents the sensitivity and specificity to distinguish subjects with advanced fibrosis (1) from those without advanced fibrosis (0). The predictive model was trained on probands with NAFLD-cirrhosis (n=24) and non-NAFLD controls (n=47). We validated the prediction on a cohort comprising of NAFLD patients without advanced fibrosis (n=17) and first first-degree relatives of NAFLD-cirrhosis probands (n=32). The model predicted the presence of advanced fibrosis with an accuracy of 86%.

2.9 References

1. Estes C, Razavi H, Loomba R, Younossi Z, Sanyal AJ. 2018. Modeling the epidemic of nonalcoholic fatty liver disease demonstrates an exponential increase in burden of disease. Hepatology 67:123–133.

2. Loomba R, Sanyal AJ. 2013. The global NAFLD epidemic. Nat Rev Gastroenterol Hepatol 10:686–690.

3. Dulai PS, Singh S, Patel J, Soni M, Prokop LJ, Younossi Z, Sebastiani G, Ekstedt M, Hagstrom H, Nasr P, Stal P, Wong VW-S, Kechagias S, Hultcrantz R, Loomba R. 2017. Increased risk of mortality by fibrosis stage in nonalcoholic fatty liver disease: Systematic review and meta-analysis. Hepatology 65:1557–1565.

4. Friedman SL, Neuschwander-Tetri BA, Rinella M, Sanyal AJ. 2018. Mechanisms of NAFLD development and therapeutic strategies. Nat Med 24:908–922.

5. Loomba R, Seguritan V, Li W, Long T, Klitgord N, Bhatt A, Dulai PS, Caussy C, Bettencourt R, Highlander SK, Jones MB, Sirlin CB, Schnabl B, Brinkac L, Schork N, Chen C-H, Brenner DA, Biggs W, Yooseph S, Craig Venter J, Nelson KE. 2017. Gut Microbiome-Based Metagenomic Signature for Non-invasive Detection of Advanced Fibrosis in Human Nonalcoholic Fatty Liver Disease. Cell Metabolism.

6. Hoyles L, Fernández-Real J-M, Federici M, Serino M, Abbott J, Charpentier J, Heymes C, Luque JL, Anthony E, Barton RH, Chilloux J, Myridakis A, Martinez-Gili L, Moreno-Navarrete JM, Benhamed F, Azalbert V, Blasco-Baque V, Puig J, Xifra G, Ricart W, Tomlinson C, Woodbridge M, Cardellini M, Davato F, Cardolini I, Porzio O, Gentileschi P, Lopez F, Foufelle F, Butcher SA, Holmes E, Nicholson JK, Postic C, Burcelin R, Dumas M-E. 2018. Molecular phenomics and metagenomics of hepatic steatosis in non-diabetic obese women. Nat Med 24:1070–1080.

7. Tripathi A, Debelius J, Brenner DA, Karin M, Loomba R, Schnabl B, Knight R. 2018. The gut-liver axis and the intersection with the microbiome. Nat Rev Gastroenterol Hepatol 15:397–411.

8. Boursier J, Mueller O, Barret M, Machado M, Fizanne L, Araujo-Perez F, Guy CD, Seed PC, Rawls JF, David LA, Hunault G, Oberti F, Calès P, Diehl AM. 2016. The severity of nonalcoholic fatty liver disease is associated with gut dysbiosis and shift in the metabolic function of the gut microbiota. Hepatology 63:764–775.

9. Ponziani FR, Bhoori S, Castelli C, Putignani L, Rivoltini L, Del Chierico F, Sanguinetti M, Morelli D, Paroni Sterbini F, Petito V, Reddel S, Calvani R, Camisaschi C, Picca A, Tuccitto A, Gasbarrini A, Pompili M, Mazzaferro V. 2019. Hepatocellular Carcinoma Is Associated With Gut Microbiota Profile and Inflammation in Nonalcoholic Fatty Liver Disease. Hepatology 69:107–120.

10. Chen Y-M, Liu Y, Zhou R-F, Chen X-L, Wang C, Tan X-Y, Wang L-J, Zheng R-D, Zhang H-W, Ling W-H, Zhu H-L. 2016. Associations of gut-flora-dependent metabolite trimethylamine-N-oxide, betaine and choline with non-alcoholic fatty liver disease in adults. Sci Rep 6:19076.

11. Ren Z, Li A, Jiang J, Zhou L, Yu Z, Lu H, Xie H, Chen X, Shao L, Zhang R, Xu S, Zhang H, Cui G, Chen X, Sun R, Wen H, Lerut JP, Kan Q, Li L, Zheng S. 2019. Gut microbiome analysis as a tool towards targeted non-invasive biomarkers for early hepatocellular carcinoma. Gut 68:1014–1023.

12. Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, Guo J, Le Chatelier E, Yao J, Wu L, Zhou J, Ni S, Liu L, Pons N, Batto JM, Kennedy SP, Leonard P, Yuan C, Ding W, Chen Y, Hu X, Zheng B, Qian G, Xu W, Ehrlich SD, Zheng S, Li L. 2014. Alterations of the human gut microbiome in liver cirrhosis. Nature 513:59–64.

13. Caussy C, Soni M, Cui J, Bettencourt R, Schork N, Chen C-H, Ikhwan MA, Bassirian S, Cepin S, Gonzalez MP, Mendler M, Kono Y, Vodkin I, Mekeel K, Haldorson J, Hemming

A, Andrews B, Salotti J, Richards L, Brenner DA, Sirlin CB, Loomba R, Familial NAFLD Cirrhosis Research Consortium. 2017. Nonalcoholic fatty liver disease with cirrhosis increases familial risk for advanced fibrosis. J Clin Invest 127:2697–2704.

14. Struben VM, Hespenheide EE, Caldwell SH. 2000. Nonalcoholic steatohepatitis and cryptogenic cirrhosis within kindreds. Am J Med 108:9–13.

15. Loomba R, Abraham M, Unalp A, Wilson L, Lavine J, Doo E, Bass NM, Nonalcoholic Steatohepatitis Clinical Research Network. 2012. Association between diabetes, family history of diabetes, and risk of nonalcoholic steatohepatitis and fibrosis. Hepatology 56:943–951.

16. Schwimmer JB, Celedon MA, Lavine JE, Salem R, Campbell N, Schork NJ, Shiehmorteza M, Yokoo T, Chavez A, Middleton MS, Sirlin CB. 2009. Heritability of Nonalcoholic Fatty Liver Disease. Gastroenterology.

17. Abdelmalek MF, Liu C, Shuster J, Nelson DR, Asal NR. 2006. Familial Aggregation of Insulin Resistance in First-Degree Relatives of Patients With Nonalcoholic Fatty Liver Disease. Clinical Gastroenterology and Hepatology.

18. Willner IR, Waters B, Patil SR, Reuben A, Morelli J, Riely CA. 2001. Ninety patients with nonalcoholic steatohepatitis: insulin resistance, familial tendency, and severity of disease. Am J Gastroenterol 96:2957–2961.

19. Cui J, Chen C-H, Lo M-T, Schork N, Bettencourt R, Gonzalez MP, Bhatt A, Hooker J, Shaffer K, Nelson KE, Long MT, Brenner DA, Sirlin CB, Loomba R, For The Genetics Of Nafld In Twins Consortium. 2016. Shared genetic effects between hepatic steatosis and fibrosis: A prospective twin study. Hepatology 64:1547–1558.

20. Loomba R, Schork N, Chen C-H, Bettencourt R, Bhatt A, Ang B, Nguyen P, Hernandez C, Richards L, Salotti J, Lin S, Seki E, Nelson KE, Sirlin CB, Brenner D. 2015. Heritability of Hepatic Fibrosis and Steatosis Based on a Prospective Twin Study. Gastroenterology.

21. Chalasani N, Guo X, Loomba R, Goodarzi MO, Haritunians T, Kwon S, Cui J, Taylor KD, Wilson L, Cummings OW, Chen Y-DI, Rotter JI, Nonalcoholic Steatohepatitis Clinical Research Network. 2010. Genome-wide association study identifies variants associated with histologic features of nonalcoholic Fatty liver disease. Gastroenterology 139:1567–76, 1576.e1–6.

22. Speliotes EK, Yerges-Armstrong LM, Wu J, Hernaez R, Kim LJ, Palmer CD, Gudnason V, Eiriksdottir G, Garcia ME, Launer LJ, Nalls MA, Clark JM, Mitchell BD, Shuldiner AR, Butler JL, Tomas M, Hoffmann U, Hwang S-J, Massaro JM, O'Donnell CJ, Sahani DV, Salomaa V, Schadt EE, Schwartz SM, Siscovick DS, NASH CRN, GIANT Consortium, MAGIC Investigators, Voight BF, Carr JJ, Feitosa MF, Harris TB, Fox CS, Smith AV, Kao WHL, Hirschhorn JN, Borecki IB, GOLD Consortium. 2011. Genome-

wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits. PLoS Genet 7:e1001324.

23. Dongiovanni P, Stender S, Pietrelli A, Mancina RM, Cespiati A, Petta S, Pelusi S, Pingitore P, Badiali S, Maggioni M, Mannisto V, Grimaudo S, Pipitone RM, Pihlajamaki J, Craxi A, Taube M, Carlsson LMS, Fargion S, Romeo S, Kozlitina J, Valenti L. 2018. Causal relationship of hepatic fat with liver damage and insulin resistance in nonalcoholic fatty liver. J Intern Med 283:356–370.

24. Romeo S, Kozlitina J, Xing C, Pertsemlidis A, Cox D, Pennacchio LA, Boerwinkle E, Cohen JC, Hobbs HH. 2008. Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. Nat Genet 40:1461–1465.

25. Goodrich JK, Davenport ER, Beaumont M, Jackson MA, Knight R, Ober C, Spector TD, Bell JT, Clark AG, Ley RE. 2016. Genetic Determinants of the Gut Microbiome in UK Twins. Cell Host Microbe 19:731–743.

26. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI. 2009. A core gut microbiome in obese and lean twins. Nature.

27. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, Heath AC, Warner B, Reeder J, Kuczynski J, Gregory Caporaso J, Lozupone CA, Lauber C, Clemente JC, Knights D, Knight R, Gordon JI. 2012. Human gut microbiome viewed across age and geography. Nature.

28. Caussy C, Reeder SB, Sirlin CB, Loomba R. 2018. Noninvasive, Quantitative Assessment of Liver Fat by MRI-PDFF as an Endpoint in NASH Trials. Hepatology.

29. Loomba R, Wolfson T, Ang B, Hooker J, Behling C, Peterson M, Valasek M, Lin G, Brenner D, Gamst A, Ehman R, Sirlin C. 2014. Magnetic resonance elastography predicts advanced fibrosis in patients with nonalcoholic fatty liver disease: a prospective study. Hepatology 60:1920–1928.

30. Hsu C, Caussy C, Imajo K, Chen J, Singh S, Kaulback K, Le M-D, Hooker J, Tu X, Bettencourt R, Yin M, Sirlin CB, Ehman RL, Nakajima A, Loomba R. 2019. Magnetic Resonance vs Transient Elastography Analysis of Patients With Nonalcoholic Fatty Liver Disease: A Systematic Review and Pooled Analysis of Individual Participants. Clin Gastroenterol Hepatol 17:630–637.e8.

31. Ajmera V, Park CC, Caussy C, Singh S, Hernandez C, Bettencourt R, Hooker J, Sy E, Behling C, Xu R, Middleton MS, Valasek MA, Faulkner C, Rizo E, Richards L, Sirlin CB, Loomba R. 2018. Magnetic Resonance Imaging Proton Density Fat Fraction Associates With Progression of Fibrosis in Patients With Nonalcoholic Fatty Liver Disease. Gastroenterology 155:307–310.e2.

32. Caussy C, Alquiraish MH, Nguyen P, Hernandez C, Cepin S, Fortney LE, Ajmera V, Bettencourt R, Collier S, Hooker J, Sy E, Rizo E, Richards L, Sirlin CB, Loomba R. 2018. Optimal threshold of controlled attenuation parameter with MRI-PDFF as the gold standard for the detection of hepatic steatosis. Hepatology 67:1348–1359.

33. Collingridge DS. 2013. A Primer on Quantitized Data Analysis and Permutation Testing. Journal of Mixed Methods Research.

34. Jiang L, Amir A, Morton JT, Heller R, Arias-Castro E, Knight R. 2017. Discrete False-Discovery Rate Improves Identification of Differentially Abundant Microbes. mSystems 2.

35. Caussy C, Hsu C, Lo M-T, Liu A, Bettencourt R, Ajmera VH, Bassirian S, Hooker J, Sy E, Richards L, Schork N, Schnabl B, Brenner DA, Sirlin CB, Chen C-H, Loomba R, Genetics of NAFLD in Twins Consortium. 2018. Link between gut-microbiome derived metabolite and shared gene-effects with hepatic steatosis and fibrosis in NAFLD. Hepatology 68:918–932.

36. Zarrinpar A, Gupta S, Maurya MR, Subramaniam S, Loomba R. 2016. Serum microRNAs explain discordance of non-alcoholic fatty liver disease in monozygotic and dizygotic twins: a prospective study. Gut 65:1546–1554.

37. Chalasani N, Younossi Z, Lavine JE, Diehl AM, Brunt EM, Cusi K, Charlton M, Sanyal AJ. 2012. The diagnosis and management of non-alcoholic fatty liver disease: Practice Guideline by the American Association for the Study of Liver Diseases, American College of Gastroenterology, and the American Gastroenterological Association. Hepatology.

38. Permutt Z, Le T-A, Peterson MR, Seki E, Brenner DA, Sirlin C, Loomba R. 2012. Correlation between liver histology and novel magnetic resonance imaging in adult patients with non-alcoholic fatty liver disease - MRI accurately quantifies hepatic steatosis in NAFLD. Aliment Pharmacol Ther 36:22–29.

39. Patel NS, Peterson MR, Brenner DA, Heba E, Sirlin C, Loomba R. 2013. Association between novel MRI-estimated pancreatic fat and liver histology-determined steatosis and fibrosis in non-alcoholic fatty liver disease. Aliment Pharmacol Ther 37:630–639.

40. Caussy C, Chen J, Alquiraish MH, Cepin S, Nguyen P, Hernandez C, Yin M, Bettencourt R, Cachay ER, Jayakumar S, Fortney L, Hooker J, Sy E, Valasek MA, Rizo E, Richards L, Brenner DA, Sirlin CB, Ehman RL, Loomba R. 2018. Association Between Obesity and Discordance in Fibrosis Stage Determination by Magnetic Resonance vs Transient Elastography in Patients With Nonalcoholic Liver Disease. Clin Gastroenterol Hepatol 16:1974–1982.e7.

41. Noureddin M, Lam J, Peterson MR, Middleton M, Hamilton G, Le T-A, Bettencourt R, Changchien C, Brenner DA, Sirlin C, Loomba R. 2013. Utility of magnetic resonance

imaging versus histology for quantifying changes in liver fat in nonalcoholic fatty liver disease trials. Hepatology 58:1930–1940.

42. Reeder SB. 2013. Emerging quantitative magnetic resonance imaging biomarkers of hepatic steatosis. Hepatology.

43. Cui J, Ang B, Haufe W, Hernandez C, Verna EC, Sirlin CB, Loomba R. 2015. Comparative diagnostic accuracy of magnetic resonance elastography vs. eight clinical prediction rules for non-invasive diagnosis of advanced fibrosis in biopsy-proven non-alcoholic fatty liver disease: a prospective study. Aliment Pharmacol Ther 41:1271–1280.

44. Kim D, Ray Kim W, Talwalkar JA, Kim HJ, Ehman RL. 2013. Advanced Fibrosis in Nonalcoholic Fatty Liver Disease: Noninvasive Assessment with MR Elastography. Radiology.

45. Yin M, Glaser KJ, Talwalkar JA, Chen J, Manduca A, Ehman RL. 2016. Hepatic MR Elastography: Clinical Performance in a Series of 1377 Consecutive Examinations. Radiology.

46. Dulai PS, Sirlin CB, Loomba R. 2016. MRI and MRE for non-invasive quantitative assessment of hepatic steatosis and fibrosis in NAFLD and NASH: Clinical trials to clinical practice. J Hepatol 65:1006–1016.

47. Park CC, Nguyen P, Hernandez C, Bettencourt R, Ramirez K, Fortney L, Hooker J, Sy E, Savides MT, Alquiraish MH, Valasek MA, Rizo E, Richards L, Brenner D, Sirlin CB, Loomba R. 2017. Magnetic Resonance Elastography vs Transient Elastography in Detection of Fibrosis and Noninvasive Measurement of Steatosis in Patients With Biopsy-Proven Nonalcoholic Fatty Liver Disease. Gastroenterology 152:598–607.e2.

48. Cui J, Heba E, Hernandez C, Haufe W, Hooker J, Andre MP, Valasek MA, Aryafar H, Sirlin CB, Loomba R. 2016. Magnetic resonance elastography is superior to acoustic radiation force impulse for the Diagnosis of fibrosis in patients with biopsy-proven nonalcoholic fatty liver disease: A prospective study. Hepatology.

49. Shalapour S, Lin X-J, Bastian IN, Brain J, Burt AD, Aksenov AA, Vrbanac AF, Li W, Perkins A, Matsutani T, Zhong Z, Dhar D, Navas-Molina JA, Xu J, Loomba R, Downes M, Yu RT, Evans RM, Dorrestein PC, Knight R, Benner C, Anstee QM, Karin M. 2017. Inflammation-induced IgA+ cells dismantle anti-liver cancer immunity. Nature 551:340–345.

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

51. Marotz C, Amir A, Humphrey G, Gaffney J, Gogul G, Knight R. 2017. DNA extraction for streamlined metagenomics of diverse environmental samples. Biotechniques 62:290–293.

52. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, Gonzalez A, Kosciolek T, McCall L-I, McDonald D, Melnik AV, Morton JT, Navas J, Quinn RA, Sanders JG, Swafford AD, Thompson LR, Tripathi A, Xu ZZ, Zaneveld JR, Zhu Q, Caporaso JG, Dorrestein PC. 2018. Best practices for analysing microbiomes. Nat Rev Microbiol 16:410–422.

53. Breiman L. 2001. Machine Learning.

Chapter 3. Intermittent hypoxia and hypercapnia, a hallmark of obstructive sleep apnea, alters the gut microbiome and metabolome

Obstructive sleep apnea (OSA) is a common disorder characterized by episodic obstruction to breathing due to upper airway collapse during sleep. OSA has been associated with adverse cardiovascular and metabolic outcomes, although data regarding potential causal pathways are still evolving. As O_2 and CO_2 affect the ecology of the gut microbiota and the microbiota has been shown to contribute to various cardio-metabolic disorders, we hypothesized that OSA alters the gut ecosystem which exacerbates the downstream physiological consequences. Here, we model human OSA and its cardiovascular consequences using *Ldlr*$^{-/-}$ mice fed a high-fat diet and exposed to intermittent hypoxia and hypercapnia (IHH). The gut microbiome and metabolome were characterized longitudinally (using 16S rRNA amplicon sequencing and untargeted LC-MS/MS mass-spectrometry) and seen to co-vary during IHH. Joint analysis of microbiome and metabolome data revealed marked compositional changes in both microbial (>10%, most remarkably, Clostridia) and molecular (>22%) species in the gut. Moreover, molecules altered in abundance included microbe-dependent bile acids, enterolignans and fatty acids, highlighting the impact of IHH on host-commensal co-metabolism in the gut. Thus, we present the first evidence that IHH perturbs the gut microbiome functionally, setting the stage for understanding its involvement in associated cardio-metabolic disorders.

3.1 Introduction

Intestinal dysbiosis marks various cardiovascular diseases comorbid with OSA. It has not been systematically studied if dysbiosis due to hypoxic stress in OSA is causally linked to these comorbidities. We take advantage of a longitudinal study design and paired '-omics to investigate correlations in microbial and molecular dynamics in the gut to ascertain the contribution of microbes on intestinal metabolism. We observe microbe-dependent changes in the gut metabolome that will guide future research on unrecognized mechanistic links between gut microbes and comorbidities of OSA. Additionally, we highlight novel, non-invasive biomarkers for OSA-linked pathologies.

Obstructive sleep apnea (OSA) afflicts nearly 12% of the adult population in the USA with a cost burden of nearly $149.6 billion, according to a recent study commissioned by the American Academy of Sleep Medicine (1). Timely diagnosis and treatment of OSA improves not only sleep and cognitive function but also management of comorbid cardiometabolic diseases (CMDs). Therefore, identifying downstream consequences of OSA would aid in development of effective treatment modalities, reducing overall health care utilization.

OSA is marked by changes in oxygen and carbon dioxide-inspired concentrations which impacts the gut microbial community (2). Since the gut microbiota play a key role in metabolism of dietary precursors including lipids, cholesterol and choline, it impacts the cardiometabolic health of the host (3). Gut dysbiosis has already been linked to an array of metabolic disorders such as hypertension, T2 diabetes, hepatic steatosis and atherosclerosis (4, 5). Additionally, previous work has identified specific gut bacteria to be significantly correlated with plasma cholesterol and apolipoprotein levels (6). Thus, probing this commensal ecosystem may provide a valuable avenue of investigation to understand the mechanism of pathogenesis of cardiovascular consequences of

OSA. In this study, we investigated the taxonomic and molecular alterations in gut microbiome that potentially mediate the interplay between OSA and related CMDs.

3.2 Results

We used atherosclerosis-prone (*Ldlr*-) adult mice fed high-fat diet (HFD) enriched in cholesterol and milk fat (resembling western dietary practices) to evaluate atherosclerosis risk in OSA. We previously demonstrated that IHH increases atherosclerosis plaque formation in this model (7). As episodic hypoxia and hypercapnia mimic the changes in blood gases that occur in OSA-driven downstream consequences (8), these mice were exposed to IHH (treatment group; n=8) or air (control group; n=8) and examined longitudinally for 6 weeks (Methods, Figure 3.S1). Fecal samples, representative of the gut ecosystem, were collected at baseline and twice each week thereafter, and the microbiome and metabolome were profiled using 16S rRNA amplicon sequencing and LC-MS/MS-based untargeted mass-spectrometry, respectively. These data were processed (Methods) to obtain relative abundances of microbial and molecular species per sample (referred to as feature tables henceforth), which were used for comparing OSA-mimicking and control mice.

First, we performed principal coordinate analysis (PCoA) on the microbiome and metabolome feature tables to identify major factors driving the clustering of samples. Figure 3.1 shows the PCoA plotted against time to visualize the dynamics of clustering based on gut microbiome (unweighted UniFrac distances (9); Figure 3.1a and metabolome (Gower distances (9, 10)); Figure 3.1b,c as duration of IHH-exposure increases. Here, the first fecal sample represents the baseline gut composition before animals were switched to a HFD. There is a rapid shift in both microbial and molecular composition due to HFD alone, consistent with similar previous findings

(11–13). Moreover, starting from a highly congruent gut composition, IHH-exposed mice significantly diverge from controls with increasing exposure duration (PERMANOVA test performed per time point, Table S1[3]). This demonstrates that prolonged IHH-exposure (analogous to chronic OSA) cumulatively perturbs the gut microbiome and metabolome. We tested the relationship between the two omics datasets by superimposing the principal coordinates computed from microbiome and metabolome data (Procrustes analysis (14)**;** Figure 3.1d, e). The ordination spaces are correlated (Mantel test r-statistic = 0.36, $p < 0.001$), and changes in metabolome and microbiome of samples within the treatment groups over time are proportional, suggesting microbe-dependent changes in intestinal metabolism on chronic OSA.

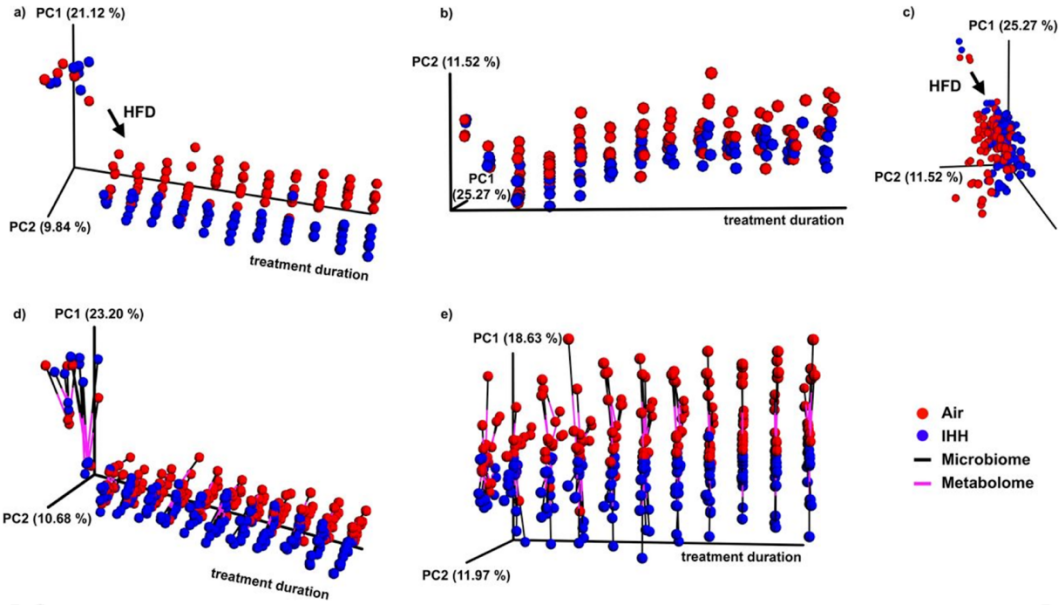

Figure 3.1 *Principal coordinate analysis (PCoA) and Procrustes analysis of gut microbiome and metabolome*. **a)** PCoA of microbiome (16S rRNA sequencing) data using unweighted UniFrac distances **b, c)** PCoA of metabolome (untargeted LC-MS/MS mass-spectrometry) data using Gower distances **d, e)** Procrustes analysis of microbiome and metabolome datasets **d)** with baseline samples **e)** without baseline samples. Here, coordinates for a sample obtained using microbiome data (black line) are connected to coordinates for the same sample obtained using metabolome data (pink line). This analysis stretches, rotates and superimposes ordinations generated from one dataset over the other, while preserving distances within each individual matrix. The goal is to find the best fit between two matrices to infer whether one dataset coherently captures the properties of the other. (IHH: intermittent hypoxia and hypercapnia)

[3] https://msystems.asm.org/content/3/3/e00020-18

We then tested for specific microbes and metabolites that changed with OSA. More than 80 (of ~730) microbial features differed significantly between the IHH-exposed group and controls (by permutation test with discrete FDR correction (15)). Figure 3.S2a presents a global overview of these changes in gut microbiota per sample (sorted by duration of treatment). Table S2[3] provides a list of these differentially represented bacteria that potentially contribute to alterations in gut metabolism due to IHH. Figure 3.2a-f displays trends in relative abundances of bacteria showing the largest differences, which belong to the Mogibacteriaceae (family), *Oscillospira* (genus), Lachnospiraceae (family) and Clostridiaceae (family). Previous studies have consistently associated these taxonomic groups with metabolic and inflammatory disturbances in the host (16, 17), which suggests that related mechanisms may be at play in driving the consequences of hypoxic and hypercapnic stress.

[3] https://msystems.asm.org/content/3/3/e00020-18

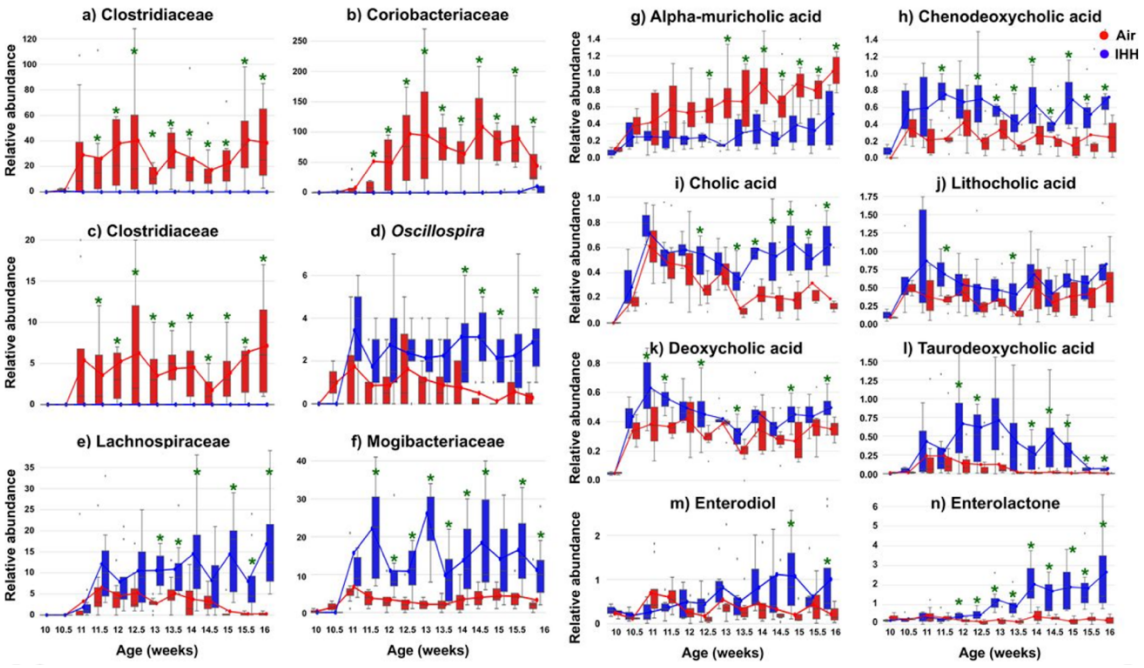

Figure 3.2 *Changes in the gut microbes and molecules due to IHH exposure*. **a-f)** Top differentially abundant sOTUs elevated in the control group (a,b,c) and treatment group (d,e,f). The sOTUs belonging to the families Clostridiaceae (a,c) and Coriobacteriaceae (b) were elevated in controls, whereas those belonging to the genus *Oscillospira* (d) and families Lachnospiraceae (e), Mogibacteriaceae (f) were higher in IHH-exposed mice. **g-n)** Trends in abundance of significantly differential bile acids. These differential bile acids include unconjugated primary bile acids: alpha-muricholic acid (g), chenodeoxycholic acid (h) and cholic acid (i), secondary bile acids: lithocholic acid (j) deoxycholic acid (k) and, conjugated secondary bile acid: taurodeoxycholic acid (l). **m,n)** Trends in abundance of significantly differential xenoestrogens, enterodiol (m) and enterolactone (n). (IHH: intermittent hypoxia and hypercapnia)

Using the same statistical approach, we found that more than 380 (out of ~1700) molecular species differed significantly in relative abundance in animals exposed to IHH. Figure 3.S2b provides a global representation of these differentially abundant molecules in samples belonging to treatment and control groups and sorted by treatment duration (Table S3[3] provides a comprehensive list of these molecules). To gain insight into the structures of these differentially abundant metabolites, we performed molecular networking using Global Natural Products Social Molecular Networking (GNPS) (18). The molecular network is constructed using a cosine similarity measure between tandem mass spectral data, then visualized using Cytoscape (19)

[3] https://msystems.asm.org/content/3/3/e00020-18

(Figure 3.S3). Each node in the network, which represents a consensus MS/MS spectrum, was searched against public libraries in GNPS. In total, we annotated about 400 molecular compounds in GNPS including bile acids, fatty acids and phytoestrogens. Additionally, all key compounds discussed in this work were defined to the highest level of annotation according to the metabolomics standards initiative using commercial standards (Figure 3.S4,5, Table S4[3]) (20). Interestingly, the top differentially abundant features detected between IHH and control mice, included molecules known to depend on gut microbes for their production. Below, we discuss some of these metabolites and their implications with respect to consequences of OSA.

Alterations in bile acids: We observed significant alterations in bile acids (BAs) between IHH-exposed mice and control groups (Table S3[3]). Figure 3.2g-n displays these trends in primary (Figure 3.2g-i) and secondary (Figure 3.2j-l) bile acids with increasing IHH exposure duration. Primary BAs are amphipathic molecules synthesized in the liver from cholesterol. These are conjugated to glycine or taurine and released in the biliary tract. Together with other biliary components, these facilitate in emulsification and transportation of dietary fats, cholesterol, and fat-soluble vitamins. About 95% of the BAs are reabsorbed in the terminal ileum and recycled. The remaining 5% reach the colon and are deconjugated, dehydrogenated and dehydroxylated by the intestinal bacteria to form secondary BAs (21). BAs, including microbially-generated BAs, are potent signaling molecules that interact with the Farnesoid X receptor (FXR) (expressed in the liver and intestine) which modulates BA synthesis by the liver (22). Perturbations in gut microbial populations disrupt normal signaling properties that regulate BA production, and can profoundly alter the BA composition in the gut. A range of diseases, including cardiometabolic diseases, are

[3] https://msystems.asm.org/content/3/3/e00020-18

characterized by aberrant BA profiles (23), and prolonged perturbations in the BA pool could also be a factor in mediating the consequences of OSA.

Elevations in phytoestrogens: The dietary hormones, enterolactone (mammalian lignan) and enterodiol (oxidation product of enterolactone) were significantly elevated in the exposed mice compared to controls. Figure 3.2m, n shows the trends in their abundances with increasing duration of IHH-exposure. These molecules are phytoestrogens i.e. plant-derived hormones that structurally mimic estrogen and are produced by intestinal microbiota on bioconversion of dietary lignans. Owing to their affinity to estrogen receptors (producing estrogenic or/and antiestrogenic effects (24)), they perturb many hormone-dependent systems in the body and have been linked to adverse metabolic, reproductive and neurological outcomes (25). Sex-specific differences in OSA diagnostic symptoms and risk factors suggest hormonal involvement (26, 27). However, the contribution of microbes in maintaining hormonal homeostasis has not yet been investigated. Therefore, these findings motivate novel avenues of research for biomarkers and therapeutic targets to manage the metabolic consequences of OSA.

Alterations in fatty-acids: In addition to changes in bile acids and phytoestrogens, we also detected differentially abundant fatty acid-related chemical families (Table S3[3]). For example, we noted a significant reduction in molecular features matched to elaidic acid. Elaidic acid is an unsaturated fatty acid that increases plasma cholesteryl ester transfer protein (CETP) activity that modulates systemic levels of LDL and HDL cholesterol. A decrease in elaidic acid in the IHH-exposed group suggests reduction in plasma CETP activity, a mechanism associated with adverse cardiovascular effects (28). Similarly, phytomonic, jasmonic, hexadecanoic, linoleic acid, and conjugated linoleic acids were also reduced in exposed mice compared to controls. Out of these,

[3] https://msystems.asm.org/content/3/3/e00020-18

phytomonic acid and conjugated linoleic acid are known to be microbially produced (29, 30), suggesting that changes in the microbiome could be contributing to these changes in metabolome.

In summary, we demonstrate that IHH, a hallmark of OSA, changes the microbiota and the chemistry in the gut. We have highlighted the changes of bile acids, phytoestrogens and fatty acids under OSA-related conditions leading to CMDs. The present results reveal a previously unrecognized mechanistic link between OSA and gut microbes. It suggests that targeting gut microbiota and their metabolites may serve as a potential therapeutic approach for the treatment of cardiometabolic consequences of OSA patients.

3.3 Materials and methods

Animals

Atherosclerosis-prone ten-week old male *Ldlr*- mice on C57BL/6J background (Stock Numbers 002207; The Jackson Laboratory, Bar Harbor, ME) were used in this study (31). *Ldlr* deficiency was confirmed by PCR according to the vendor's instructions. All animal protocols were approved by the Animal Care Committee of the University of California San Diego and followed the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health.

High Fat Diet Treatment

Starting at 10 weeks of age, male mice were provided with a high fat diet (HFD) containing 1.25% cholesterol and 21% milk fat (4.5 Kcal/g; TD96121; Harlan-Teklad Madison, WI) for 6 weeks while being exposed to either IHH or room air.

Intermittent Hypoxia and Hypercapnia Exposure

Intermittent hypoxia and hypercapnia (IHH) was maintained in a computer-controlled atmosphere chamber system (OxyCycler, Reming Bioinstruments, Redfield, NY) as previously described (32). IHH exposure was introduced to the mice in short periods (~4 min) of synchronized reduction of O_2 (from 21% to 8%) and elevation of CO_2 (from ~0.5% to 8%) separated by alternating periods (~4 min) of normoxia ($[O_2]$ = 21%) and normocapnia ($[CO_2]$ = ~0.5%) with 1-2 min ramp intervals for 10 hours per day during the light cycle, for 6 weeks. This treatment protocol mimics the severe clinical condition observed in obstructive sleep apnea patients. Mice on the same HFD but in room air were used as controls.

As the experimental setup requires IHH-exposed mice in a controlled atmosphere chamber and controls in room air, we ensured that the effect of treatment is not confounded by the effect of distinct housing conditions. To do so, we used two cages per treatment group, and we compared the relative effect size of treatment and cages with redundancy analysis (RDA), which estimates the independent effect size of each covariate on microbiome composition variation based on unweighted UniFrac Distance (33). The RDA results showed that treatment had a higher effect size than the cages, more specifically, that treatment contributed to 11.6% of the microbiome community variation, while cages had an independent effect size of around 9.8%; with respect to the metabolome, treatment contributed to 6.2% of the variation, while cages only about 0.7%

LC-MS/MS data acquisition

Prior to LC-MS/MS analysis, fecal samples were prepared using the following extraction procedures. For extraction, 500 μL of 50/50 methanol/H_2O was added to all fecal samples and vortexed. Fecal pellets in extraction solvent were placed in an ultrasonic bath and sonicated for 30 minutes to break apart the pellet, then allowed to incubate for an additional 30

minutes. Extracted samples were the centrifuged to separate insoluble material and 450 uL of each liquid extract was subsequently transferred to a 96-deep-well plate and dried completely using centrifugal evaporation (Centrivap, Labconco, Kansas City, MO). The dried extracts were resuspended in 150 μL of methanol/H2O (1/1, v/v) including 1 μM amitriptyline as an autosampler injection standard. After resuspension, the samples were transferred into 96-well plates and analyzed on a Vanquish ultra-high performance liquid chromatography (UPLC) system coupled to a Q-Exactive orbital ion trap (Thermo Fisher Scientific, Bremen, Germany). For the chromatographic separation, a C18 core-shell column (Kinetex, 50 x 2 mm, 1.7 um particle size, 100 A pore size, Phenomenex, Torrance, USA) with a flowrate of 0.5 mL/min (Solvent A: H2O + 0.1 % formic acid (FA), Solvent B: Acetonitrile (ACN) + 0.1 % FA) was used. After injection, the samples were eluted during a linear gradient from 0-0.5 min, 5 % B, 0.5-4 min 5-50 % B, 4-5 min 50-99 % B, followed by a 2 min washout phase at 99% B and a 2 min re-equilibration phase at 5 % B. For online MS/MS measurements, the flow was directed to heated ESI source (HESI). The electrospray ionization (ESI) parameters were set to 35 L/min sheath gas flow, 10 L/min auxiliary gas flow, 2 L/min sweep gas flow and 400 °C auxiliary gas temperature. The spray voltage was set to 3.5 kV and the inlet capillary was set to 250 °C. 50 V S-lens radio frequency (RF) level was applied. Product ion spectra were recorded in data dependent acquisition (DDA) mode. Both MS1 survey scans (m/z 150-1500) and up to 5 MS/MS scans of the most abundant ions per duty cycle were measured with a resolution (R) of 17,500 with 1 micro-scan in positive mode. The maximum ion injection time was set to 100 ms. MS/MS precursor selection windows were set to *m/z* 3 with *m/z* 0.5 offset. Normalized collision energy was stepwise increased from 20 to 30 to 40 % with z = 2 as default charge state. MS/MS experiments were automatically triggered at the apex of a peak within 2 to 15 s from their first occurrence. Dynamic exclusion was set to 5 s.

LC-MS/MS data analysis

Feature detection: Thermo raw datasets were converted to mzXML in centroid mode using MSConvert (part of proteowizard) (34, 35). All mzXML files were cropped with *m/z* range of 75.00 - 1000.00 Da. MS1-based feature detection and MS2-based molecular networking was performed using GNPS workflow (https://gnps.ucsd.edu/) (18). The parameters used are detailed here: http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=c6438af750784d919dcd0ee0a783b4fc. Feature extraction parameters were optimized using MZmine2 (http://mzmine.sourceforge.net/) (36) with a signal threshold of 2.0e5 and 0.3 s minimum peak width. The mass tolerance was set to 10 ppm and the maximum allowed retention time deviation was set to 10 s. For chromatographic deconvolution the local minimum search algorithm was used with minimum relative peak height of 1% and minimum retention time range of 0.6 s. Maximum peak width was set to 1 min. After isotope peak removal, the peak lists of all samples were aligned with the above-mentioned retention time and mass tolerances. After the creation of a feature matrix containing the feature retention times, exact mass and peak areas of the corresponding extracted ion chromatograms, metadata of the samples (treatment type and duration) was added. The signal intensity of the features was normalized (probabilistic quotient normalization or PQN)(37) to internal standard (mz 278.189; RT 3.81 minutes) for subsequent analysis.

MS/MS annotations: Molecular features, in the form of MS/MS spectra, were putatively identified using MS2-based spectral library matches. False Discovery Rate (FDR) was estimated using a decoy database approach (38) in GNPS and was found to be less than 1 % above a cosine similarity score of 0.6. (GNPS job link: https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=feac48de4c9f45d485403e3feb7a470d)

Therefore, we use a cosine score of 0.65 here. For level 1 annotation, as defined by the 2007

metabolomics standards initiative, differentially abundant metabolites we purchased authentic standards of alpha-muricholic acid, chenodeoxycholic acid, cholic acid, lithocholic acid, deoxycholic acid, taurodeoxycholic acid and xenoestrogens, enterodiol and enterolactone from Cayman chemical, MI, USA and analyzed using identical LC-MS/MS method described above. We then compared and verified the matching of exact masses, fragmentation patterns and retention times of those compounds to ensure correct annotations (Figure 3.S2, 3).

Statistical analysis: QIIME 1.9.1 was used to perform principal coordinate analysis (PCoA) (*beta_diversity.py;* Gower dissimilarity metric (39)) and PERMANOVA test (*compare_categories.py*). The PCoA plots were visualized in EMPeror (40). Differential abundance analysis was performed using discrete FDR (15).

16S rRNA sequencing

DNA extraction and 16S rRNA amplicon sequencing was done using EMP standard protocols (http://www.earthmicrobiome.org/protocols-and-standards/16s) (41). In brief, DNA was extracted using the MO BIO PowerSoil DNA extraction kit (Carlsbad, CA). Amplicon PCR was performed on the V4 region of the 16S rRNA gene using the primer pair 515f–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 MoBio UltraClean PCR Clean-up kit and sequenced on the Illumina HiSeq 2500 sequencing platform. Sequence data were demultiplexed and minimally quality filtered using the QIIME 1.9.1 script *split_libraries_fastq.py* with Phred quality threshold of 3 and default parameters to generate per-study FASTA sequence files.

16S marker gene data analysis

Feature detection and identification: The raw sequence data were processed using Deblur workflow (42) with default parameters in Qiita (https://qiita.ucsd.edu/). This generated a sub-Operational Taxonomic Units (sOTU) abundance per sample BIOM table (42, 43). Taxonomies for sOTUs were assigned using a sklearn-based taxonomy classifier (feature-classifier plugin) in QIIME 2 (44). The sOTU table was rarefied to a depth of 2000 sequences/sample to control for sequencing effort (45). A phylogeny was inferred using SATe-enabled Phylogenetic Placement (46) which was used to insert 16S Deblur sOTUs into the Greengenes 13_8 at 99% phylogeny.

Statistical analysis: QIIME 2 was used to perform principal coordinate analysis (PCoA) (unweighted UniFrac distances (47)). QIIME 1.9.1 was used for PERMANOVA test (*compare_categories.py*), Mantel test (*compare_distance_matrices.py*) and Procrustes analysis (*transform_coordinate_matrices.py*). The PCoA and Procrustes plots were visualized in EMPeror. (40) Differential abundance analysis was performed using discrete FDR (15).

3.4 Acknowledgements

We acknowledge NIH Grants GMS10RR029121 and 5P41GM103484-07 for the shared instrumentation infrastructure that enabled this work.

3.5 Author contributions

Chapter 3, in full, is a reprint of previously published material: Tripathi, A., Melnik, A. V., Xue, J., Poulsen, O., Meehan, M. J., Humphrey, G., Jiang, L., Ackermann, G., McDonald, D., Zhou, D., Knight, R., Dorrestein, P. C., & Haddad, G. G. (2018). Intermittent Hypoxia and

Hypercapnia, a Hallmark of Obstructive Sleep Apnea, Alters the Gut Microbiome and Metabolome. In *mSystems* (Vol. 3, Issue 3).

A.T. wrote the initial manuscript, managed, analyzed and interpreted the data. A.V.M. contributed to the manuscript, analyzed and interpreted the data. J.X. contributed to the manuscript and the study design. O.P. contributed to the study design and collected samples for analysis. M.J.M and G.H. performed mass-spectrometry and amplicon sequencing, respectively. L.J. analyzed and interpreted the data. G.A. curated the metadata. D.M. contributed to the manuscript and interpreted the data. D.Z. contributed to the study design. R.K., P.C.D., G.G.H. conceived and designed the study, interpreted the data and contributed to the manuscript. The co-authors listed above supervised or provided support for the research and have given permission for the inclusion of the work in this dissertation.

3.6 Competing interests

The authors declare no competing financial interests.

3.7 Data and code availability

The data generated in this study are available publicly. Metabolomics data: MSV000081482. Commercial standards: MSV000081853. Microbiome data: EBI accession: ERP106495. Data analysis has been documented in Jupyter notebooks available on Github (https://github.com/knightlab-analyses/haddad_osa)

3.8 Supplemental figures

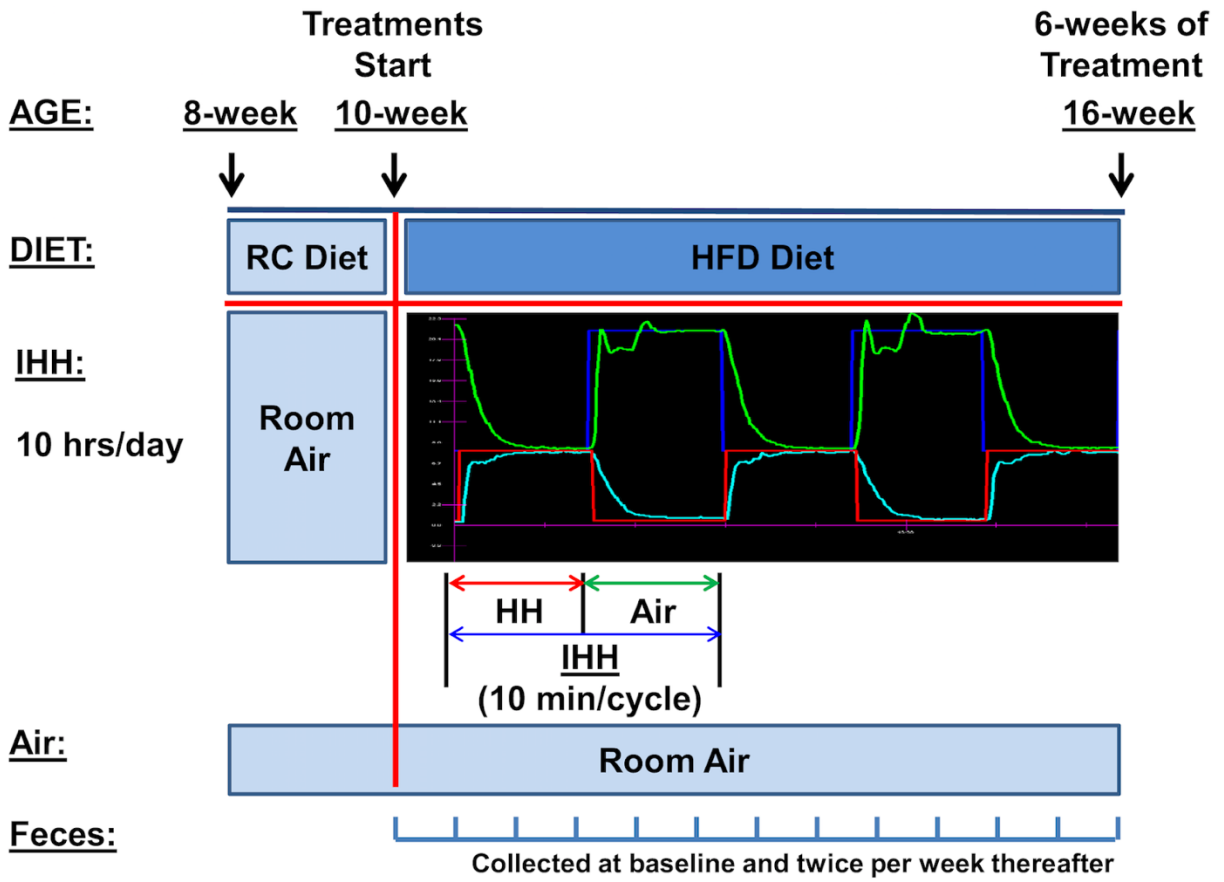

Figure 3.S1 *Schematic illustration of treatment paradigm and sample collection*. Groups of 8-week old male *Ldlr*- mice were transferred to the treatment room for 2 weeks of acclimatization with room air (RA) and regular chow (RC) food. At 10 weeks of age, mice were switched to high fat diet (HFD) and treated with or without intermittent hypoxia and hypercapnia (IHH). The IHH treatment group received 10 hrs/day IHH in the light cycle for 6 weeks (The blue line was the O_2 set point and the green was the actual level of O_2. The red line was the CO_2 set point and light blue was the actual level of CO_2). The control groups remained in room air for the same period. Fecal pellets were collected at baseline and twice per week thereafter and were used for microbiome and metabolome analyses. (IHH: intermittent hypoxia and hypercapnia)

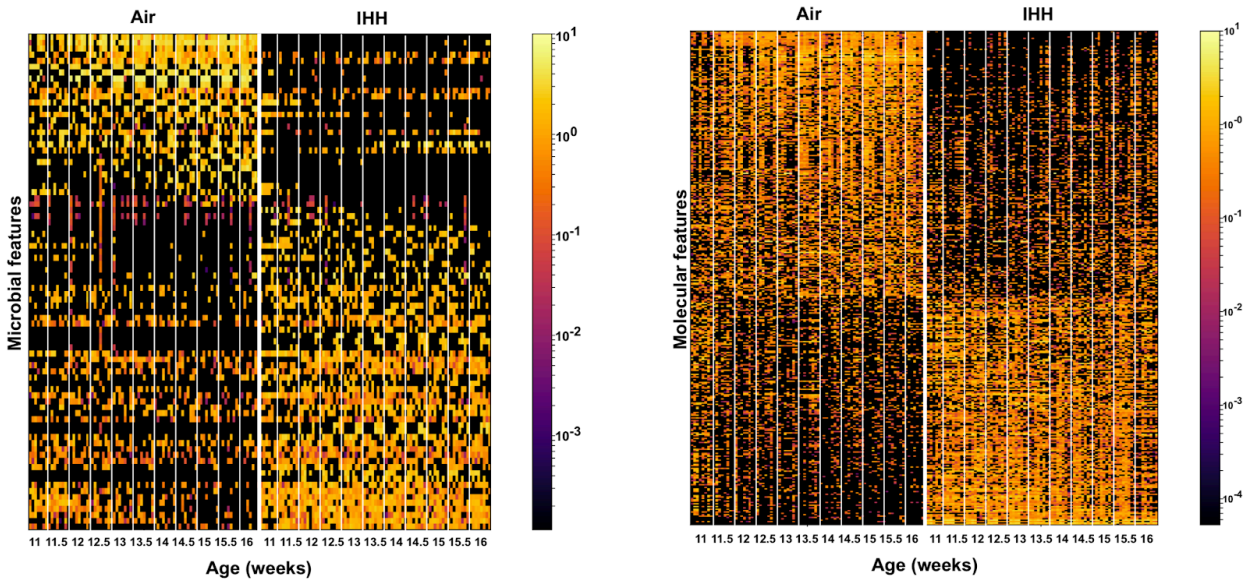

Figure 3.S2 *Global overview of changes in gut microbiota and metabolome*. **a)** Heatmap of 87 differential microbial sOTUs between IHH-exposed and control mice. **b)** Heatmap of 382 differential molecular features between IHH-exposed and control mice. (IHH: intermittent hypoxia and hypercapnia)

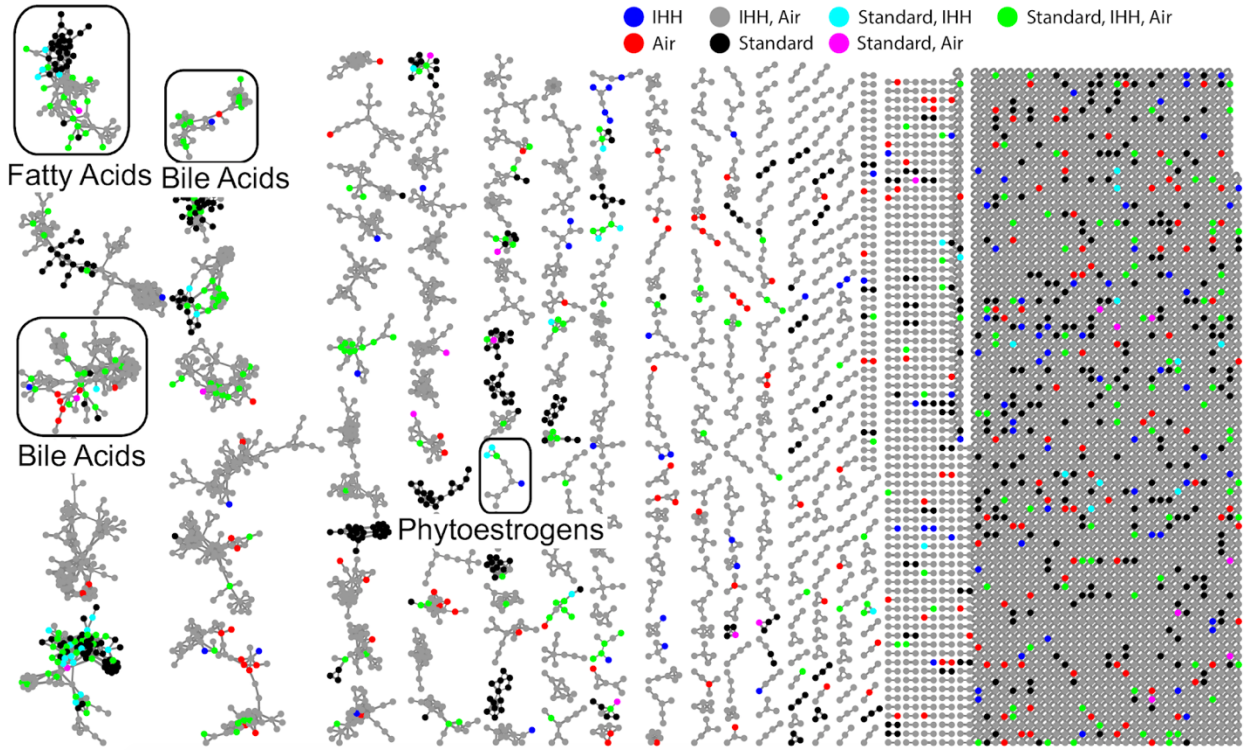

Figure 3.S3 *Molecular network of LC-MS/MS metabolomic data generated on GNPS (rendered using Cytoscape 3.4* (19)**).** Highlighted by boxes are clusters in which differentially abundant metabolites of interest are observed. Color-coding represents the treatment group and its overlap with authentic standards. (IHH: intermittent hypoxia and hypercapnia)

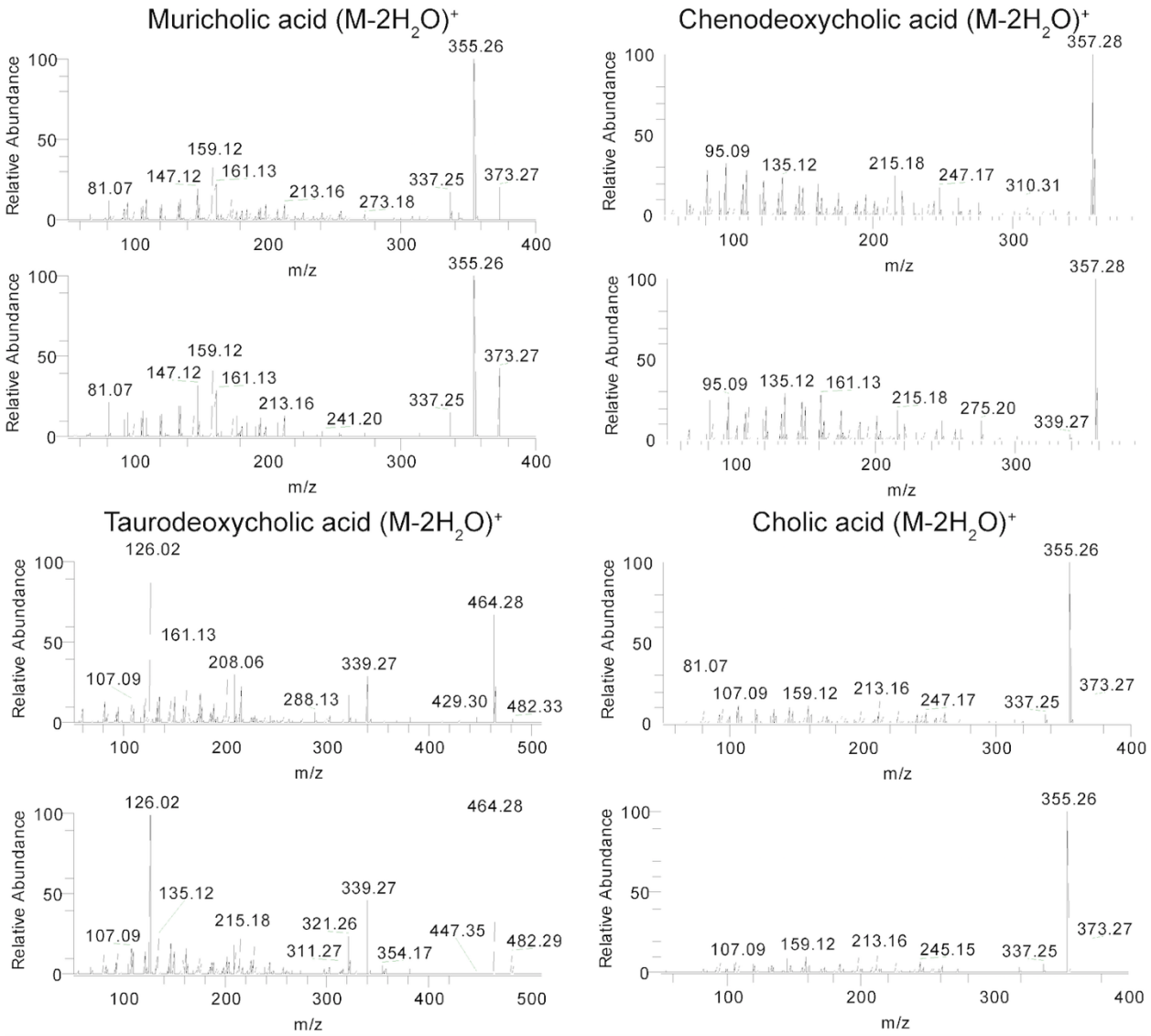

Figure 3.S4. *Comparisons of MS/MS fragmentation spectra I.* MS/MS fragmentation spectra displayed for annotated molecules. Fragmentation spectra originating from the most abundant ion is picked for each molecule to display (refer to table S4[3]). MS/MS spectrum observed in samples and commercial standards shown on the top and on the bottom for each compound, respectively.

[3] https://msystems.asm.org/content/3/3/e00020-18

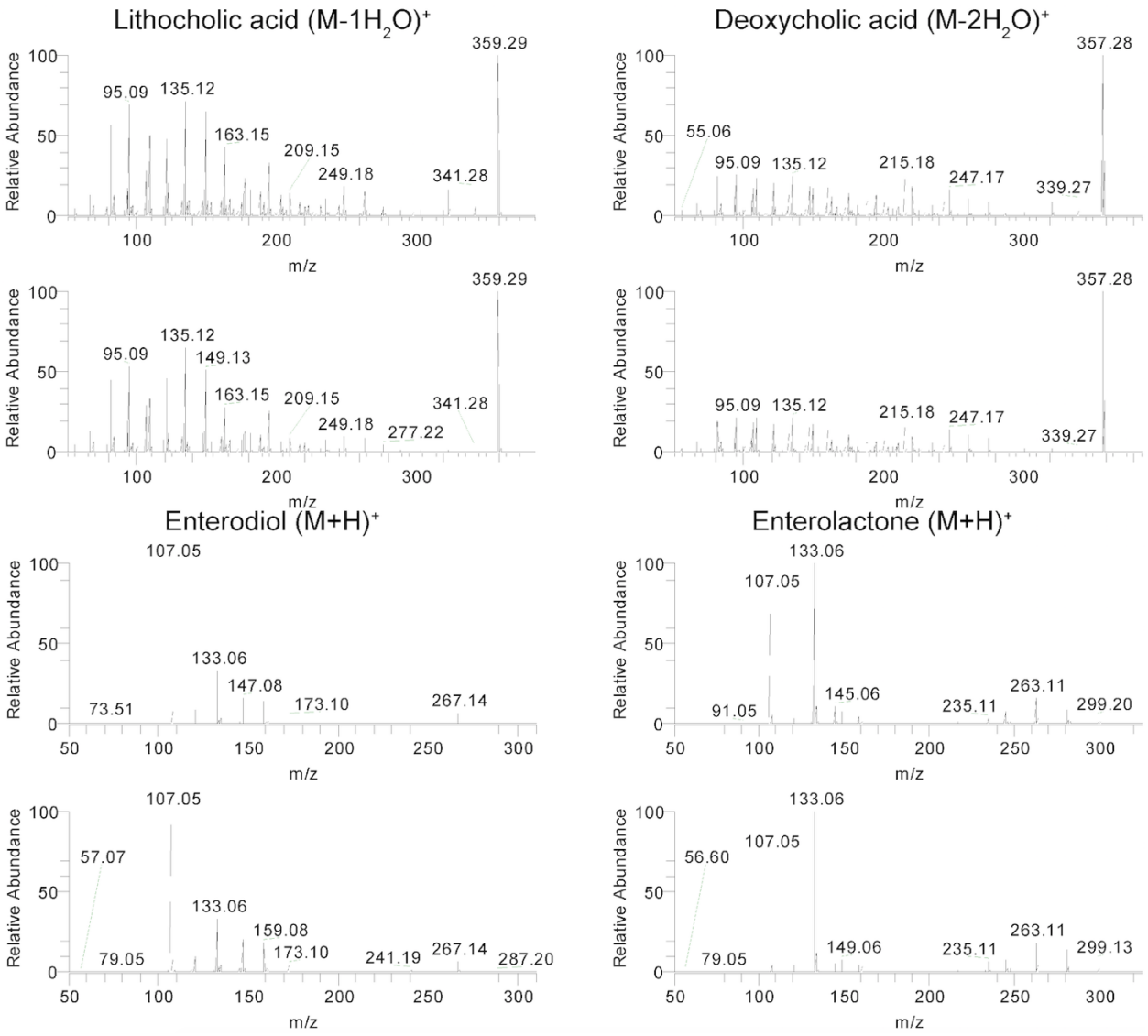

Figure 3.S5. *Comparisons of MS/MS fragmentation spectra II*. MS/MS fragmentation spectra displayed for annotated molecules. Fragmentation spectra originating from the most abundant ion is picked for each molecule to display (refer to the table S4[3]). MS/MS spectrum observed in samples and commercial standards shown on the top and on the bottom for each compound, respectively.

[3] https://msystems.asm.org/content/3/3/e00020-18

3.9 References

1. Watson NF. 2016. Health Care Savings: The Economic Value of Diagnostic and Therapeutic Care for Obstructive Sleep Apnea. J Clin Sleep Med 12:1075–1077.

2. Moreno-Indias I, Torres M, Montserrat JM, Sanchez-Alcoholado L, Cardona F, Tinahones FJ, Gozal D, Poroyko VA, Navajas D, Queipo-Ortuño MI, Farré R. 2014. Intermittent hypoxia alters gut microbiota diversity in a mouse model of sleep apnoea. Eur Respir J 45:1055–1065.

3. Garcia-Rios A, Torres-Peña JD, Perez-Jimenez F, Perez-Martinez P. 2017. Gut Microbiota: A New Marker of Cardiovascular Disease. Curr Pharm Des 23:3233–3238.

4. He M, Shi B. 2017. Gut microbiota as a potential target of metabolic syndrome: the role of probiotics and prebiotics. Cell Biosci 7:54.

5. Okubo H, Nakatsu Y, Kushiyama A, Yamamotoya T, Matsunaga Y, Inoue M-K, Fujishiro M, Sakoda H, Ohno H, Yoneda M, Ono H, Asano T. 2017. Gut microbiota as a therapeutic target for metabolic disorders. Curr Med Chem.

6. Koren O, Spor A, Felin J, Fåk F, Stombaugh J, Tremaroli V, Behre CJ, Knight R, Fagerberg B, Ley RE, Bäckhed F. 2011. Human oral, gut, and plaque microbiota in patients with atherosclerosis. Proc Natl Acad Sci U S A 108 Suppl 1:4592–4598.

7. Xue J, Zhou D, Poulsen O, Imamura T, Hsiao Y-H, Smith TH, Malhotra A, Dorrestein P, Knight R, Haddad GG. 2017. Intermittent Hypoxia and Hypercapnia Accelerate Atherosclerosis, Partially via Trimethylamine-Oxide. Am J Respir Cell Mol Biol.

8. Sforza E, Roche F. 2016. Chronic intermittent hypoxia and obstructive sleep apnea: an experimental and clinical approach. Hypoxia (Auckl) 4:99–108.

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

10. Gower JC. 1971. A General Coefficient of Similarity and Some of Its Properties. Biometrics 27:857.

11. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA, Biddinger SB, Dutton RJ, Turnbaugh PJ. 2014. Diet rapidly and reproducibly alters the human gut microbiome. Nature 505:559–563.

12. Esko T, Hirschhorn JN, Feldman HA, Hsu Y-HH, Deik AA, Clish CB, Ebbeling CB, Ludwig DS. 2017. Metabolomic profiles as reliable biomarkers of dietary composition. Am J Clin Nutr 105:547–554.

13. Turnbaugh PJ, Ridaura VK, Faith JJ, Rey FE, Knight R, Gordon JI. 2009. The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. Sci Transl Med 1:6ra14.

14. Gower JC. 1975. Generalized procrustes analysis. Psychometrika 40:33–51.

15. Jiang L, Amir A, Morton JT, Heller R, Arias-Castro E, Knight R. 2017. Discrete False-Discovery Rate Improves Identification of Differentially Abundant Microbes. mSystems 2:e00092–17.

16. Pascal V, Pozuelo M, Borruel N, Casellas F, Campos D, Santiago A, Martinez X, Varela E, Sarrabayrouse G, Machiels K, Vermeire S, Sokol H, Guarner F, Manichanh C. 2017. A microbial signature for Crohn's disease. Gut 66:813–822.

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

18. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T, Porto C, Bouslimani A, Melnik AV, Meehan MJ, Liu W-T, Crüsemann M, Boudreau PD, Esquenazi E, Sandoval-Calderón M, Kersten RD, Pace LA, Quinn RA, Duncan KR, Hsu C-C, Floros DJ, Gavilan RG, Kleigrewe K, Northen T, Dutton RJ, Parrot D, Carlson EE, Aigle B, Michelsen CF, Jelsbak L, Sohlenkamp C, Pevzner P, Edlund A, McLean J, Piel J, Murphy BT, Gerwick L, Liaw C-C, Yang Y-L, Humpf H-U, Maansson M, Keyzers RA, Sims AC, Johnson AR, Sidebottom AM, Sedio BE, Klitgaard A, Larson CB, P CAB, Torres-Mendoza D, Gonzalez DJ, Silva DB, Marques LM, Demarque DP, Pociute E, O'Neill EC, Briand E, Helfrich EJN, Granatosky EA, Glukhov E, Ryffel F, Houson H, Mohimani H, Kharbush JJ, Zeng Y, Vorholt JA, Kurita KL, Charusanti P, McPhail KL, Nielsen KF, Vuong L, Elfeki M, Traxler MF, Engene N, Koyama N, Vining OB, Baric R, Silva RR, Mascuch SJ, Tomasi S, Jenkins S, Macherla V, Hoffman T, Agarwal V, Williams PG, Dai J, Neupane R, Gurr J, Rodríguez AMC, Lamsa A, Zhang C, Dorrestein K, Duggan BM, Almaliti J, Allard P-M, Phapale P, Nothias L-F, Alexandrov T, Litaudon M, Wolfender J-L, Kyle JE, Metz TO, Peryea T, Nguyen D-T, VanLeer D, Shinn P, Jadhav A, Müller R, Waters KM, Shi W, Liu X, Zhang L, Knight R, Jensen PR, Palsson BO, Pogliano K, Linington RG, Gutiérrez M, Lopes NP, Gerwick WH, Moore BS, Dorrestein PC, Bandeira N. 2016. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol 34:828–837.

19. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504.

20. Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, Fan TW-M, Fiehn O, Goodacre R, Griffin JL, Hankemeier T, Hardy N, Harnly J, Higashi R, Kopka J, Lane AN, Lindon JC, Marriott P, Nicholls AW, Reily MD, Thaden JJ, Viant MR. 2007.

Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). Metabolomics 3:211–221.

21. Chiang JYL. 2013. Bile acid metabolism and signaling. Compr Physiol 3:1191–1212.

22. Ramírez-Pérez O, Cruz-Ramón V, Chinchilla-López P, Méndez-Sánchez N. 2017. The Role of the Gut Microbiota in Bile Acid Metabolism. Ann Hepatol 16:15–20.

23. Joyce SA, Gahan CGM. 2017. Disease-Associated Changes in Bile Acid Profiles and Links to Altered Gut Microbiota. Dig Dis 35:169–177.

24. Patisaul HB, Jefferson W. 2010. The pros and cons of phytoestrogens. Front Neuroendocrinol 31:400–419.

25. Gore AC, Chappell VA, Fenton SE, Flaws JA, Nadal A, Prins GS, Toppari J, Zoeller RT. 2015. Executive Summary to EDC-2: The Endocrine Society's Second Scientific Statement on Endocrine-Disrupting Chemicals. Endocr Rev 36:593–602.

26. McKinney J, Ortiz-Young D, Jefferson F. 2015. Gender differences in obstructive sleep apnea and the associated public health burden. Sleep Biol Rhythms 13:196–209.

27. Jehan S, Auguste E, Zizi F, Pandi-Perumal SR, Gupta R, Attarian H, Jean-Louis G, McFarlane SI. 2016. Obstructive Sleep Apnea: Women's Perspective. J Sleep Med Disord 3.

28. Abbey M, Nestel PJ. 1994. Plasma cholesteryl ester transfer protein activity is increased when trans-elaidic acid is substituted for cis-oleic acid in the diet. Atherosclerosis 106:99–107.

29. Hofmann K, Tausig F. 1955. On the identity of phytomonic and lactobacillic acids; a reinvestigation of the fatty acid spectrum of Agrobacterium (Phytomonas) tumefaciens. J Biol Chem 213:425–432.

30. Van Nieuwenhove CP, Teran V, Nelina S. 2012. Conjugated Linoleic and Linolenic Acid Production by Bacteria: Development of Functional FoodsProbiotics.

31. Ishibashi S, Brown MS, Goldstein JL, Gerard RD, Hammer RE, Herz J. 1993. Hypercholesterolemia in low density lipoprotein receptor knockout mice and its reversal by adenovirus-mediated gene delivery. J Clin Invest 92:883–893.

32. Xue J, Zhou D, Poulsen O, Imamura T, Hsiao Y-H, Smith TH, Malhotra A, Dorrestein P, Knight R, Haddad GG. 2017. Intermittent Hypoxia and Hypercapnia Accelerate Atherosclerosis, Partially via Trimethylamine-Oxide. Am J Respir Cell Mol Biol.

33. Falony G, Joossens M, Vieira-Silva S, Wang J, Darzi Y, Faust K, Kurilshikov A, Bonder MJ, Valles-Colomer M, Vandeputte D, Tito RY, Chaffron S, Rymenans L, Verspecht C,

De Sutter L, Lima-Mendez G, D'hoe K, Jonckheere K, Homola D, Garcia R, Tigchelaar EF, Eeckhaudt L, Fu J, Henckaerts L, Zhernakova A, Wijmenga C, Raes J. 2016. Population-level analysis of gut microbiome variation. Science 352:560–564.

34. Adusumilli R, Mallick P. 2017. Data Conversion with ProteoWizard msConvert, p. 339–368. In Methods in Molecular Biology.

35. Kessner D, Chambers M, Burke R, Agus D, Mallick P. 2008. ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics 24:2534–2536.

36. Pluskal T, Castillo S, Villar-Briones A, Oresic M. 2010. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics 11:395.

37. Dieterle F, Ross A, Schlotterbeck G, Senn H. 2006. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. Anal Chem 78:4281–4290.

38. Scheubert K, Hufsky F, Petras D, Wang M, Nothias L-F, Dührkop K, Bandeira N, Dorrestein PC, Böcker S. 2017. Significance estimation for large scale metabolomics annotations by spectral matching. Nat Commun 8:1494.

39. Gower JC. 1971. A General Coefficient of Similarity and Some of Its Properties. Biometrics 27:857.

40. Vázquez-Baeza Y, Pirrung M, Gonzalez A, Knight R. 2013. EMPeror: a tool for visualizing high-throughput microbial community data. Gigascience 2:16.

41. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, Gormley N, Gilbert JA, Smith G, Knight R. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J 6:1621–1624.

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

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

44. Caporaso JG, Gregory Caporaso J, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight

R. 2010. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7:335–336.

45. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, Lozupone C, Zaneveld JR, Vázquez-Baeza Y, Birmingham A, Hyde ER, Knight R. 2017. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome 5:27.

46. Mirarab S, Nguyen N, Warnow T. 2011. SEPP: SATé-Enabled Phylogenetic PlacementBiocomputing 2012.

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

Chapter 4. Intermittent hypoxia and hypercapnia reproducibly change the gut microbiome and metabolome across rodent model systems

Studying perturbations in the gut ecosystem using animal models of disease continues to provide valuable insights into the role of the microbiome in various physiological and pathological conditions. However, understanding whether these changes are consistent across animal models of different genetic backgrounds, and hence potentially translatable to human populations remains a major unmet challenge in the field. Nonetheless, in relatively limited cases have the same interventions been studied in two animal models in the same laboratory. Moreover, such studies typically examine only one data layer and one-time point. Here, we show the power of utilizing time series microbiome (measured by 16S rRNA amplicon profiling) and metabolome (measured by untargeted LC-MS/MS) data to relate two different mouse models of atherosclerosis: *ApoE*$^{-}$ and *Ldlr*$^{-}$ that are exposed to intermittent hypoxia and hypercapnia (IHH) longitudinally (for 10 weeks and 6 weeks, respectively) to model chronic obstructive sleep apnea. Using Random Forest classifiers trained on each data layer, we show excellent accuracy values in predicting IHH-exposure within *ApoE*$^{-}$ and *Ldlr*$^{-}$ knockout models, and in cross-applying predictive features found in one animal model to the other. Some of the key microbes and metabolites that predicted IHH-exposure across animal models included bacterial species from the family Clostridiaceae, muricholic acid (a bile acid) and vaccenic acid (a fatty acid), providing a refined set of biomarkers reproducibly associated with this intervention. The results highlight that time series, multi-omics data can be used to relate different animal models of disease to one another using supervised

machine learning techniques, and can provide a pathway towards identifying robust microbiome and metabolome features that underpin translation from animal models to understanding human disease.

4.1 Introduction

Reproducibility of microbiome research is a major topic of contemporary interest. Although it is often possible to distinguish individuals with specific diseases within a study, the differences are often inconsistent across cohorts, often due to systematic variation in analytical conditions. Here we study the same intervention in two different mouse models of cardiovascular disease (atherosclerosis) by profiling the microbiome and metabolome in stool specimens over time. We demonstrate that shared microbial and metabolic changes are involved in both models with the intervention. We then introduce a pipeline for finding similar results in other studies. This work will help find common features identified across different model systems, which are most likely to apply in humans.

Obstructive sleep apnea (OSA) is a common sleep disorder marked by obstructed breathing due to episodic upper airway collapse. Chronic OSA is associated with adverse cardio-metabolic outcomes such as atherosclerosis (1); however, potential causal pathways remain elusive. We previously modeled human OSA and its cardiovascular consequences in Ldlr knockout (*Ldlr*$^{-/-}$; atherosclerosis model) mice by exposing individuals to intermittent hypoxia and hypercapnia (IHH), a hallmark of OSA (2). IHH is a clinically important exposure because it markedly promotes atherosclerotic lesions in the pulmonary arteries and aorta in not only *Ldlr*$^{-/-}$ mice but also ApoE knockout (*ApoE*$^{-/-}$) mice, another widely-used atherosclerosis model (3, 4), thereby mimicking the adverse cardiovascular changes that occur in OSA patients (5). In *Ldlr*$^{-/-}$ mice, we

reported significant shifts in the bacterial and chemical composition of the gut on IHH-exposure. The key chemical alterations included changes in microbe-dependent metabolites such as gut-derived estrogen-like molecules (phytoestrogens) and bile acids. These observations revealed an unrecognized link between IHH and gut microbes, thereby holding immense potential for translation in OSA patients. However, a key challenge in microbiome research is understanding if different animal models, or animal models and human subjects are characterized by common changes in the microbiome and metabolome (6, 7). As a first step, finding reproducible alterations across multiple animal models would provide confidence in the generalizability of the findings and accrue evidence for clinical relevance.

Here, we use machine learning predictive models to address the reproducibility of the perturbations associated with IHH exposure in the gut ecosystem using both *Ldlr*$^{-}$ and *ApoE*$^{-}$ mouse models (Figure 4.S1). To model OSA and its cardiovascular conditions, all mice were either exposed to IHH (treatment group) or air (control group) and fed a high-fat diet (HFD) (3, 4). Individuals were studied longitudinally for 6 weeks (*Ldlr*$^{-}$) or 10 weeks (*ApoE*$^{-}$) to understand the impact prolonged IHH-exposure (analogous to chronic OSA in humans). Furthermore, multiple cages per treatment group were used to untangle the effect of treatment with the effect of distinct housing conditions (8). Starting with 10 weeks of age (baseline), fecal pellets were collected twice every week, and profiled for microbiome and metabolome using 16S rRNA amplicon sequencing and liquid chromatography-tandem mass spectrometry (LC-MS/MS)-based untargeted mass spectrometry, respectively. These data layers were processed per recommended practices (9) to obtain relative abundances of microbial and molecular species per sample for all downstream analyses (see methods).

Predictive models that classify microbiome and metabolome responses to interventions have proven extremely useful in disease diagnosis and biomarker discovery (10), (11). Yet, these have been surprisingly hard to generalize across populations or model systems (12), (13, 14). In this work, we use Random Forest (RF) classification to investigate the cross-applicability of our previous findings in *Ldlr*$^{-/-}$ mice (2) to *ApoE*$^{-/-}$ mice and vice-versa. RF is an ensemble machine learning algorithm that fits many decision trees on random subsamples of original data, and then aggregates the results of each decision tree to improve the prediction accuracy (15). The level of accuracy is often expressed using the area under the curve (AUC) of true-positive versus false-positive rates, known as a receiver operating characteristic (ROC). RF has consistently been reported to perform well in high-dimensional datasets i.e. datasets with many features (microbial reads or metabolites) such as ours, making it our algorithm of choice for this work (16–18) . We had previously shown that machine learning classifiers trained on Inflammatory Bowel Disease (IBD) cases and healthy controls in humans can distinguish between IBD cases and controls in dogs using cross-sectional microbiome data (19). To our knowledge, however, this type of cross-model classification task has not been performed with metabolomics data, or with data collected longitudinally.

4.2 Results

Unsupervised comparison of the gut microbiome and metabolome in *ApoE*$^{-/-}$ and *Ldlr*$^{-/-}$ mouse models: First, we performed Principal Coordinate Analysis (PCoA) to get a visual overview of the characteristic microbiome and metabolome of the two animal models. PCoA is an unsupervised method routinely used to explore major factors that drive the clustering of data points in high-dimensional datasets by projecting the samples in a reduced-dimensional space (as 2D or

3D graph) (20). Figure 4.1 displays the PCoA results plotted along time to visualize the dynamics of diet and IHH-associated changes in the gut ecosystem. This analysis shows that the *Ldlr*$^{-}$ and *ApoE*$^{-}$ mice in our study have very distinct gut microbial (Figure 4.1a and b) and chemical signature (Figure 4.1c and d) which is captured by the first principal axis (axis 1) in both data layers. These plots also capture a rapid shift in the baseline gut microbial and chemical composition in response to HFD which has also been reported previously (21, 22). We performed PCoA without baseline samples, to better visualize the impact of IHH-exposure alone (Figure 4.S2). We observed that despite underlying differences in the two genotypes, axis 2 consistently captured IHH-induced shifts in both the gut microbiome and metabolome highlighting common shifts in the gut ecosystem due to IHH exposure.

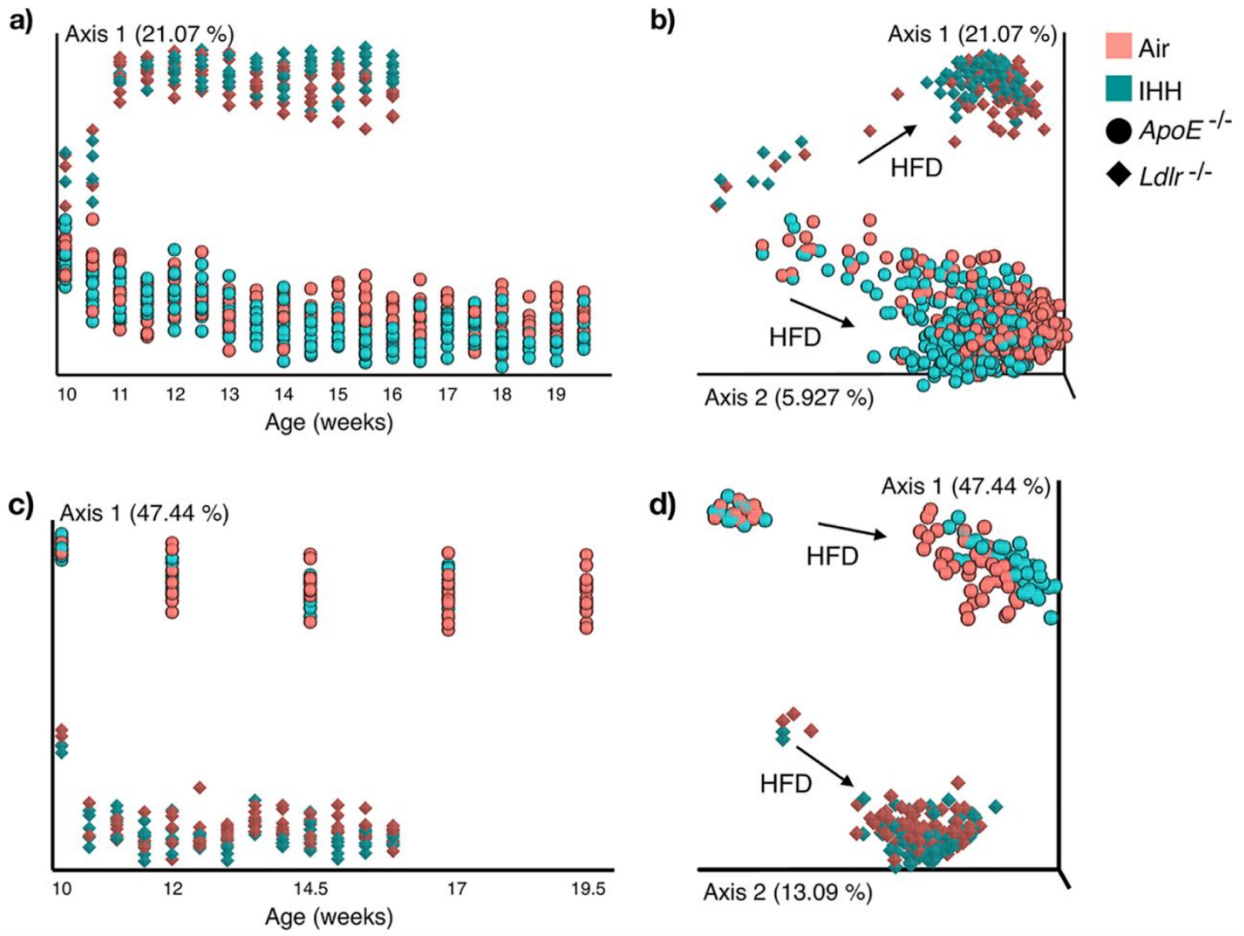

Figure 4.1. *Principal-coordinate analysis (PCoA) of the gut microbiome and metabolome in ApoE and Ldlr mouse models*. **(a and b)** PCoA of the microbiome (16S rRNA sequencing) data using unweighted UniFrac distances. **(c and d)** PCoA of the metabolome (untargeted LC-MS/MS) data using Bray-Curtis distances. The ordination is visualized along the duration of treatment (starting at 10 weeks of age, with an interval of 0.5 week). Axis 1, principal coordinate 1; IHH, intermittent hypoxia and hypercapnia; HFD, high-fat diet.

When comparing the overall sharedness of microbial features, out of 635 unique 16S deblured sequences (23) in *Ldlr*- (and 582 in *ApoE*-) gut microbiome, the two animal models only shared 248 unique sequences. The chemical space was also distinct with the two models sharing only 137 out of 267 and 374 metabolomic features in *Ldlr*- and *ApoE*- mice, respectively (see methods). Interestingly, *Ldlr*- and *ApoE*- mice have more similar microbiomes (PERMANOVA (24) pseudo-F 13.2605, $p<0.001$) at baseline (10 weeks of age) i.e. before the HFD-induced shift is observed, compared to later time points (pseudo-F 19.9059 at 12 weeks of age). We note a similar divergence in the gut metabolome (pseudo-F 46.9112 and 66.1165 at baseline and 12 weeks

of age) of the two animal models as well. Together, these results suggest a differential impact of high-fat diet on the gut ecosystem of the *Ldlr*- and *ApoE*- mice, making the mouse models more distinct over time.

It is important to note that the two animal models are temporally separated for sample collection and data acquisition (for both 16S rRNA sequencing and LC-MS/MS mass-spectrometry), which likely contributes to the strong distinction between the models observed here. We quantified the effects of covariates such as genotypes (or processing batches), age of individuals, housing conditions and individual variability on the microbiome and metabolome composition by performing effect size analysis on our dataset (see 'Effect size analyses' section in methods). While the largest effect on the microbial and chemical composition was linked to the mouse model, the type of exposure (IHH or air) impacted each data layer within both models significantly. Moreover, the effect sizes varied based on the animal model highlighting the distinctive characteristics of the gut ecosystem in the two models (Table S1[4]).

Gut microbiome- and metabolome-based prediction of IHH exposure within and across animal models: Our unsupervised analysis showed that the gut ecosystems of *ApoE*- and *Ldlr*- mice, despite being inherently distinct, consistently shift in response to IHH-exposure. We applied supervised machine learning in order to capture the consistent shifts associated with IHH-exposure in both animal models (Figure 4.S2). Specifically, we built RF classifiers using IHH-associated microbial and chemical composition in *ApoE*- and tested its performance in predicting IHH-exposure in *Ldlr*- and vice-versa. This informed us if the changes we observe in one model are reproducible to the other, which would make the findings more relevant for translation in OSA patients.

[4] https://msystems.asm.org/content/4/2/e00058-19

To examine the predictive potential of microbiome data, we trained RF classifiers on relative abundances of 16S tag sequences shared between the two mouse models (443 features after removing low prevalence sequences [see methods]). Within each animal model, the classifiers yielded nearly perfect prediction (99% area under the receiver operator characteristic curve or AUC) of IHH-exposure (Figure 4.2). We then predicted the same in *Ldlr*$^{-/-}$ using RF trained on microbiome signature in *ApoE*$^{-/-}$ (Figure 4.2a) and vice-versa (Figure 4.2c), still achieving very high cross-model prediction accuracies (95% and 89% AUC, respectively). Similarly, we used metabolomics data for training RF models on relative abundance of MS1 spectral ions (377 features after removing low prevalence ions [see methods]). Metabolome-based RF classifiers also predicted IHH-exposure within animal models accurately (99% AUC), and maintained impressive cross-model prediction accuracies (97% when trained on *Ldlr*$^{-/-}$ and tested on *ApoE*$^{-/-}$ [Figure 4.2b]; 84% vice-versa [Figure 4.2d]). Together, these analyses suggest that IHH-exposure alters both the gut microbial and chemical composition distinguishably in each animal model. Moreover, the changes induced by IHH-exposure are consistent across *Ldlr*$^{-/-}$ and *ApoE*$^{-/-}$ models, despite the underlying differences between the two genotypes (Figure 4.1, Figure 4.S2). It is worth noting that we hugely benefited from our longitudinal sample collection scheme as we had more data points available for learning, despite limited number of animals (i.e. n=8 for *Ldlr*$^{-/-}$ and n=12 for *ApoE*$^{-/-}$) per group. We accounted for the longitudinal samples from the same individual in our analyses by ensuring that observations for each individual appeared either in the training or validation dataset but not both. This prevented over-optimistic cross-validation accuracy scores as a result of the model overfitting to the characteristics of the individual itself rather than the treatment. (The relatively lower accuracy of *ApoE*$^{-/-}$-based metabolomics RF model can be attributed to fewer numbers of samples compared to *Ldlr*$^{-/-}$; Figure 4.S1).

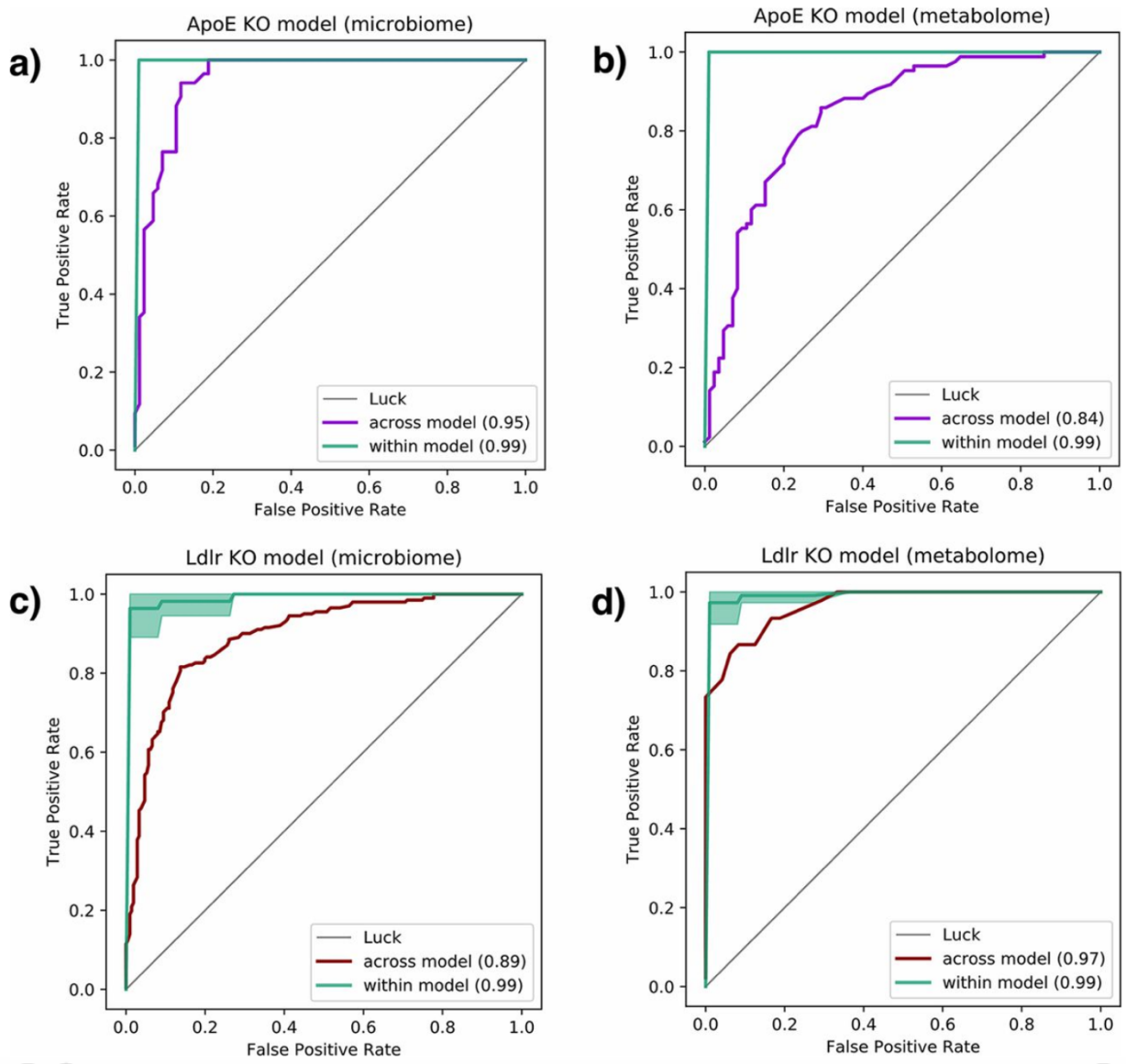

Figure 4.2. *Receiver operating characteristic (ROC) curves evaluating ability to predict exposure to IHH using Random Forest model*. Green curves represent classification accuracy within each mouse model. Purple ROC curves correspond to a model trained using gut microbiome **(a)** and metabolome **(b)** data from *ApoE*$^{-}$ mouse model to predict IHH exposure in *Ldlr*$^{-}$ mice. Red curves show the same for microbiome **(c)** and metabolome **(d)** data from *Ldlr*$^{-}$ mice tested on *ApoE*$^{-}$ mice. IHH, intermittent hypoxia and hypercapnia.

Longitudinal dynamics of IHH-associated changes in the gut ecosystem: Next, we used these longitudinal data sets to learn how the duration of IHH-exposure impacts the gut microbiome and metabolome over time; and if this is consistent across the mouse models. The goal was to compare the dynamics of changes in the gut ecosystem with chronic IHH exposure in the *ApoE*

and *Ldlr*- mice. We tested this by assessing the capability of the RF classifier to distinguish IHH samples from control at each time point. In *ApoE*- mice, the classification AUC using gut microbiome data is high (constantly 1) at each time point starting at 11 weeks of age. The microbiome in *Ldlr*- mice, however, appears more predictive only at later time points, with its classification AUC improving from 0.71 at week 11 to more than 0.99 beyond week 14. We also observed a similar lag in gut metabolome changes in *Ldlr*- compared to *ApoE*- animals (Table S2[4]). Importantly, this is concordant with our previous finding that the atherosclerotic lesions evolved slowly and mildly in *Ldlr*- mice as compared to *ApoE*- mice (4). Therefore, observing this trend in both 'omics layers provides supporting evidence that the atherosclerosis phenotype in these animals is linked to perturbations in their gut ecosystem. Moreover, the gut microbiome and metabolome changes occur quickly after IHH-exposure, before atherosclerotic lesions were observed, which was reported to be 4 weeks for *ApoE*- and 6 weeks for *Ldlr*- post IHH exposure (4).

Reproducible biomarkers of IHH exposure: The subsequent goal of this analysis was to narrow down the list of fecal biomarkers that are reproducibly predictive of IHH-exposure, thereby guiding future mechanistic and clinical studies. The RF classifiers used to distinguish IHH-exposed and control animals described above provided us with a ranked list of bacterial and chemical features important for prediction (classifier trained on *ApoE*- microbiome and metabolome data: Tables S3; classifier trained on *Ldlr*- data: Table S4)[4]. We examined the features that were top ranked predictors (top 30 ranks) in both Ldlr- and ApoE-based classifiers. To investigate if there were some key biomarkers that could single-handedly distinguish IHH from control, we used the abundance of each of these features individually to plot ROC curves and

[4]https://msystems.asm.org/content/4/2/e00058-19

compute AUCs. Indeed, some of these microbial (Figure 4.3a) and chemical (Figure 4.3d) features could alone detect IHH-exposure within each mouse model highly accurately (AUC>0.75; see Table S5, 6 for the AUC values per feature and model)[4].

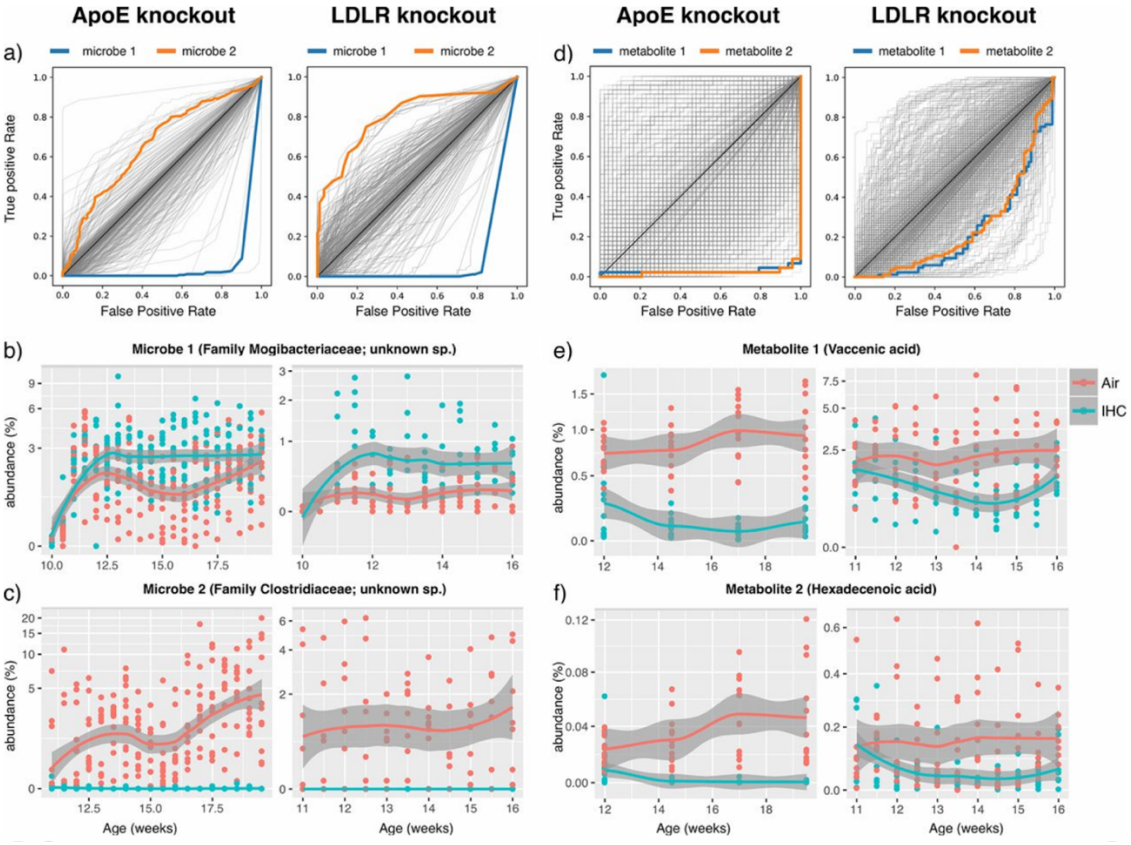

Figure 4.3. *Individual microbes and metabolites that distinguish the IHH from control groups in both ApoE and Ldlr mice*. **(a)** The ROC curves using each individual microbe abundance. Each curve represents the sensitivity and specificity as a function of the abundance of a single microbe to distinguish IHH and control groups. The curves for microbes enriched in IHH are above the diagonal line while those for microbes depleted in IHH are below the diagonal line. The two microbial features that have highest AUC (the values in the parentheses) in both mouse models are highlighted by color (both from the order of Clostridiales). The AUC here is defined as the area under the curve if the curve is above the diagonal line, or one minus the area under the curve for the curve below the diagonal line. **(b, c)** The abundance trends of these 2 microbes in each mouse model over time. **(d, e, f)** Similar plots for metabolites. ROC, receiver operating characteristic; IHH, intermittent hypoxia and hypercapnia; AUC, Area under the ROC curve.

We used our longitudinal data to compare trends of these predictive features in IHH-exposed and control groups in both animal models (Figure 4.3b, c, e, f). These predictors included

[4] https://msystems.asm.org/content/4/2/e00058-19

bacterial strains from the families Clostridiaceae and molecules identified as muricholic acid (bile acid; level 1 identification [(25)]) and vaccenic acid (fatty acid; level 2 identification). The goal was to investigate if these microbial and chemical species changed in the same direction on IHH-exposure in both *ApoE*$^{-}$ and *Ldlr*$^{-}$ mice or had idiosyncratic responses to IHH exposure based on the genetic background of the host. Figure 4.3c, e, and f show trends in these consistently altered features. These microbes and metabolites highlight key IHH-related changes in the gut microenvironment, and could guide subsequent reconstitution experiments in germ-free mice to establish causality. It is noteworthy that one unclassified species from the order Clostridiales (Figure 4.3b), despite being highly predictive within each animal, was depleted in IHH in *ApoE*$^{-}$ mice but enriched in *Ldlr*$^{-}$ mice. This, together with the high cross-genotype prediction accuracy using all features (Figure 4.2), suggests that although the microbiome and metabolome changes induced by IHH are reproducible across mouse models overall, there do exist animal model-specific changes as well. Hence, multi-animal model studies such as this are highly advantageous in precisely identifying biomarkers associated with an intervention of interest.

4.3 Discussion

We examined the reproducibility of IHH-associated alterations in the gut microbiome and metabolome of *Ldlr*$^{-}$ and *ApoE*$^{-}$ mouse models, crucial for understanding links between OSA and associated cardiovascular pathologies. As both APOE and LDLR are important in clearing cholesterol and triglyceride-rich particles from the blood, both models show elevated plasma cholesterol levels. However, they develop atherosclerotic plaques to different extents under high-fat dietary conditions (26–28). Concordant with these phenotypic differences, we highlight throughout that the gut ecosystem of the two models is also intrinsically distinct. As technical

variables such as origin of animals, housing conditions, experimental batches and data acquisition protocols are important considerations for meta-analyses such as ours (29, 30), we ensured that all animals were housed and handled in the same facility and data were acquired using identical protocols to minimize confounding effects. Furthermore, we used supervised machine learning to identify features specifically associated with IHH-exposure in both animal models reproducibly.

To our knowledge, the impact of IHH on the gut microbiome and metabolome in the context of atherosclerosis has not been investigated before, making our work exploratory in nature. Intermittent hypoxia alone (without hypercapnia or HFD) has been reported to significantly alter the microbiome in wild-type mice (31) and guinea-pigs (32) which lends support to our findings with IHH-exposure. Another study modeled human OSA and its cardiovascular consequences in HFD-fed rats by inflating a tracheal balloon during the sleep cycle (33). The authors reported that HFD and OSA synergistically caused hypertension and gut-dysbiosis in these rats. This study also noted perturbations in members of the order Clostridiales in response to HFD. Curiously, we also observe a member of this order to be highly predictive of IHH, yet changing in different directions in *ApoE*$^{-}$ and *Ldlr*$^{-}$ animals, hinting that this may be due to the differential impact of HFD on the two models (Figure 4.1b). In addition to genotype-specific changes, we also report consistent changes to unclassified strains belonging to the families Ruminococcaceae, Mogibacteriaceae, Lachnospiraceae and Clostridiaceae (Figure 4.S3). These taxonomic groups have been associated with cardiovascular, metabolic and inflammatory conditions previously (34–36), which indicates shared mechanistic pathways in OSA-associated cardiovascular conditions. Furthermore, our work is the first to profile OSA-associated changes in the gut metabolome at this scale. We observed reproducible perturbations in clinically relevant biomolecules in both *ApoE*$^{-}$ and *Ldlr*$^{-}$ mice. For example, Vaccenic acid, a trans-fatty acid that has been reported to lower LDL cholesterol and

triglyceride levels in rats (37) was found to decrease under IHH-exposure in both models. Similarly, bile acid molecules such as muricholic acid and taurocholic acid were more abundant in IHH-exposed versus control animals. Bile acids are crucial for not only facilitating transport of dietary fats and cholesterol in the host but also regulating host energy expenditure, glucose homeostasis and anti-inflammatory immune responses (38–42). Many metabolic and cardiovascular conditions (43) have been associated with aberrant bile acid profiles, suggesting that prolonged perturbations in these key molecules could contribute to downstream adverse cardiovascular consequences of OSA as well.

In summary, our work provides reproducible candidate biomarkers of IHH-exposure in animal models (and potentially OSA in humans) and will be most applicable to designing diagnostic and treatment modalities. Furthermore, by identifying consistent alterations across different model systems, we outline a general pipeline to select for biomarkers and therapeutic targets applicable to other intervention studies as well. We have made these information rich datasets publicly available to promote collaborative progress in this area of research.

4.4 Materials and Methods

Animals

Atherosclerosis-prone ten-week old male *Ldlr*$^{-}$ (n=16) and *ApoE*$^{-}$ (n=24) mice on C57BL/6J background (Stock Numbers 002207 and 002052 respectively; The Jackson Laboratory, Bar Harbor, ME) were used in this study (44). *Ldlr* and *ApoE* deficiencies were confirmed by PCR according to the vendor's instructions. Animals were either exposed to intermittent hypoxia and hypercapnia (n=8 and n=12 for *Ldlr*$^{-}$ and *ApoE*$^{-}$ animals, respectively) or air (control group) and fed with a high fat diet. All animal protocols were approved by the Animal Care Committee of the

University of California San Diego and followed the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health.

High-Fat Diet Treatment

Mice were fed with regular chow consisting of 0.01% cholesterol and 4.4% fat (TD.8604; Envigo-Teklad, Madison, WI) until initiation of dietary and IHH treatments. Starting at 10 weeks of age, male mice were provided with a high fat diet (HFD) containing 1.25% cholesterol and 21% milk fat (4.5 Kcal/g; TD.96121; Envigo-Teklad Madison, WI) while being exposed to either IHH or room air. Body weight of each mouse was measured twice a week. Food intake of animals in each cage was recorded twice a week.

Intermittent Hypoxia and Hypercapnia Exposure

Intermittent hypoxia and hypercapnia (IHH) was maintained in a computer-controlled atmosphere chamber system (OxyCycler, Reming Bioinstruments, Redfield, NY) as previously described (4). IHH exposure was introduced to the mice in short periods (~4 min) of synchronized reduction of O_2 (from 21% to 8%) and increasing of CO_2 (from ~0.5% to 8%) separated by alternating periods (~4 min) of normoxia ([O_2] = 21%) and normocapnia ([CO_2] = ~0.5%) with 1–2 min ramp intervals for 10 hours per day during the light cycle. This treatment protocol mimics the severe clinical condition observed in obstructive sleep apnea patients. Mice on the same HFD but in room air were used as controls. Fecal samples were collected at baseline and twice each week for 6 weeks (*Ldlr*$^{-/-}$) or 10 weeks (*ApoE*$^{-/-}$).

16S rRNA sequence processing

We performed 16S sequencing on fecal samples from *Ldlr*- and *ApoE*- mice for all the time points. DNA extraction and 16S rRNA amplicon sequencing were done using Earth Microbiome Project (EMP) standard protocols (http://www.earthmicrobiome.org/protocols-and-standards/16s).(45) In brief, DNA was extracted using the MO BIO PowerSoil DNA extraction kit (Carlsbad, CA). Amplicon PCR was performed on the V4 region of the 16S rRNA gene (Platinum Hot Start PCR 2X Master Mix, Invitrogen RED 13000014) 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 HiSeq 2500 sequencing platform. Sequence data were demultiplexed and minimally quality filtered using the QIIME 1.9.1 script split_libraries_fastq.py, with a Phred quality threshold of 3 and default parameters to generate per-study FASTA sequence files.

The raw sequence data were processed using the Deblur workflow (23) with default parameters in Qiita (46). This generated a sub-operational taxonomic unit (sOTU) abundance per sample (BIOM format) (47). Taxonomies for sOTUs were assigned using the sklearn-based taxonomy classifier trained on the Greengenes 13_8 99% OTUs (feature classifier plug-in) in QIIME 2. (48) The sOTU table was rarefied to a depth of 2,000 sequences/sample to control for sequencing effort. (49) A phylogeny was inferred using SATé-enabled phylogenetic placement, (50) which was used to insert 16S Deblur sOTUs into Greengenes 13_8 at a 99% phylogeny.

LC-MS/MS data processing

We acquired LC-MS/MS data on fecal samples from *Ldlr* (for 10 through 16 weeks of age) and *ApoE* (at ages 10, 12, 14.5, 17, and 19.5 weeks) mice using identical protocol. Details of data acquisition parameters are specified in (2). Briefly, fecal pellets (30 to 50 mg approximately) were extracted in 500 μl of 50:50 methanol-H2O solvent, followed by centrifugation to separate insoluble material. The extracts were dried completely by centrifugal evaporation (CentriVap centrifugal vacuum concentrator; Labconco, Kansas City, MO) and resuspended in 150 μl of methanol-H2O (1:1, vol/vol). After resuspension, the samples were analysed on a Vanquish ultrahigh-performance liquid chromatography (UPLC) system coupled to a Q Exactive orbital ion trap (Thermo Fisher Scientific, Bremen, Germany). A C18 core shell column (Kinetex column, 50 by 2 mm, 1.7-μm particle size, 100-Å pore size; Phenomenex, Torrance, CA) with a flow rate of 0.5 ml/min (solvent A, H2O-0.1% formic acid [FA]; solvent B, acetonitrile-0.1% FA) was used for chromatographic separation (2).

The raw data sets were converted to *m/z* extensible markup language (mzXML) in centroid mode using MSConvert (part of ProteoWizard)(51)(52). All mzXML files were cropped with an *m/z* range of 75.00 to 1,000.00 Da. Feature extraction was performed in MZmine2 (http://mzmine.sourceforge.net/) (53) with a signal intensity threshold of 2.0e5 and minimum peak width of 0.3-s. The maximum allowed mass and retention time tolerances were 10 ppm and 10 s, respectively. Local minimum search algorithm with a minimum relative peak height of 1% was used for chromatographic deconvolution; maximum peak width was set to 1 min. The detected peaks were aligned across all samples using the above-mentioned retention time and mass tolerances producing the final feature table used in these analyses. (see MZmine2 batch file:

https://github.com/knightlab-analyses/crossmodel_prediction/blob/master/data/metabolome/fileS7.mzmine2_batch.xml).

We performed molecular networking (54, 55) in GNPS (https://gnps.ucsd.edu/) to putatively identify molecular features using MS/MS-based spectral library matches. The parameters used for molecular networking are at the following URL: https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=3dbc660b9bdd4f699d31750d99b25463. Additionally, we purchased analytical standards for bile acids of interest (based on previous work (2, 55); alpha/beta-muricholic acid, chenodeoxycholic acid, cholic acid, lithocholic acid, deoxycholic acid, taurodeoxycholic acid) from Cayman Chemical (Ann Arbor, MI). We analyzed them using the same LC-MS/MS method described above to compare and verify the exact masses, fragmentation patterns, and retention times to ensure level 1 annotations, as defined by the 2007 metabolomics standards initiative (25).

Sharedness of microbial and metabolomic features across animal models

We calculated the sharedness of microbial features as follows. To quality-control the 16S sequences obtained per animal model, we retained only reads that were prevalent within each model i.e. above a sum relative abundance threshold of 10E-06 and present in at least 1% of the samples, thus avoiding sequencing noise. The number of such reads in *Ldlr*- and *ApoE*- animals was 635 and 582, respectively. Out of these, 248 sequences were shared between the two models. Therefore, the percentage of microbiome features shared between the animal models was 39% of unique microbial features found in *Ldlr*- (and 42% of those in *ApoE*-) models.

For metabolomic data, we quality-controlled the chemical features by retaining those above a sum relative abundance threshold of 10E-01 and present in at least 10% of all samples for each

animal model individually. There were 267 and 374 such features in *Ldlr*$^{-}$ and *ApoE*$^{-}$ animals, respectively. Out of these, 137 metabolites were shared between the two models. Thus, the percentage shared between the animal models were 51% of total features in *Ldlr*$^{-}$ (and 36% of those in *ApoE*$^{-}$) knockout models.

Effect size analyses

Effect sizes were calculated over the individual genotype, mice, cage number, age, exposure type. For each of these covariates, we applied the mixed directional FDR (mdFDR) (56) methodology to test for the significance of each pairwise comparison among the groups. For each significant pairwise comparison via PERMANOVA (24), we computed the effect size using Cohen's d (57) or the absolute difference between the mean of each group divided by the pooled standard deviation. As diversity estimators we used unweighted UniFrac and Bray-Curtis distances matrices for the 16S rRNA sequencing and LC-MS/MS mass-spectrometry, respectively.

For the microbiome data layer (Table S1[4]), when taking both genotypes together, we see that the first three largest effect sizes are mouse number, age and cage number, followed by genotype and exposure type. It is important to note that the maximum difference (effect size) on the first three covariates are related to genotype differences. For example, the maximum difference in mouse number is between two mice [mouse numbers 105 (*ApoE*$^{-}$) vs. 32 (*Ldlr*$^{-}$); Figure 4.S1] that belong to two different genotypes and exposure types. To untangle the effect of genotype, we stratified our dataset by genotypes and calculated effect sizes of each of the covariates within each model. It is noteworthy that effects of covariates are ranked differently within each model, hinting

[4] https://msystems.asm.org/content/4/2/e00058-19

towards underlying differences in the characteristics of the microbial community. Nevertheless, the effect of exposure is ranked comparably across models.

Similarly, we calculate effect sizes of the above mentioned covariates for the metabolome data layer (Table S1[4]). When taking both genotypes together, consistent with the microbiome results, mouse number, age and cage number have the largest effect sizes, and the groups with the maximum effects belong to different genotypes [e.g. mouse number 114 (*ApoE*) vs. 17 (*Ldlr*)]. We then stratified the data by genotype and observed that different covariates had distinct effects within each genotype. Interestingly, our analysis shows that unlike in *Ldlr*- mice, individual variability was not significant in *ApoE*- mice.

Supervised classification

Random Forest (RF) classifier was trained and evaluated with cross validation for each mouse model, using microbial or chemical features as predictors. During cross validation, all the samples from the same mouse appeared only in either training or validation data but not both to avoid over-optimistic cross-validation accuracy scores as a result of the classifier learning idiosyncrasies of the individual itself rather than the treatment. The classifiers trained for each mouse model were then applied on the samples of the other mouse model for cross-genotype prediction. For the longitudinal prediction, we trained and evaluated a RF classifier on the samples collected at each time point for AUC computation. To assess the capability of individual 16S sequences and metabolites to separate IHH-exposed from control animals, we used the abundance of each feature as the score to plot ROC curve and compute AUC, and highlighted the features that can single-handedly distinguish IHH on ROC plots. These analyses were done using the scikit-learn Python package.

4.5 Acknowledgements

We thank Lingjing Jiang for very helpful suggestions and discussions regarding statistical analyses. We acknowledge NIH Grants GMS10RR029121 and 5P41GM103484-07 for the shared instrumentation and computational infrastructure that enabled this work.

4.6 Author contributions

Chapter 4, in full, is a reprint of previously published material: Tripathi, A., Xu, Z. Z., Xue, J., Poulsen, O., Gonzalez, A., Humphrey, G., Meehan, M. J., Melnik, A. V., Ackermann, G., Zhou, D., Malhotra, A., Haddad, G. G., Dorrestein, P. C., & Knight, R. (2019). Intermittent Hypoxia and Hypercapnia Reproducibly Change the Gut Microbiome and Metabolome across Rodent Model Systems. *mSystems*, *4*(2).

A.T. and Z.Z.X. contributed equally to this article. All authors worked together to finalize and approve this manuscript.The co-authors listed above supervised or provided support for the research and have given permission for the inclusion of the work in this dissertation.

4.7 Competing interests

The authors declare no competing financial interests.

4.8 Data and code availability

The data generated in this study are available publicly under the following accession numbers: for metabolomics data: MSV000081482 (Ldlr knockout animal), MSV000082813 (ApoE knockout animal), MSV000081853 (commercial standards); and for microbiome data: ERP106495 (Ldlr knockout animals; EBI database) and ERP110592 (ApoE knockout animals).

Data analysis has been documented in Jupyter notebooks available on GitHub (https://github.com/knightlab-analyses/crossmodel_prediction)

4.9 Supplemental figures

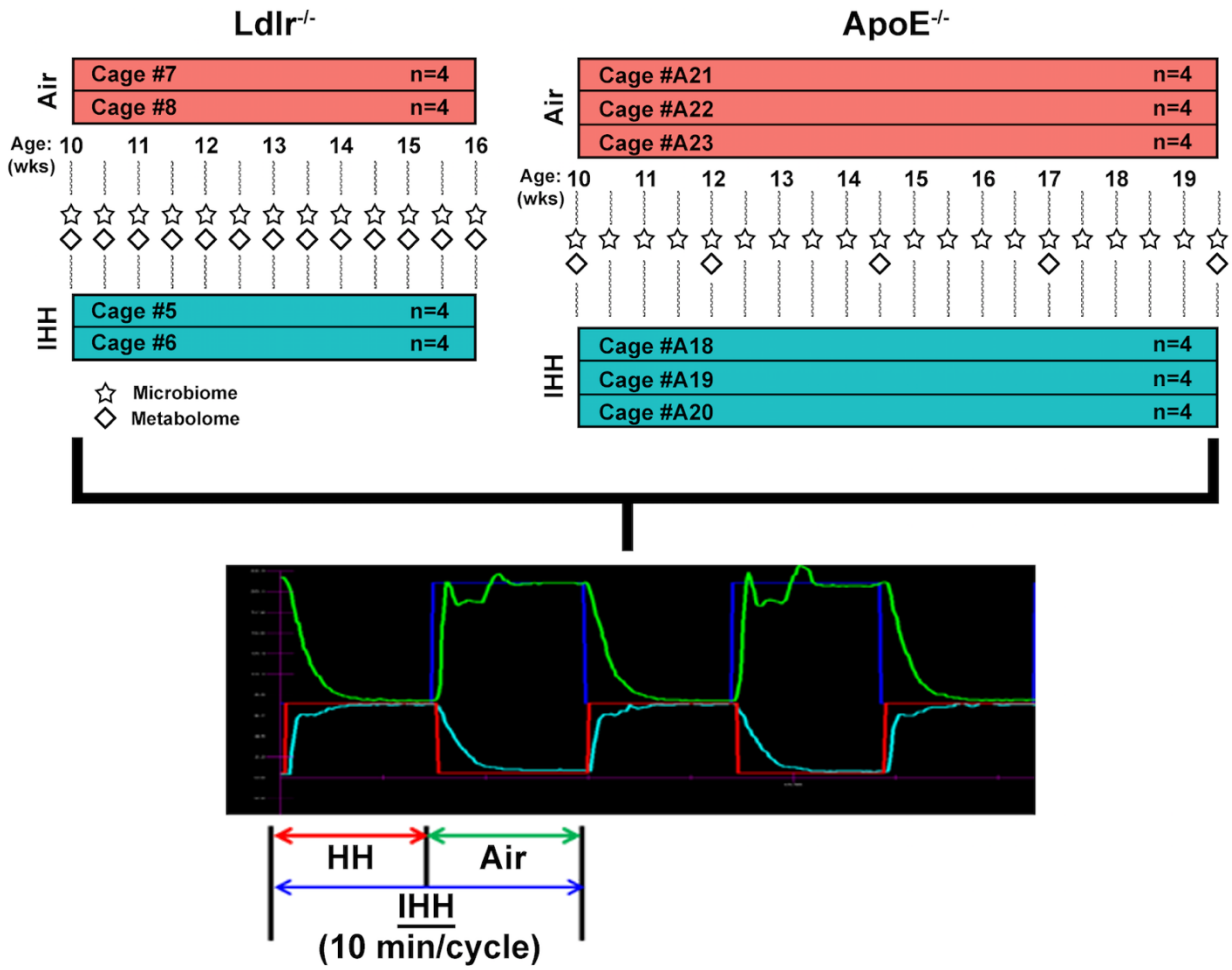

Figure 4.S1 *Schematic illustration of treatment paradigm and sample collection*. Groups of 8-week old male *Ldlr*$^{-/-}$ mice or *ApoE*$^{-/-}$ mice were transferred to the treatment room for 2 weeks of acclimatization with room air and regular chow food. At 10 weeks of age, mice were switched to high-fat diet (HFD) and treated with or without intermittent hypoxia and hypercapnia (IHH). The IHH treatment group received 10 hrs/day IHH in the light cycle for 6 weeks in *Ldlr*$^{-/-}$ mice or for 10 weeks in *ApoE*$^{-/-}$ mice (The blue line is the O_2 set point and the green was the actual level of O_2. The red line is the CO_2 set point and light blue was the actual level of CO_2). The control groups remained in room air for the same period. All the mice are reared in the same animal facility room thus have the same microbial exposure. Fecal pellets were collected at baseline and twice per week thereafter and were used for microbiome and metabolome analyses. All the time points were analyzed except for metabolome of *ApoE*$^{-/-}$ mice (only fecal samples at the age of 10, 12, 14.5, 17 and 19.5 weeks were analyzed). Mice belonging to each treatment group were split between multiple cages to untangle cage effects with the effect of the treatment. For *Ldlr*$^{-/-}$ mice, IHH: mouse numbers 17-20 and 21-24 were kept in cage numbers 5 and 6, respectively; Air: mouse 25-28 and 29-32 were kept in cage numbers 7 and 8, respectively. For *ApoE*$^{-/-}$ mice, IHH: mouse numbers 97-100, 101-104, and 105-108 were kept in cage numbers A18, A19, and A20, respectively; Air: mouse numbers 109-112, 113-116, and 117-120 were kept in cage numbers A21, A23 and A23, respectively. Comprehensive sample metadata is available publicly [see data availability].

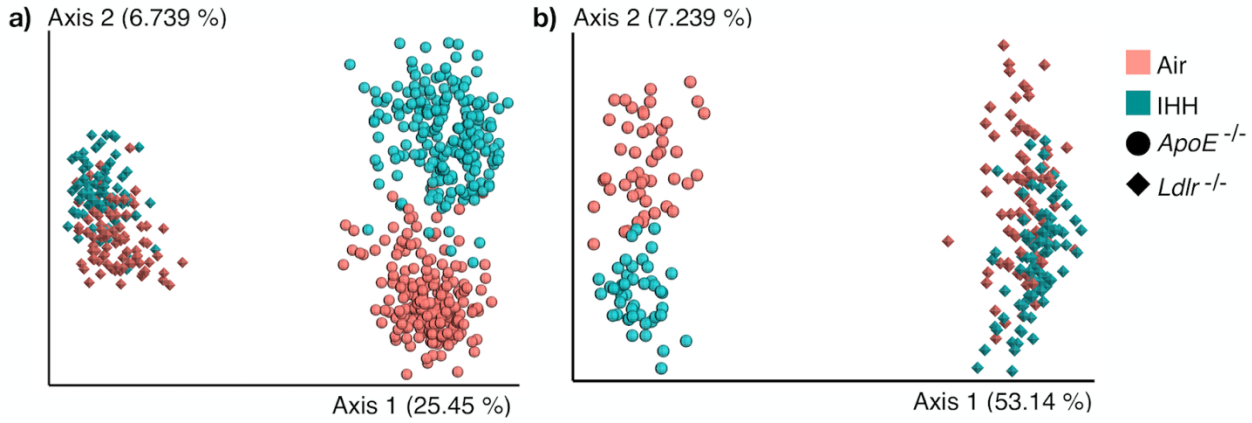

Figure 4.S2 *Principal-coordinate analysis (PCoA) of the gut microbiome and metabolome in ApoE and Ldlr mouse models without baseline samples*. (a) PCoA of the gut microbiome using unweighted UniFrac after removal of samples collected at 10 and 10.5 weeks of age. (b) Similar PCoA for gut metabolome using Bray-Curtis distances. Baseline samples were removed from this plot to better visualize the impact of IHH-exposure alone. Axis 1 explains the variability due to the animal models and axis 2 consistently captures the IHH-associated changes in the microbiome and metabolome in both models. Axis 1, principal coordinate 1; IHH, intermittent hypoxia and hypercapnia

4.10 References

1. McNicholas WT, Bonsigore MR, Management Committee of EU COST ACTION B26. 2007. Sleep apnoea as an independent risk factor for cardiovascular disease: current evidence, basic mechanisms and research priorities. Eur Respir J 29:156–178.

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

3. Douglas RM, Bowden K, Pattison J, Peterson AB, Juliano J, Dalton ND, Gu Y, Alvarez E, Imamura T, Peterson KL, Witztum JL, Haddad GG, Li AC. 2013. Intermittent hypoxia and hypercapnia induce pulmonary artery atherosclerosis and ventricular dysfunction in low density lipoprotein receptor deficient mice. J Appl Physiol 115:1694–1704.

4. Xue J, Zhou D, Poulsen O, Imamura T, Hsiao Y-H, Smith TH, Malhotra A, Dorrestein P, Knight R, Haddad GG. 2017. Intermittent Hypoxia and Hypercapnia Accelerate Atherosclerosis, Partially via Trimethylamine-Oxide. Am J Respir Cell Mol Biol 57:581–588.

5. Lui MM-S, Sau-Man M. 2012. OSA and atherosclerosis. J Thorac Dis 4:164–172.

6. Franklin CL, Ericsson AC. 2017. Microbiota and reproducibility of rodent models. Lab Anim 46:114–122.

7. Poussin C, Sierro N, Boué S, Battey J, Scotti E, Belcastro V, Peitsch MC, Ivanov NV, Hoeng J. 2018. Interrogating the microbiome: experimental and computational considerations in support of study reproducibility. Drug Discov Today.

8. Goodrich JK, Di Rienzi SC, Poole AC, Koren O, Walters WA, Gregory Caporaso J, Knight R, Ley RE. 2014. Conducting a Microbiome Study. Cell 158:250–262.

9. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, Gonzalez A, Kosciolek T, McCall L-I, McDonald D, Melnik AV, Morton JT, Navas J, Quinn RA, Sanders JG, Swafford AD, Thompson LR, Tripathi A, Xu ZZ, Zaneveld JR, Zhu Q, Caporaso JG, Dorrestein PC. 2018. Best practices for analysing microbiomes. Nat Rev Microbiol 16:410–422.

10. Zhou Y, Xu ZZ, He Y, Yang Y, Liu L, Lin Q, Nie Y, Li M, Zhi F, Liu S, Amir A, González A, Tripathi A, Chen M, Wu GD, Knight R, Zhou H, Chen Y. 2018. Gut Microbiota Offers Universal Biomarkers across Ethnicity in Inflammatory Bowel Disease Diagnosis and Infliximab Response Prediction. mSystems 3.

11. Loomba R, Seguritan V, Li W, Long T, Klitgord N, Bhatt A, Dulai PS, Caussy C, Bettencourt R, Highlander SK, Jones MB, Sirlin CB, Schnabl B, Brinkac L, Schork N, Chen C-H, Brenner DA, Biggs W, Yooseph S, Venter JC, Nelson KE. 2017. Gut Microbiome-Based Metagenomic Signature for Non-invasive Detection of Advanced Fibrosis in Human Nonalcoholic Fatty Liver Disease. Cell Metab 25:1054–1062.e5.

12. Sze MA, Schloss PD. 2016. Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome. MBio 7.

13. Walters WA, Xu Z, Knight R. 2014. Meta-analyses of human gut microbes associated with obesity and IBD. FEBS Lett 588:4223–4233.

14. Forslund K, Hildebrand F, Nielsen T, Falony G, Le Chatelier E, Sunagawa S, Prifti E, Vieira-Silva S, Gudmundsdottir V, Pedersen HK, Arumugam M, Kristiansen K, Voigt AY, Vestergaard H, Hercog R, Costea PI, Kultima JR, Li J, Jørgensen T, Levenez F, Dore J, MetaHIT consortium, Nielsen HB, Brunak S, Raes J, Hansen T, Wang J, Ehrlich SD, Bork P, Pedersen O. 2015. Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. Nature 528:262–266.

15. Breiman L. 2001. 10.1023/A:1010933404324. Machine Learning.

16. Knights D, Kuczynski J, Charlson ES, Zaneveld J, Mozer MC, Collman RG, Bushman FD, Knight R, Kelley ST. 2011. Bayesian community-wide culture-independent microbial source tracking. Nat Methods 8:761–763.

17. Yazdani M, Taylor BC, Debelius JW, Li W, Knight R, Smarr L. 2016. Using machine learning to identify major shifts in human gut microbiome protein family abundance in disease2016 IEEE International Conference on Big Data (Big Data).

18. Caruana R, Munson A, Niculescu-Mizil A. 2006. Getting the Most Out of Ensemble SelectionSixth International Conference on Data Mining (ICDM'06).

19. Vázquez-Baeza Y, Hyde ER, Suchodolski JS, Knight R. 2016. Dog and human inflammatory bowel disease rely on overlapping yet distinct dysbiosis networks. Nat Microbiol 1:16177.

20. Borg I, Groenen P. 2003. Modern Multidimensional Scaling: Theory and Applications. Journal of Educational Measurement 40:277–280.

21. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA, Biddinger SB, Dutton RJ, Turnbaugh PJ. 2014. Diet rapidly and reproducibly alters the human gut microbiome. Nature 505:559–563.

22. Esko T, Hirschhorn JN, Feldman HA, Hsu Y-HH, Deik AA, Clish CB, Ebbeling CB, Ludwig DS. 2017. Metabolomic profiles as reliable biomarkers of dietary composition. Am J Clin Nutr 105:547–554.

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

24. Anderson MJ. 2017. Permutational Multivariate Analysis of Variance (PERMANOVA), p. 1–15. In Wiley StatsRef: Statistics Reference Online.

25. Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, -M. Fan TW, Fiehn O, Goodacre R, Griffin JL, Hankemeier T, Hardy N, Harnly J, Higashi R, Kopka J, Lane AN, Lindon JC, Marriott P, Nicholls AW, Reily MD, Thaden JJ, Viant MR. 2007. Proposed minimum reporting standards for chemical analysis. Metabolomics 3:211–221.

26. Ishibashi S, Brown MS, Goldstein JL, Gerard RD, Hammer RE, Herz J. 1993. Hypercholesterolemia in low density lipoprotein receptor knockout mice and its reversal by adenovirus-mediated gene delivery. J Clin Invest 92:883–893.

27. Piedrahita JA, Zhang SH, Hagaman JR, Oliver PM, Maeda N. 1992. Generation of mice carrying a mutant apolipoprotein E gene inactivated by gene targeting in embryonic stem cells. Proc Natl Acad Sci U S A 89:4471–4475.

28. Which JAX mouse model is best for atherosclerosis studies: Apoe or Ldlr knockout mice? The Jackson Laboratory.

29. Ericsson AC, Gagliardi J, Bouhan D, Spollen WG, Givan SA, Franklin CL. 2018. The influence of caging, bedding, and diet on the composition of the microbiota in different regions of the mouse gut. Sci Rep 8:4065.

30. Kim D, Hofstaedter CE, Zhao C, Mattei L, Tanes C, Clarke E, Lauder A, Sherrill-Mix S, Chehoud C, Kelsen J, Conrad M, Collman RG, Baldassano R, Bushman FD, Bittinger K. 2017. Optimizing methods and dodging pitfalls in microbiome research. Microbiome 5:52.

31. Moreno-Indias I, Torres M, Montserrat JM, Sanchez-Alcoholado L, Cardona F, Tinahones FJ, Gozal D, Poroyko VA, Navajas D, Queipo-Ortuño MI, Farré R. 2015. Intermittent hypoxia alters gut microbiota diversity in a mouse model of sleep apnoea. Eur Respir J 45:1055–1065.

32. Lucking EF, O'Connor KM, Strain CR, Fouhy F, Bastiaanssen TFS, Burns DP, Golubeva AV, Stanton C, Clarke G, Cryan JF, O'Halloran KD. 2018. Chronic intermittent hypoxia disrupts cardiorespiratory homeostasis and gut microbiota composition in adult male guinea-pigs. EBioMedicine 38:191–205.

33. Durgan DJ, Ganesh BP, Cope JL, Ajami NJ, Phillips SC, Petrosino JF, Hollister EB, Bryan RM Jr. 2016. Role of the Gut Microbiome in Obstructive Sleep Apnea-Induced Hypertension. Hypertension 67:469–474.

34. Kasselman LJ, Vernice NA, DeLeon J, Reiss AB. 2018. The gut microbiome and elevated cardiovascular risk in obesity and autoimmunity. Atherosclerosis 271:203–213.

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

36. Pascal E al V. A microbial signature for Crohn's disease. - PubMed - NCBI.

37. Wang Y, Jacome-Sosa MM, Ruth MR, Goruk SD, Reaney MJ, Glimm DR, Wright DC, Vine DF, Field CJ, Proctor SD. 2009. Trans-11 vaccenic acid reduces hepatic lipogenesis and chylomicron secretion in JCR:LA-cp rats. J Nutr 139:2049–2054.

38. Schaap FG, Trauner M, Jansen PLM. 2013. Bile acid receptors as targets for drug development. Nat Rev Gastroenterol Hepatol 11:55–67.

39. Perino A, Schoonjans K. 2015. TGR5 and Immunometabolism: Insights from Physiology and Pharmacology. Trends Pharmacol Sci 36:847–857.

40. Zarrinpar A, Loomba R. 2012. Review article: the emerging interplay among the gastrointestinal tract, bile acids and incretins in the pathogenesis of diabetes and non-alcoholic fatty liver disease. Aliment Pharmacol Ther 36:909–921.

41. Broeders EPM, Nascimento EBM, Havekes B, Brans B, Roumans KHM, Tailleux A, Schaart G, Kouach M, Charton J, Deprez B, Bouvy ND, Mottaghy F, Staels B, van Marken

Lichtenbelt WD, Schrauwen P. 2015. The Bile Acid Chenodeoxycholic Acid Increases Human Brown Adipose Tissue Activity. Cell Metab 22:418–426.

42. Thomas C, Gioiello A, Noriega L, Strehle A, Oury J, Rizzo G, Macchiarulo A, Yamamoto H, Mataki C, Pruzanski M, Pellicciari R, Auwerx J, Schoonjans K. 2009. TGR5-mediated bile acid sensing controls glucose homeostasis. Cell Metab 10:167–177.

43. Joyce SA, Gahan CGM. 2017. Disease-Associated Changes in Bile Acid Profiles and Links to Altered Gut Microbiota. Dig Dis 35:169–177.

44. Ishibashi S, Brown MS, Goldstein JL, Gerard RD, Hammer RE, Herz J. 1993. Hypercholesterolemia in low density lipoprotein receptor knockout mice and its reversal by adenovirus-mediated gene delivery. J Clin Invest 92:883–893.

45. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, Gormley N, Gilbert JA, Smith G, Knight R. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J 6:1621–1624.

46. Gonzalez A, Navas-Molina JA, Kosciolek T, McDonald D, Vázquez-Baeza Y, Ackermann G, DeReus J, Janssen S, Swafford AD, Orchanian SB, Sanders JG, Shorenstein J, Holste H, Petrus S, Robbins-Pianka A, Brislawn CJ, Wang M, Rideout JR, Bolyen E, Dillon M, Caporaso JG, Dorrestein PC, Knight R. 2018. Qiita: rapid, web-enabled microbiome meta-analysis. Nat Methods 15:796–798.

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

48. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. 2010. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7:335–336.

49. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, Lozupone C, Zaneveld JR, Vázquez-Baeza Y, Birmingham A, Hyde ER, Knight R. 2017. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome 5:27.

50. Mirarab S, Nguyen N, Warnow T. 2011. SEPP: SATé-Enabled Phylogenetic PlacementBiocomputing 2012.

51. Adusumilli R, Mallick P. 2017. Data Conversion with ProteoWizard msConvert. Methods Mol Biol 1550:339–368.

52. Mirzaei H, Carrasco M. 2016. Modern Proteomics – Sample Preparation, Analysis and Practical Applications. Springer.

53. Katajamaa M, Miettinen J, Oresic M. 2006. MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. Bioinformatics 22:634–636.

54. Watrous J, Roach P, Alexandrov T, Heath BS, Yang JY, Kersten RD, van der Voort M, Pogliano K, Gross H, Raaijmakers JM, Moore BS, Laskin J, Bandeira N, Dorrestein PC. 2012. Mass spectral molecular networking of living microbial colonies. Proc Natl Acad Sci U S A 109:E1743–52.

55. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T, Porto C, Bouslimani A, Melnik AV, Meehan MJ, Liu W-T, Crüsemann M, Boudreau PD, Esquenazi E, Sandoval-Calderón M, Kersten RD, Pace LA, Quinn RA, Duncan KR, Hsu C-C, Floros DJ, Gavilan RG, Kleigrewe K, Northen T, Dutton RJ, Parrot D, Carlson EE, Aigle B, Michelsen CF, Jelsbak L, Sohlenkamp C, Pevzner P, Edlund A, McLean J, Piel J, Murphy BT, Gerwick L, Liaw C-C, Yang Y-L, Humpf H-U, Maansson M, Keyzers RA, Sims AC, Johnson AR, Sidebottom AM, Sedio BE, Klitgaard A, Larson CB, P CAB, Torres-Mendoza D, Gonzalez DJ, Silva DB, Marques LM, Demarque DP, Pociute E, O'Neill EC, Briand E, Helfrich EJN, Granatosky EA, Glukhov E, Ryffel F, Houson H, Mohimani H, Kharbush JJ, Zeng Y, Vorholt JA, Kurita KL, Charusanti P, McPhail KL, Nielsen KF, Vuong L, Elfeki M, Traxler MF, Engene N, Koyama N, Vining OB, Baric R, Silva RR, Mascuch SJ, Tomasi S, Jenkins S, Macherla V, Hoffman T, Agarwal V, Williams PG, Dai J, Neupane R, Gurr J, Rodríguez AMC, Lamsa A, Zhang C, Dorrestein K, Duggan BM, Almaliti J, Allard P-M, Phapale P, Nothias L-F, Alexandrov T, Litaudon M, Wolfender J-L, Kyle JE, Metz TO, Peryea T, Nguyen D-T, VanLeer D, Shinn P, Jadhav A, Müller R, Waters KM, Shi W, Liu X, Zhang L, Knight R, Jensen PR, Palsson BO, Pogliano K, Linington RG, Gutiérrez M, Lopes NP, Gerwick WH, Moore BS, Dorrestein PC, Bandeira N. 2016. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol 34:828–837.

56. Guo W, Sarkar SK, Peddada SD. 2010. Controlling false discoveries in multidimensional directional decisions, with applications to gene expression data on ordered categories. Biometrics 66:485–492.

57. Cohen J. 1992. A power primer. Psychol Bull 112:155–159.

Chapter 5. Chemically-informed analyses of metabolomics mass spectrometry data with Qemistree

Untargeted mass spectrometry is employed to detect small molecules in complex biospecimens, generating data that are difficult to interpret. We developed Qemistree, a data exploration strategy based on hierarchical organization of molecular fingerprints predicted from fragmentation spectra, represented in the context of sample metadata and chemical ontologies. By expressing molecular relationships as a tree, we can apply ecological tools, designed around the relatedness of DNA sequences, to study chemical composition.

5.1 Introduction

Molecular networking (1), introduced in 2012, was one of the first data organization approaches to visualize the relationships between fragmentation spectra for similar molecules from tandem mass spectrometry data in the context of metadata. It formed the basis for the web-based mass spectrometry infrastructure, Global Natural Products Social Molecular Networking (2) (GNPS, https://gnps.ucsd.edu/) which sees ~200,000 new accessions per month. Molecular networking is used for a range of applications (3) in drug discovery, environmental monitoring, medicine, and agriculture. While molecular networking is useful for visualizing closely related molecular families, the inference of chemical relationships at a dataset-wide level and in the context of diverse metadata requires complementary representation strategies. To address this need, we developed an approach that uses fragmentation trees (4) and supervised machine learning (5) to calculate all pairwise chemical relationships and visualizes it in the context of sample metadata and molecular annotations. We show that a chemical tree enables the application of

various tree-based tools, originally developed for analyzing DNA sequencing data (6–9), for exploring mass-spectrometry data.

We introduce Qemistree, pronounced *chemis-tree*, a software that constructs a chemical tree from fragmentation spectra based on predicted molecular fingerprints (10). Molecular fingerprints are vectors where each position encodes a substructural property of the molecule. Recent methods allow us to predict molecular fingerprints from tandem mass spectra (11–15). In Qemistree, we use SIRIUS (16) and CSI:FingerID (13) to obtain predicted molecular fingerprints. The users first perform feature detection (17, 18) to generate a list of observed ions, referred to as chemical features henceforth, to be analyzed by Qemistree (Figure 5.S1). SIRIUS then determines the molecular formula of each feature using the isotope and fragmentation patterns, and estimates the best fragmentation tree explaining the fragmentation spectrum. Subsequently, CSI:FingerID operates on the fragmentation trees using kernel support vector machines to predict molecular properties (2936 properties; Supplementary Table 1[5]). We use these molecular fingerprints to calculate pairwise distances between chemical features that are hierarchically clustered to generate a tree representing their structural relationships. Although alternative approaches to hierarchically cluster features based on cosine similarity of fragmentation spectra exist (19–21), we use molecular fingerprints as it allows us to compare features based on a diverse range of structural properties predicted by CSI:FingerID. Additionally, as CSI:FingerID was shown to perform well for automatic *in silico* structural annotation (22), we leverage it to search molecular structural databases to provide complementary insights into structures when no match is obtained against spectral libraries. Subsequently, we use ClassyFire (23) to assign a 5-level chemical taxonomy

[5]https://www.biorxiv.org/content/10.1101/2020.05.04.077636v1.supplementary-material

(kingdom, superclass, class, subclass, and direct parent) to all molecules annotated via spectral library matching and *in silico* prediction.

Phylogenetic tools such as iTOL (24) can be used to visualize Qemistree trees interactively in the context of sample information and feature annotations for easy data exploration. The outputs of Qemistree can also be plugged into other workflows in QIIME 2 (25) (many of which were originally developed for microbiome sequence analysis) or in R, Python etc. for system-wide metabolomic data analyses (6, 7, 9, 26). Qemistree is available to the microbiome community as a QIIME 2 plugin (https://github.com/biocore/q2-qemistree) and the metabolomics community as a workflow on GNPS (2) (https://ccms-ucsd.github.io/GNPSDocumentation/qemistree/). The chemical tree from the GNPS workflow can be explored interactively (e.g. https://qemistree.ucsd.edu/).

5.2 Results

To verify that molecular fingerprint-based trees correctly capture the chemical relationships between molecules, we generated an evaluation dataset with two human fecal samples, a tomato seedling sample, and a human serum sample. Mixtures of these samples were prepared by combining them in gradually increasing proportions to generate a set of diverse but related metabolite profiles and untargeted tandem mass spectrometry was used to profile the chemical composition of these samples. Mass-spectrometry was performed twice using different chromatographic gradients causing a non-uniform retention time shift between the two runs. The data processing of these two experiments leads to the same molecules being detected as different chemical features in downstream analysis. In Figure 5.1a we highlight how these technical variations make the same samples appear chemically disjointed.

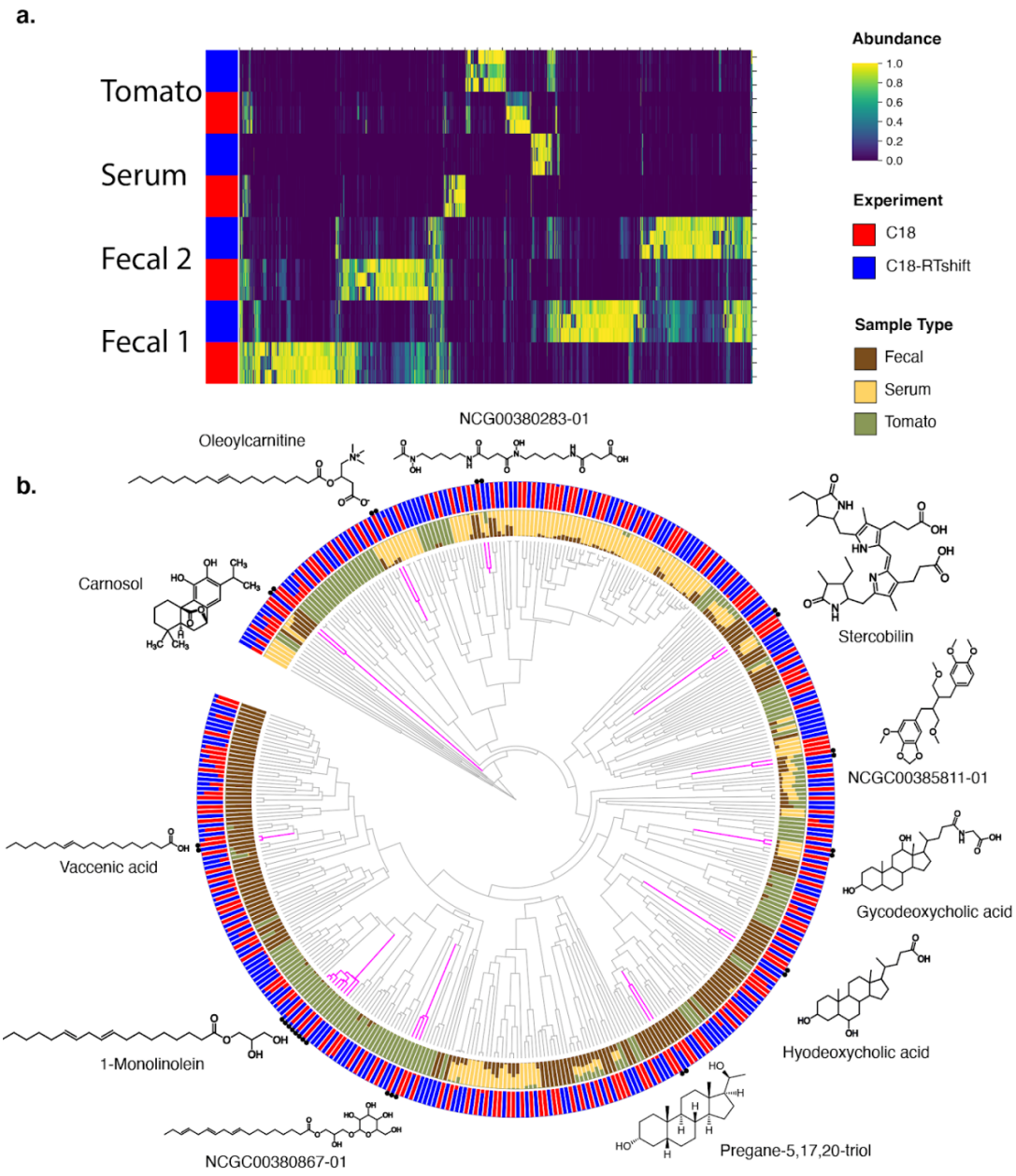

Figure 5.1 *Qemistree mitigates aspects of technical artifacts by co-clustering structurally similar molecules across mass spectrometry runs*. **a)** Sample (y-axis) by molecule (x-axis) heatmap of 2 fecal samples, tomato seedling samples, and serum samples in the evaluation dataset grouped by chromatography conditions. **b)** A chemical tree based on predicted molecular fingerprints representing the structural relationships between compounds detected in the evaluation dataset. Outer ring shows the relative abundance of molecules stratified by mass spectrometry run; inner ring shows the same stratified by fecal, serum and tomato samples in the evaluation dataset. Structurally similar molecules detected as different chemical features due to shift in retention time across mass spectrometry runs are clustered together; we highlight some examples of these artificially duplicated features around the tree. All structures shown are spectral reference library matches obtained from feature-based molecular networking (17) in GNPS: (https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=efda476c72724b29a91693a108fa5a9d; Metabolomics Standard Initiative (MSI) level 3 annotation) (27).

Using Qemistree, we map each of the spectra in the two chromatographic conditions (batches) to a molecular fingerprint, and organize these in a tree structure (Figure 5.1b). Because molecular fingerprints are independent of retention time shifts, spectra are clustered based on their chemical similarity. This tree structure can be decorated using sample type descriptions, chromatographic conditions, and spectral library matches obtained from molecular networking in GNPS. Figure 5.1 shows that similar chemical features are detected exclusively in one of the two batches. However, based on the molecular fingerprints, these chemical features were arranged as neighboring tips in the tree regardless of the retention time shifts. This result shows how Qemistree can reconcile and facilitate the comparison of datasets acquired on different chromatographic gradients.

We demonstrate the use of a chemical hierarchy in performing chemically-informed comparisons of metabolomics profiles. In standard metabolomic statistical analyses, each molecule is assumed unrelated to the other molecules in the dataset. Some of the pitfalls of this assumption are highlighted in Figure 5.2a. Consider a scenario where we want to compare samples 1-3. An analysis schema that does not account for the chemical relationships among the molecules in these samples (Figure 5.2a, left), will assume that the sugars in samples 2 and 3 are as chemically related to the lipids in sample 1 as they are to each other. This would lead to the naive conclusion that samples 1 and 2, and samples 2 and 3 are equally distinct, yet they are not from a chemical perspective. On the other hand, if we account for the fact that sugar molecules are more chemically related to one another than they are to lipids, we can obtain a chemically-informed sample-to-sample comparison. Sedio and coworkers developed the chemical structural compositional similarity (CSCS) metric (28) to account for relationships between molecules based on the similarity of their fragmentation spectra. While CSCS compares samples based on modified

cosine scores obtained from molecular networking, we calculate chemical relationships based on structurally-informed molecular fingerprints. We express these relationships in the form of a hierarchy which enables the use of other tree-based tools for downstream data analyses. For example, in Figure 5.2a, we show that by using a tree of structural relationships between molecules in samples 1-3, we can apply UniFrac (9), a tree-informed distance metric and demonstrate that the composition of sample 1 is distinct from samples 2 and 3.

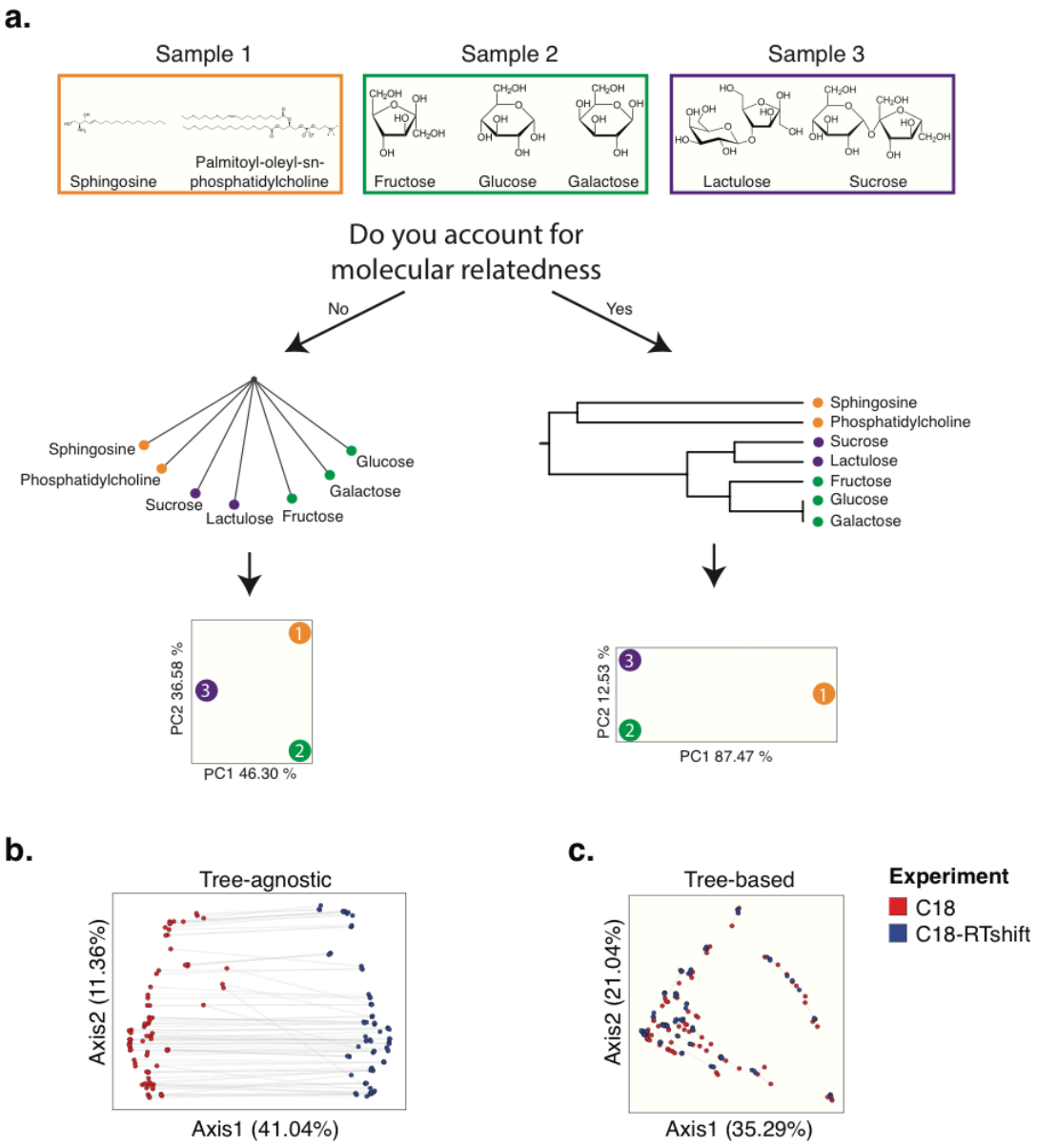

Figure 5.2 *The pitfalls of assuming equal relatedness of molecules and the advantages of a chemical tree for sample comparison*. **a)** A scenario where the goal is to compare the chemical composition in samples 1 (sphingosine and phosphatidylcholine), 2 (glucose, galactose, and fructose), and 3 (sucrose and lactulose). When we do not account for the chemical relationships between the molecules, i.e. assume that the lipid molecules in sample 1 are equally related to the sugars in samples 2 and 3 (left), we conclude that samples 1, 2, and 3 are similarly distinct. If we account for sugar molecules being more chemically related to one another than sugars are to lipid molecules (right), we can obtain a chemically-meaningful distance between samples. This is exemplified through a principal coordinates analysis (PCoA) of the computed UniFrac (9) (tree-based) distances among samples; we see that samples 2 and 3 are more similar to each other, and sample 1 which is chemically distinct is separated along the primary axis of variation, when distances are computed using the chemical tree. **b, c)** PCoA of samples in the evaluation dataset colored by chromatography conditions. PCoA plot using tree-agnostic (Bray-Curtis (29)) distances which do not account for the chemical relationship between features detected across chromatography conditions (b) and tree-based (Weighted UniFrac (9)) distances which are based on the hierarchical relationships between molecules in the evaluation dataset (c).

The importance of comparing samples by accounting for their molecular relatedness is highlighted when we contrast the results from ignoring the tree structure (Figure 5.2b) to those which integrate it (Figure 5.2c). With the structural context provided by Qemistree, the differences between replicates across batches are comparable to the within-batch differences (Figure 5.S2). The retention time shift in this dataset leads to a strong technical signal that obscures the biological relationships among the samples (permutational ANOVA; tree agnostic (29) pseudo-F=120.75, p=0.001 vs. tree informed (9) pseudo-F=18.2239, p=0.001). We observed and remediated a similar pattern originating from plate-to-plate variation in a recently published study investigating the metabolome and microbiome of captive cheetahs (30) (Figure 5.S3). In this study, placing the molecules in a tree using Qemistree reduced the observed technical variation (Figure 5.S3a, c), and highlighted the dietary effect that was expected (Figure 5.S3b, d). These results show how systematic and spurious molecular differences can be mitigated in an unsupervised manner using chemically-informed distance measures based on a tree structure.

As a case study, we used Qemistree to explore chemical diversity in a set of food samples collected as a part of the Global FoodOmics initiative (http://globalfoodomics.org). We selected a diverse range of food ingredients to represent animal, plant, and fungal groupings(31). We first performed feature-based molecular networking using MZmine (17, 18) to obtain spectral library matches for a subset of the chemical features (~20% annotated with cosine cutoff > 0.7). Understanding the chemical relationships between different foods is challenging because most molecules within foods are unannotated. Using Qemistree, we collated GNPS spectral library matches and *in silico* predictions from CSI:FingerID to annotate ~91% of the chemical features (total 634 features after quality filtering) with molecular structures. Using ClassyFire (23), we assigned a chemical taxonomy to 60% of these structures; the remaining 40% returned no

ClassyFire taxonomy. Labeling annotations allowed us to retrieve subtrees of distinct chemical classes (Figure 5.3a) such as flavonoids, alkaloids, phospholipids, acyl-carnitines, and O-glycosyl compounds in food products. We propagated ClassyFire annotations of chemical features (tree tips) to each internal node of the tree and labeled the nodes by pie charts depicting the distribution in chemical superclasses (Figure 5.S4a) and classes (Figure 5.S4b) of its tips. The molecular fingerprint-based hierarchy of chemical features agreed well with ClassyFire taxonomy assignment, further demonstrating that molecular fingerprints can meaningfully capture structural relationships among molecules in a hierarchical manner. Furthermore, Qemistree coupled the chemical tree to sample metadata, revealing distinct chemical classes expected for each sample type. Branches representing acyl-carnitines were exclusively found in animal products (shades of blue; Figure 5.3a). In contrast, honey, although categorized as an animal product, shared most of its chemical space with plant products, reflective of the plant nectar and pollen-based diet of honey bees. We observed a clade of flavonoids in both plant products and honey (Figures 5.3a, S4b), but no other animal-based foods.

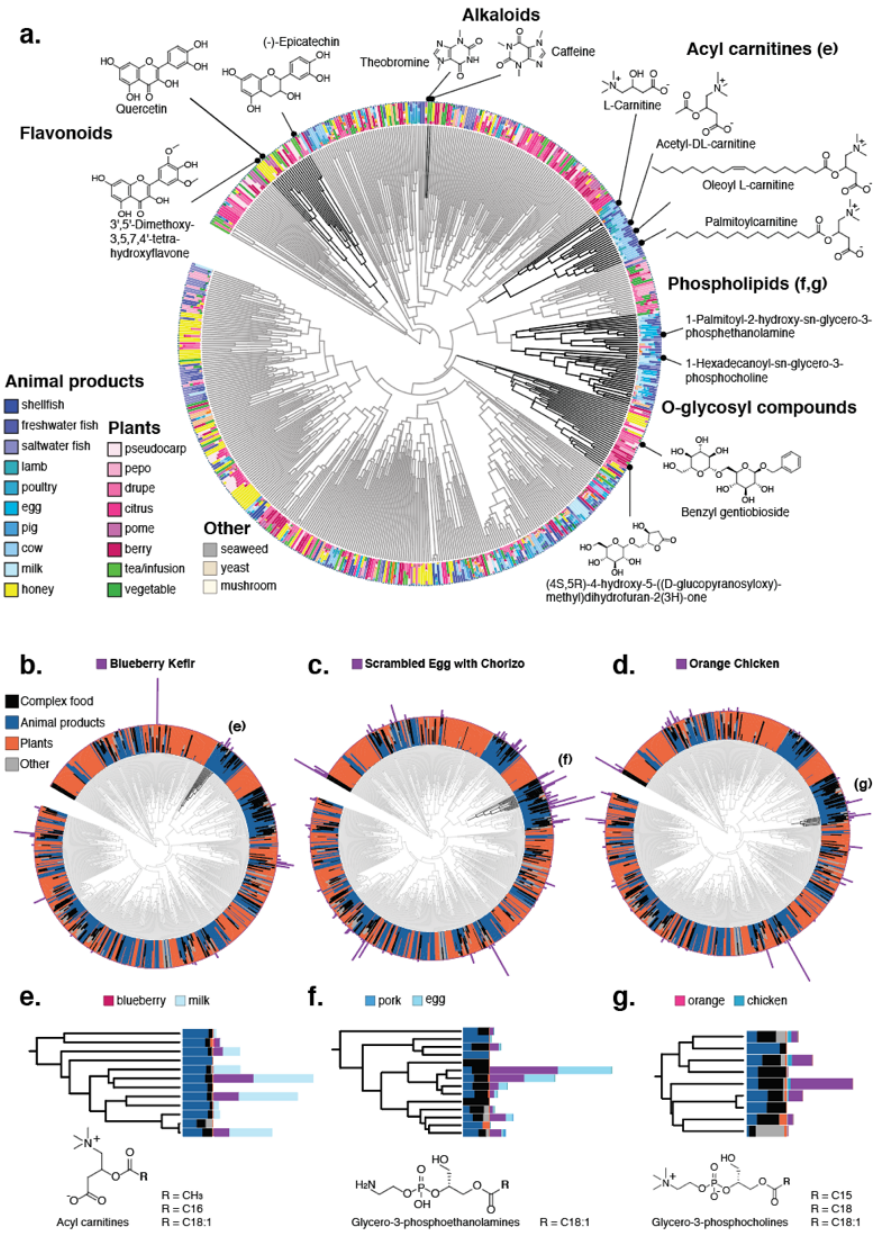

Figure 5.3 *A chemical hierarchy of food-derived compounds based on predicted molecular fingerprints*. **a)** A chemical tree based on molecular fingerprints representing the structural relationships between chemical features (tree tips) detected in food products (single ingredient i.e. simple foods; N=119). The outer ring shows the relative abundance of each compound across a diverse range of food sources. We highlight clusters of compounds that are characteristic of specific food sources. **(b-d)** A hierarchy of the compounds observed in simple foods (above) and seven complex samples: two meals of orange chicken, a cooked cucumber and the sauce from a meal (schmorgurken), sour cream, blueberry kefir, and egg scramble with chorizo (N=126). The inner ring shows the relative abundance of each compound across simple animal products, plant products, fungi and algae (other) and the 7 complex foods (black). The absolute abundances of compounds in blueberry kefir (b), scrambled eggs with chorizo (c), and orange chicken (d) (outer bars) are overlaid on the **(e)** A subtree showing the absolute abundance of acyl carnitines in blueberry kefir and its primary ingredients (blueberry and milk). Similar subtrees showing phosphoethanolamine in scrambled eggs with chorizo **(f)**, and phosphocholine in orange chicken **(g).**

While it is expected that a complex food such as blueberry kefir contains molecules from both blueberries and dairy, we can now visualize how individual ingredients and food preparation contribute to the chemical composition of complex foods. We noted that metabolite signatures that stem directly from particular ingredients, such as phosphoethanolamine from eggs, are present in egg scramble (Figure 5.3c), but not in the other two foods highlighted (Figure 5.3b, d). We can also observe the addition of ingredients in foods that were not listed as present in the initial set of ingredients. We were able to retrieve that there is black pepper in the egg scramble with chorizo and orange chicken, but that this signal is absent from the blueberry kefir (Figure 5.S5).

We show that our tree-based approach coherently captures chemical ontologies and relationships among molecules and samples in various publicly available datasets. Qemistree depends on representing chemical features as molecular fingerprints, and shares limitations with the underlying fingerprint prediction tool CSI:FingerID. For example, fingerprint prediction depends on the quality and coverage of MS/MS spectral databases available for training the predictive models, and these will improve as databases are enriched with more compound classes. Qemistree is also applicable in negative ionization mode; however, less molecular fingerprints can be confidently predicted due to less publicly available reference spectra, resulting in less extensive trees.

In summary, we introduce a new tree-based approach for computing and representing chemical features detected in untargeted metabolomics studies. A hierarchy enables us to leverage existing tree-based tools, and can be augmented with structural and environmental annotations, greatly facilitating analysis and interpretation. We anticipate that Qemistree, as a data organization strategy, will be broadly applicable across fields that perform global chemical analysis, from

medicine to environmental microbiology to food science, and well beyond the examples shown here.

5.3 Materials and methods

Qemistree algorithm

The Qemistree workflow uses MS1-based feature tables and MS1, MS2 fragment ion information (MGF file format) as inputs (Figure 5.S1). These inputs can be generated by processing untargeted mass spectrometry data using MZmine (17) following the Feature-Based Molecular Networking method (18) (example batch file that can be used to perform feature detection and generate the inputs for Qemistree can be found here: MSV000085226). The files exported from MZmine with the *Export/Submit to GNPS* and *SIRIUS Export* module, and are then imported into QIIME2 (25) as the following semantic types: FeatureTable[Frequency] (for the feature table) and MassSpectrometryFeatures (for the ion information).

PREPROCESSING:

Use mzXML files from the instrument

Perform feature detection using MZMine2

    Export sirius MGF and feature table (row m/z, row ID, feature area under the curve per sample)

Convert the feature table to FeatureTable[Frequency] for QIIME2

Create a FeatureData[Molecules] file for QIIME2 using ‘row ID’ and ‘row m/z’

Import the MGF file as MassSpectrometryFeatures for QIIME2

We use SIRIUS (version 4.0.1), ZODIAC (34) and CSI:FingerID to predict molecular substructures within mass spectrometry features in the MGF files imported as MassSpectrometryFeatures. SIRIUS computes fragmentation trees for each molecular formula

candidate of a feature (using PubChem database by default) and ranks these by score. SIRIUS uses MS1 spectrum in the MGF file to determine the candidate ion adduct(s) to be used for the fragmentation tree computation of each feature. ZODIAC takes the top SIRIUS candidates as input and re-ranks molecular formula candidates considering reciprocal compound similarities in the dataset to increase correct molecular formula assignments. Subsequently, CSI:FingerID predicts molecular fingerprints for each feature based on the molecular formula with the highest ZODIAC score.

Note that all spectra provided to the Qemistree pipeline do not necessarily produce a fingerprint. Indeed, SIRIUS does not compute fragmentation trees for multiply charged compounds and CSI:FingerID does not predict molecular fingerprints from spectra with less than 3 explained peaks. To ensure that high confidence molecular formulas are used in Qemistree, we only consider compounds with a ZODIAC score above 0.98 (described in ref. 34).

SUBSTRUCTURE PREDICTION:

For each feature with MS2 spectra in the MGF file:

Compute fragmentation trees (using Sirius)

Re-rank molecular formula candidates on the complete dataset (using Zodiac)

Predict fingerprints based on best molecular formula assignment (using CSI:FingerID)

A dataset M (i.e. a set of exports from MZmine) is a matrix of size n rows by l columns. Each row represents a molecule (m_1, m_2, … m_n), and each column represents a molecular substructure feature. As such, each molecule m_i is composed of a vector (with length l) of predicted probability values (one for each SIRIUS-generated molecular substructure). We remove from our analyses the features without a corresponding vector m_i. In our tests, we have observed that for each dataset 10-15% of the input features are discarded.

For indexing purposes, we relabel each molecule m_i with the MD5-checksum of the predicted fingerprint vector. The motivation to apply the MD5 hashing function is to assign a unique identifier to each feature, which is particularly useful when comparing datasets independently processed using Mzmine. If two distinct molecules (i, j) have identical checksums i.e. $md5(m_i) = md5(m_j)$, then we aggregate those two vectors such that all rows in M are unique. This operation is also propagated down to the table of molecular intensities, in that context intensities are added together.

To co-analyze multiple datasets $M_1, M_2, \ldots M_k$, we combine the matrices into a new dataset M^*. For any two repeated molecules m_i and m_j in M^* we merge their intensities and values as described before. Lastly, we create a hierarchy of chemical relationships T using a distance matrix D measuring the distance between all pairs of molecules in M^*. For qualitative substructure comparisons, we use the Jaccard distance metric and a threshold of 0.5. Otherwise, we use the Euclidean distance with the original probability vectors. With D, we cluster the molecules in a hierarchical fashion using the unweighted pair group method with arithmetic mean (UPGMA). The tips in the resulting tree T have a one-to-one correspondence with all the molecules m_i in M^*.

```
# HIERARCHY CREATION (meta-analysis)
For each fingerprint, feature table in DATASETS:
        Collate fingerprints into a matrix of features by fingerprints
        Match the tuple to have the exact same features and same order
Merge all the fingerprints and feature tables
(use MD5 hash  of fingerprint vectors to merge identical fingerprints)
Compute a distance matrix between features using fingerprints (quantitatively or qualitatively)
Build a hierarchical tree based on the distance matrix
```

Evaluation dataset

Sample preparation and extraction: Four samples were used in the gradient benchmarking dataset: 1) the "*serum*" sample consists of the NIST SRM 1950 reference sample made of human serum spiked with compounds (35) 2) Two human fecal samples from the American Gut Project(36) obtained from a single male individual with a 35 days interval (Sample *fecal-1* " 11-10-2013, and *fecal-2 :* 12-14-2013), and 3) the "*tomato*" seedling sample (*Solanum lycopersicum* plant) was prepared using 3 weeks post-germination specimens (fresh whole seedlings were used). The NIST SRM 1950 sample (1mL), two fecal samples (210 mg of fresh material each), and the tomato seedlings (800 mg of fresh material) were dissolved in 1 mL of 7/3 methanol/water in a 1 mL polypropylene round-bottom tube (QIAGEN), and homogenized in a tissue-lyser (Tissue Lyser II, QIAGEN) at 25 Hz for 5 min. The tubes were then centrifuged at 15,000 rpm for 15 min, and 600 μL of the supernatant was collected and loaded on solid-phase extraction cartridges (Oasis HLB, Waters) made of hydrophilic-lipophilic balance stationary phase (30 mg and 30 μm particle size), that were first activated with 100% methanol, and 100% water (1mL each). After loading the supernatants on the cartridges, washing elution was carried out with 95/5 methanol/water (1 mL), and the samples were eluted with 7/3 methanol/water (2mL), followed by 100% methanol (1mL). The samples were dried down with a vacuum concentrator (Centrivap, Labconco) and resuspended in 2.5 mL of 7/3 methanol/water containing 0.5 μM of amitriptyline as an internal standard. Samples were prepared by mixing the four different samples in various proportions. The resulting extracts were analyzed by mass spectrometry but also used to prepare mixtures of these samples in different ratios. For example, the *serum* and *tomato samples* were mixed in the following ratios: 100/0, 75/25, 50/50, 25/75, 0/100.

Liquid chromatography and mass spectrometry experiments: Samples were analyzed using ultra high performance liquid chromatography (Vanquish, Thermo Scientific) coupled to a quadrupole-Orbitrap mass spectrometer (Q Exactive, Thermo Scientific). The quadrupole-Orbitrap mass spectrometer (Q Exactive, Thermo Scientific) was fitted with an electrospray source (HESI-II) operating in positive ionisation mode. The source used the following parameters: spray voltage, +3500 V; heater temperature, 437.5°C; capillary temperature, 268.75°C; S-lens RF, 50 (arb. units); sheath gas flow rate, 52.5 (arb. units); and auxiliary gas flow rate, 13.75 (arb. units). The samples were acquired in non-targeted MS^2 acquisition mode, with up to four MS^2 scans of the most abundant ions per MS1 scan. The spectra were recorded from 0.48 to 17 min. The following parameters were used for full MS scan: resolution (35,000), Automatic Gain Control target (1.0×10^6), maximum injection time (125 ms), scan range (150-1500 m/z). For the data-dependent in MS^2, the following parameters were used: resolution (17,500), AGC target (2.5×10^5), maximum injection time (125 ms), loop count (4), isolation window (1.5 m/z) fixed first mass (70 m/z). (70-1500 m/z) and up to four MS/MS scans of the most abundant ions per duty cycle. Higher-energy collision induced dissociation was performed with a normalized collision energy of 30 (20, 35, 50). The data-dependent settings were set as follows: minimum AGC (1.25×10^4 [intensity threshold 1.0×10^5]), apex trigger 3 to 15 s, charge exclusion 3-8 and > 8, exclude isotopes (on), dynamic exclusion (14.0 s).

Mass spectrometry data processing: Thermo mass spectrometry data (.RAW) were converted to *m/z* extensible markup language (mzML) (37) in centroid mode using MSConvert ProteoWizard (38) (release 201812). The mzML files were processed with MZmine toolbox (17) (version 2.38) on Ubuntu 18.04 LTS 64-bits workstation (intel Xeon 5E-2637, 3.5 GHz, 8 cores, 64 Go of RAM) following the Feature-Based Molecular Networking method (18).

Global FoodOmics dataset

Sample preparation and extraction: Samples were collected, extracted, and MS data were acquired as a part of the Global FoodOmics project according to the sampling and data acquisition protocols described in Gauglitz et al., 2020 Food Chemistry. Briefly, 126 food samples were selected from the Global FoodOmics dataset. 119 simple food samples (simple in contrast to complex and defined as a single-ingredient food) were selected to cover a broad spectrum of fruits, vegetables, meat and fungi. Each food was represented in at least triplicate in the data subset. Additionally 7 complex samples were selected that contained simple foods from the simple food subset in their ingredient lists. The complex foods were from two separate meals of orange chicken, a cooked cucumber and the sauce from a meal (schmorgurken; in a tomato and sour cream sauce), sour cream, blueberry kefir, and egg scramble with chorizo. Sample metadata describes the food samples based on a food hierarchy beginning with plant vs. animal vs. fungus (sample_type_group1) and increasing in detail down to persian cucumber vs. cherry tomato etc. (sample_type_group6).

Briefly, samples were extracted in 95% LC-MS grade Ethanol; 5% LC-MS grade water. Samples were analyzed using the same LC-MS/MS setup and software as described above for the maXis II QTOF mass spectrometer (Bruker Daltonics), using a Phenomenex Kinetex C18 1.7 μm (100A) 100 x 2.1 column equipped with a guard cartridge (Phenomenex). The instrument tuning and internal calibrant remained the same as described above. MS spectra were acquired in a positive ion mode in the range m/z 50–1,500. The mobile phases consisted of A (100% water + 0.1% formic acid) and B (100% acetonitrile + 0.1% formic acid), and the flow rate was set to 0.5 μL/min throughout the experiment, and the column maintained at 40°C.

Mass spectrometry data processing: The mass spectrometry data (.d) were converted to .mzXML with lock mass calibration applied using CompassXport batch mode in Data Analysis 4.4 software (Bruker Daltonics, Bremen, Germany) running on a Windows 10 PC. The mass spectrometry data was processed with MZmine toolbox (17) (version 2.38) using the parameters outlined in an XML batch file (see Data availability).

Multivariate comparisons: To evaluate the benefits of using a tree for multivariate analysis, we generated pairwise sample distances using Bray-Curtis (29) (tree-agnostic) and Weighted UniFrac(9) (tree-informed). Both of these metrics compare samples quantitatively i.e. based on the abundances of each feature. Notably, UniFrac weights the distances based on the shared branches of the tree used for computation.

5.4 Acknowledgments

PCD was supported by the Gordon and Betty Moore Foundation (GBMF7622), the U.S. National Institutes of Health for the Center (P41 GM103484, R03 CA211211, R01 GM107550), and the University of Wisconsin-Madison OVCRGE; LFN was supported by the U.S. National Institutes of Health (R01 GM107550), and the European Union's Horizon 2020 program (MSCA-GF, 704786). JJJvdH was supported by an ASDI eScience grant, ASDI.2017.030, from the Netherlands eScience Center—NLeSC. KD, MF, ML and SB were supported by Deutsche Forschungsgemeinschaft (BO 1910/20).

5.5 Author contributions

Chapter 5, in full, is a preprint of: Tripathi, A., Vázquez-Baeza, Y., Gauglitz, J. M., Wang, M., Dührkop, K., Nothias-Esposito, M., Acharya, D. D., Ernst, M., van der Hooft, J. J. J., Zhu, Q.,

McDonald, D., Gonzalez, A., Handelsman, J., Fleischauer, M., Ludwig, M., Böcker, S., Nothias, L.-F., Knight, R., & Dorrestein, P. C. (n.d.). *Chemically-informed Analyses of Metabolomics Mass Spectrometry Data with Qemistree*.

AT conceived the concept and managed the project. AT and YVB developed the algorithm and wrote the code for Qemistree. AT and YVB contributed equally to the work. LFN, RK, PCD supervised method implementation. KD, MW, JJJvdH, ME, DM, and AG tested and provided suggestions on how to improve the method. MW managed the deployment of Qemistree on GNPS. AT and MW developed the GNPS-Qemistree Dashboard. DA and AT wrote the documentation for the GNPS-Qemistree workflow. YVB, QZ, and AT developed Qemistree-iTOL visualization. LFN and MNE performed the mass-spectrometry for the evaluation dataset. AT, YVB, and LFN analyzed and interpreted the evaluation data. JMG performed mass spectrometry of the Global Foodomics samples. AT, JMG analyzed and interpreted the Global Foodomics data. KD, MF, ML, and SB supported the integration of SIRIUS, Zodiac, and CSI:FingerID. AT, YVB, PCD, and RK wrote the manuscript. LFN, JMG, MNE, JJJvdH, ME, KD, QZ, DM, AG, JH, MF, ML, SB, and RK improved the manuscript. The co-authors listed above supervised or provided support for the research and have given permission for the inclusion of the work in this dissertation.

5.6 Competing interests

Mingxun Wang is a founder of Ometa Labs LLC. Pieter C. Dorrestein is a scientific advisor for Sirenas LLC. Kai Dührkop, Marcus Ludwig, Markus Fleischauer and Sebastian Böcker are founders of Bright Giant GmbH.

5.7 Data and code availability

The mass spectrometry data, metadata, and methods for the evaluation dataset have been deposited on the GNPS/MassIVE public repository (2, 33) under the accession number MSV000083306. The parameters used for molecular networking are available on GNPS: https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=efda476c72724b29a91693a108fa5a9d. The chemical hierarchy generated by Qemistree (version 2020.1.2) is available on iTOL(24): https://itol.embl.de/tree/709513416494381587432576. The mass spectrometry data, metadata, and methods for Global Foodomics dataset have been deposited on the GNPS/MassIVE public repository (2, 33) under the accession number MSV000085226. The parameters used for molecular networking are available on GNPS: https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=ceb28a199d6b4f4fbf08490d9c96d631. The chemical hierarchy generated by Qemistree (version 2020.1.2) is available on iTOL(24): https://itol.embl.de/tree/13711034118313741584046018. All source code is publicly available under BSD-2-Clause on GitHub: https://github.com/biocore/q2-qemistree. Qemistree is also available as an advanced analysis workflow on GNPS: https://ccms-ucsd.github.io/GNPSDocumentation/qemistree/

5.8 Supplemental figures

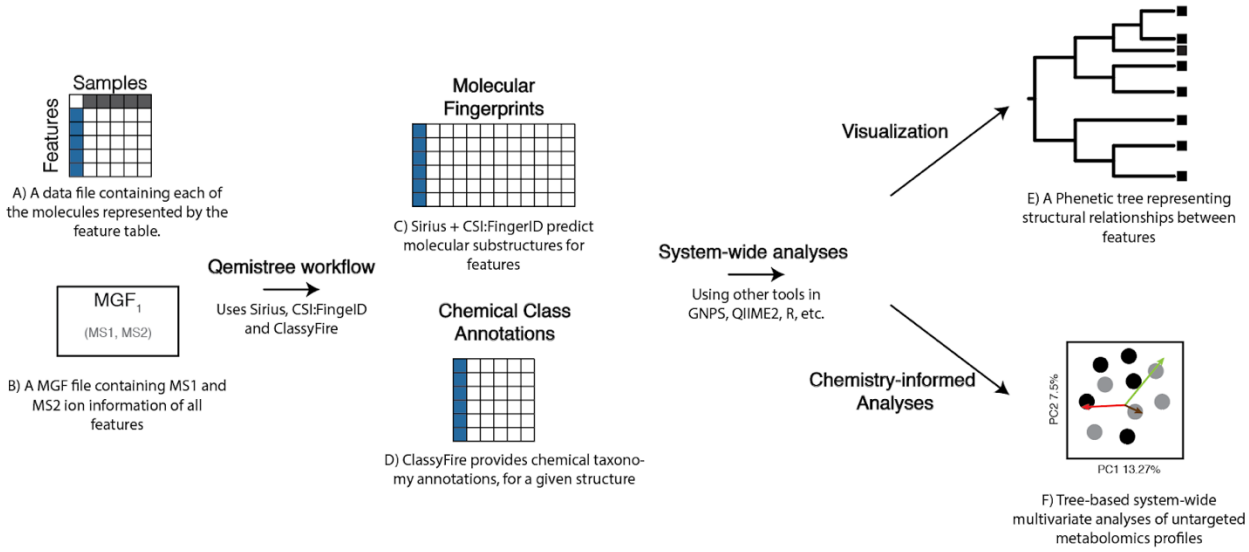

Figure 5.S1 *End-to-end Qemistree analysis using GNPS and QIIME2***.** Qemistree analysis can be performed using two required input files: 1) A table of molecule (or chemical feature) abundances per sample and 2) an MGF file with MS1 and MS2 ion information. These inputs can be generated by processing mass spectrometry files (.mzXML) through MZmine for feature detection. In Qemistree, these input files are processed through SIRIUS and CSI:FingerID to generate molecular fingerprints and *in silico* structural annotations (SMILES) per MS feature. We use the predicted molecular fingerprints to generate a phenetic tree of relationships between MS features based on sub-structural similarity. This tree can be visualized in iTOL for further data exploration. If the user inputs a sample metadata file, they can also visualize the abundances of each MS feature stratified by sample grouping of interest. Additionally, the qemistree queries ClassyFire to classify the structural annotations into chemical 'kingdom', 'superclass', 'class', 'subclass' and 'direct parent'. We further allow the users to input a file with MS/MS spectral library matches (optional) into the workflow such that these library matches (typically, 2-20% of all MS features), instead of *in silico* annotation, are used for ClassyFire queries whenever available. All the outputs of the qemistree workflow can be analyzed further using QIIME 2 tools (such as tree-based alpha and beta diversity, mmvec (26), songbird (32)) or explored in Python, R etc. as needed.

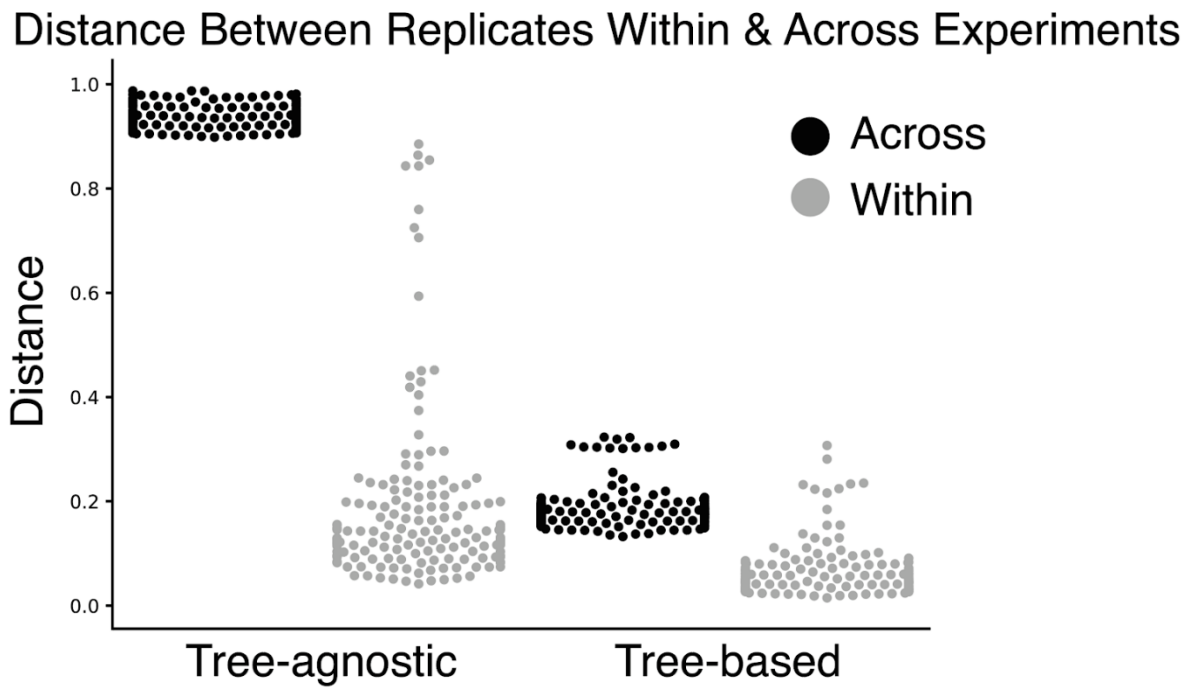

Figure 5.S2 *Qemistree reduces the differences between biological replicates across mass-spectrometry runs*. A comparison of distances between sample replicates within and across chromatography gradients when using tree-agnostic (Bray-Curtis) distances and tree-based (Weighted UniFrac) distances.

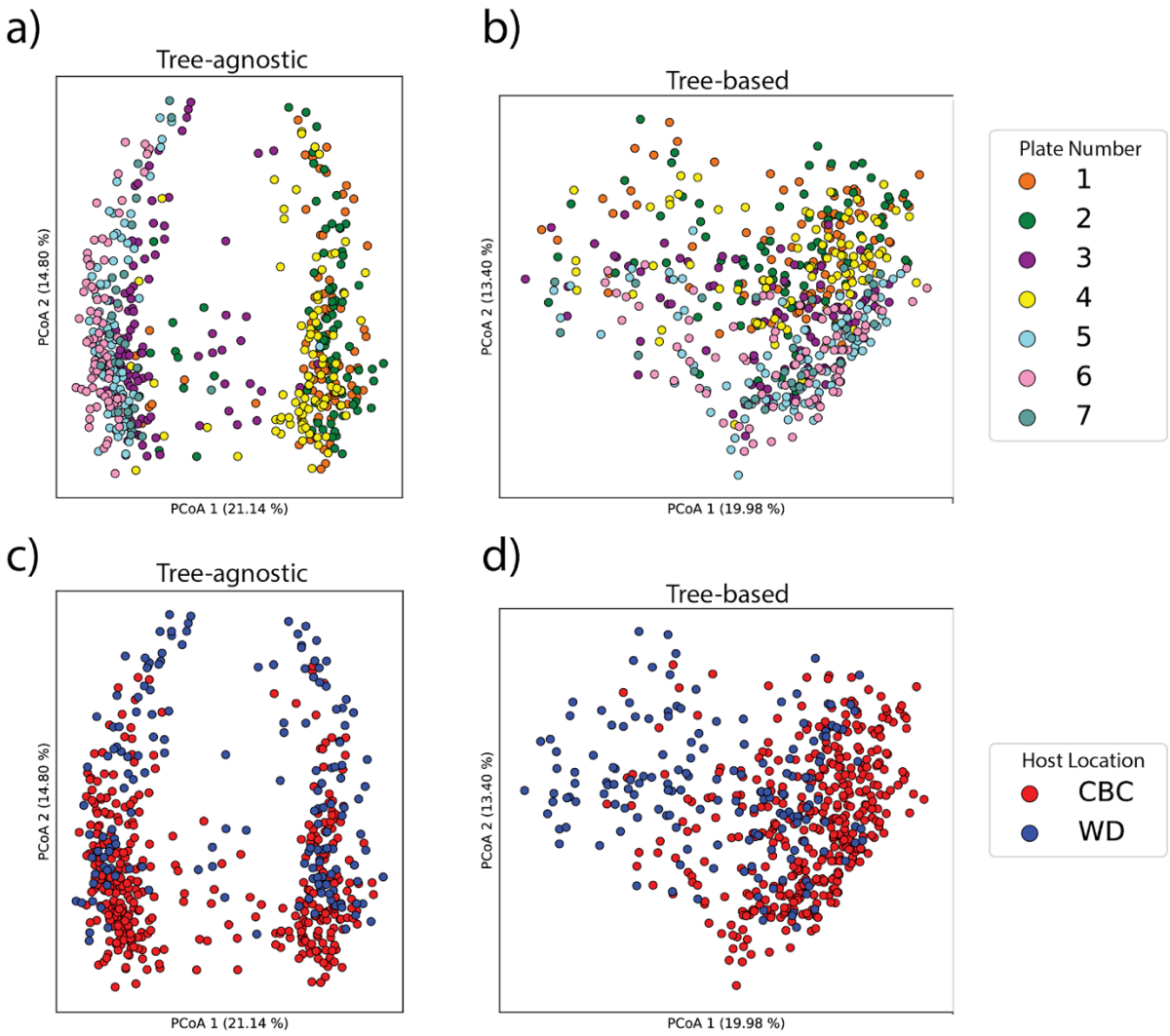

Figure 5.S3 *Qemistree mitigates plate-to-plate variation in fecal metabolomics study to highlight a biologically-relevant effect*. **a)** Principal coordinate analysis (PCoA) of tree-agnostic distances (Bray-Curtis) colored by plate number (pseudo-F=32.39, p=0.001). **b)** PCoA of tree-informed distances (Weighted UniFrac) colored by plate number (pseudo-F=15.67, p=0.001). The same PCoA of **(c)** Bray-Curtis distances (pseudo-F=33.50, p=0.001) and (d) Weighted UniFrac distances (pseudo-F=48.42, p=0.001) colored by cheetah location which governed the diet of cheetahs. Data is available on the GNPS/MassIVE public repository (2, 33) accession number MSV000082969. CBC: Cheetah Breeding Center; WD: Wildlife Discoveries

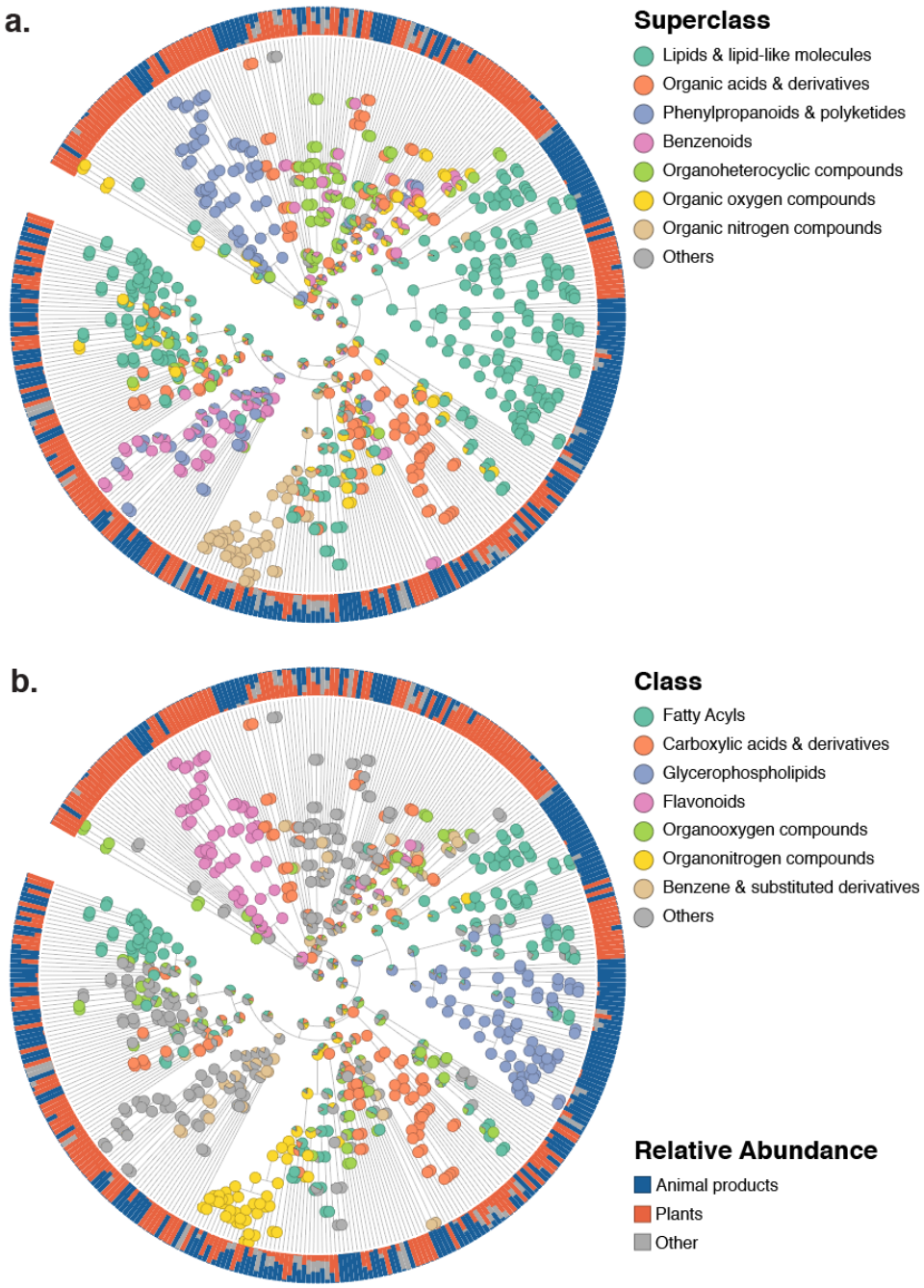

Figure 5.S4 *Qemistree highlights chemical taxonomy of food-derived compounds*. Chemical hierarchy of compounds (tree tips) detected in simple food products (single ingredient foods, N=119). The tree is pruned to only tips that were assigned a chemical class using ClassyFire. Internal nodes are labeled by pie charts of the superclass **(a)** and class **(b)** level taxonomy of children tips. For instance, if a node has 10 children tips and 8 of them are assigned to class A and 2 to class B, then the pie chart will have two colors, with the angle being 8:2. Outer ring shows the relative abundance of each compound across simple animal products, plant products, and other (fungi and algae). The chemical hierarchy can be further explored using the following iTOL (24) link: https://itol.embl.de/tree/7095134164128581587333337. For example, we observed two clades of chemical features classified as fatty acyls (b) such that the fatty acyl features found primarily in animal products are acyl carnitines (b; right) and the ones found in both plant and animal products are derivatives of linoleic acid (b; left).

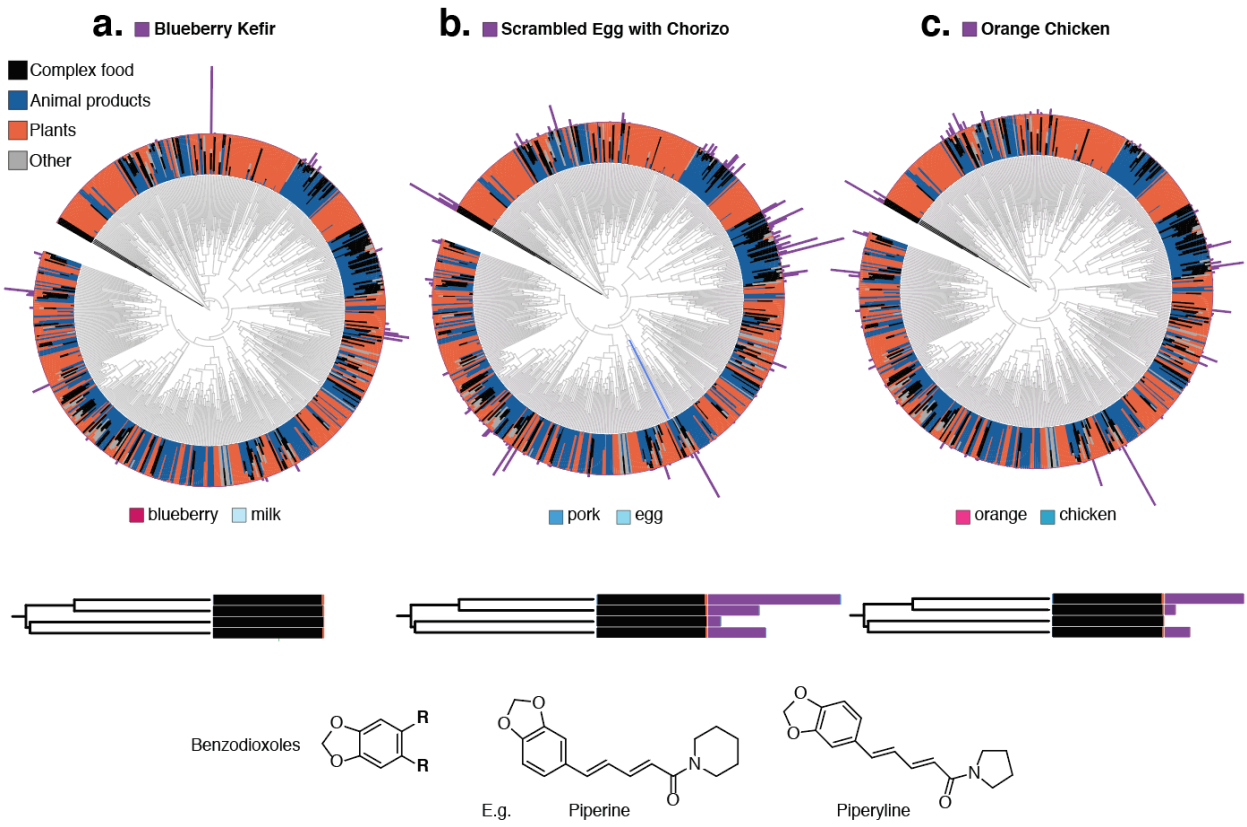

Figure 5.S5 *Chemical hierarchy of the compounds observed in simple foods and seven complex samples*. **a,b,c)** 2 meals of orange chicken, a cooked cucumber and the sauce from a meal (schmorgurken), sour cream, blueberry kefir, and egg scramble with chorizo (N=126). Analogous to Figure 5.3b-d, the inner ring shows the relative abundance of each compound across simple animal products, plant products, fungi and algae (other) and complex foods. The absolute abundances of compounds in blueberry kefir **(a)**, scrambled eggs with chorizo **(b)**, and orange chicken **(c)** (outer bars) are overlaid on the tree to illustrate the shared and unique chemistry of complex foods. We highlight a classifier subtree annotated as benzodioxoles, compounds found in black pepper (in black) that are almost exclusively detected in complex foods. We overlay the absolute abundance of benzodioxoles in complex foods and their primary ingredients. These alkaloids are detected in scrambled eggs with chorizo (b) and orange chicken (c) but not in blueberry kefir (a) or the primary ingredients of these complex foods. This indicates that they are added during cooking, a likely assumption given the prevalence of black pepper in the western diet. The presence in an egg dish and meat dish coupled with the lack of signal in blueberry kefir also corresponds with the traditional use of this spice.

5.9 References

1. Watrous J, Roach P, Alexandrov T, Heath BS, Yang JY, Kersten RD, van der Voort M, Pogliano K, Gross H, Raaijmakers JM, Moore BS, Laskin J, Bandeira N, Dorrestein PC. 2012. Mass spectral molecular networking of living microbial colonies. Proc Natl Acad Sci U S A 109:E1743–52.

2. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T, Porto C, Bouslimani A, Melnik AV, Meehan MJ, Liu W-T, Crüsemann M, Boudreau PD, Esquenazi E, Sandoval-Calderón M, Kersten RD, Pace LA, Quinn RA, Duncan KR, Hsu C-C, Floros DJ, Gavilan RG, Kleigrewe K, Northen T, Dutton RJ, Parrot D, Carlson EE, Aigle B, Michelsen CF, Jelsbak L, Sohlenkamp C, Pevzner P, Edlund A, McLean J, Piel J, Murphy BT, Gerwick L, Liaw C-C, Yang Y-L, Humpf H-U, Maansson M, Keyzers RA, Sims AC, Johnson AR, Sidebottom AM, Sedio BE, Klitgaard A, Larson CB, P CAB, Torres-Mendoza D, Gonzalez DJ, Silva DB, Marques LM, Demarque DP, Pociute E, O'Neill EC, Briand E, Helfrich EJN, Granatosky EA, Glukhov E, Ryffel F, Houson H, Mohimani H, Kharbush JJ, Zeng Y, Vorholt JA, Kurita KL, Charusanti P, McPhail KL, Nielsen KF, Vuong L, Elfeki M, Traxler MF, Engene N, Koyama N, Vining OB, Baric R, Silva RR, Mascuch SJ, Tomasi S, Jenkins S, Macherla V, Hoffman T, Agarwal V, Williams PG, Dai J, Neupane R, Gurr J, Rodríguez AMC, Lamsa A, Zhang C, Dorrestein K, Duggan BM, Almaliti J, Allard P-M, Phapale P, Nothias L-F, Alexandrov T, Litaudon M, Wolfender J-L, Kyle JE, Metz TO, Peryea T, Nguyen D-T, VanLeer D, Shinn P, Jadhav A, Müller R, Waters KM, Shi W, Liu X, Zhang L, Knight R, Jensen PR, Palsson BO, Pogliano K, Linington RG, Gutiérrez M, Lopes NP, Gerwick WH, Moore BS, Dorrestein PC, Bandeira N. 2016. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol 34:828–837.

3. Fox Ramos AE, Evanno L, Poupon E, Champy P, Beniddir MA. 2019. Natural products targeting strategies involving molecular networking: different manners, one goal. Nat Prod Rep 36:960–980.

4. Böcker S, Dührkop K. 2016. Fragmentation trees reloaded. J Cheminform 8:5.

5. Rasche F, Scheubert K, Hufsky F, Zichner T, Kai M, Svatoš A, Böcker S. 2012. Identifying the unknowns by aligning fragmentation trees. Anal Chem 84:3417–3426.

6. Washburne AD, Silverman JD, Leff JW, Bennett DJ, Darcy JL, Mukherjee S, Fierer N, David LA. 2017. Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. PeerJ 5:e2969.

7. Faith DP. 1992. Conservation evaluation and phylogenetic diversity. Biological Conservation.

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

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

10. Willett P. 2006. Similarity-based virtual screening using 2D fingerprints. Drug Discov Today 11:1046–1053.

11. Heinonen M, Shen H, Zamboni N, Rousu J. 2012. Metabolite identification and molecular fingerprint prediction through machine learning. Bioinformatics 28:2333–2341.

12. Laponogov I, Sadawi N, Galea D, Mirnezami R, Veselkov KA. 2018. ChemDistiller: an engine for metabolite annotation in mass spectrometry. Bioinformatics.

13. Dührkop K, Shen H, Meusel M, Rousu J, Böcker S. 2015. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. Proc Natl Acad Sci U S A 112:12580–12585.

14. Fan Z, Ghaffari K, Alley A, Ressom HW. 2019. Metabolite Identification Using Artificial Neural Network. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM).

15. Li Y, Kuhn M, Gavin A-C, Bork P. 2020. Identification of metabolites from tandem mass spectra with a machine learning approach utilizing structural features. Bioinformatics 36:1213–1218.

16. Dührkop K, Fleischauer M, Ludwig M, Aksenov AA, Melnik AV, Meusel M, Dorrestein PC, Rousu J, Böcker S. 2019. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. Nat Methods 16:299–302.

17. Pluskal T, Castillo S, Villar-Briones A, Oresic M. 2010. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics 11:395.

18. Nothias LF, Petras D, Schmid R, Dührkop K, Rainer J, Sarvepalli A, Protsyuk I, Ernst M, Tsugawa H, Fleischauer M, Aicheler F, Aksenov A, Alka O, Allard P-M, Barsch A, Cachet X, Caraballo M, Da Silva RR, Dang T, Garg N, Gauglitz JM, Gurevich A, Isaac G, Jarmusch AK, Kameník Z, Kang KB, Kessler N, Koester I, Korf A, Le Gouellec A, Ludwig M, Christian MH, McCall L-I, McSayles J, Meyer SW, Mohimani H, Morsy M, Moyne O, Neumann S, Neuweger H, Nguyen NH, Nothias-Esposito M, Paolini J, Phelan VV, Pluskal T, Quinn RA, Rogers S, Shrestha B, Tripathi A, van der Hooft JJJ, Vargas F, Weldon KC, Witting M, Yang H, Zhang Z, Zubeil F, Kohlbacher O, Böcker S, Alexandrov T, Bandeira

N, Wang M, Dorrestein PC. 2019. Feature-based Molecular Networking in the GNPS Analysis Environment. bioRxiv.

19. Treutler H, Tsugawa H, Porzel A, Gorzolka K, Tissier A, Neumann S, Balcke GU. 2016. Discovering Regulated Metabolite Families in Untargeted Metabolomics Studies. Anal Chem 88:8082–8090.

20. Depke T, Franke R, Brönstrup M. 2017. Clustering of MS2 spectra using unsupervised methods to aid the identification of secondary metabolites from Pseudomonas aeruginosa. Journal of Chromatography B.

21. Rawlinson C, Jones D, Rakshit S, Meka S, Moffat CS, Moolhuijzen P. 2020. Hierarchical clustering of MS/MS spectra from the firefly metabolome identifies new lucibufagin compounds. Sci Rep 10:6043.

22. Schymanski EL, Ruttkies C, Krauss M, Brouard C, Kind T, Dührkop K, Allen F, Vaniya A, Verdegem D, Böcker S, Rousu J, Shen H, Tsugawa H, Sajed T, Fiehn O, Ghesquière B, Neumann S. 2017. Critical Assessment of Small Molecule Identification 2016: automated methods. J Cheminform 9:22.

23. Feunang YD, Eisner R, Knox C, Chepelev L, Hastings J, Owen G, Fahy E, Steinbeck C, Subramanian S, Bolton E, Greiner R, Wishart DS. 2016. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. J Cheminform 8:1–20.

24. Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res 47:W256–W259.

25. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu Y-X, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS 2nd, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37:852–857.

26. Morton JT, Aksenov AA, Nothias LF, Foulds JR, Quinn RA, Badri MH, Swenson TL, Van Goethem MW, Northen TR, Vazquez-Baeza Y, Wang M, Bokulich NA, Watters A, Song SJ, Bonneau R, Dorrestein PC, Knight R. 2019. Learning representations of microbe-metabolite interactions. Nat Methods 16:1306–1314.

27. Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, -M. Fan TW, Fiehn O, Goodacre R, Griffin JL, Hankemeier T, Hardy N, Harnly J, Higashi R, Kopka J, Lane AN, Lindon JC, Marriott P, Nicholls AW, Reily MD, Thaden JJ, Viant MR. 2007. Proposed minimum reporting standards for chemical analysis. Metabolomics.

28. Sedio BE, Rojas Echeverri JC, Boya P. CA, Joseph Wright S. 2017. Sources of variation in foliar secondary chemistry in a tropical forest tree community. Ecology.

29. Bray JR, Roger Bray J, Curtis JT. 1957. An Ordination of the Upland Forest Communities of Southern Wisconsin. Ecological Monographs.

30. Gauglitz JM, Morton JT, Tripathi A, Hansen S, Gaffney M, Carpenter C, Weldon KC, Shah R, Parampil A, Fidgett A, Swafford AD, Knight R, Dorrestein PC. 2019. Metabolome-informed microbiome analysis refines metadata classifications and reveals unexpected medication transfer in captive cheetahs. bioRxiv.

31. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, Navas-Molina JA, Janssen S, Kopylova E, Vázquez-Baeza Y, González A, Morton JT, Mirarab S, Zech Xu Z, Jiang L, Haroon MF, Kanbar J, Zhu Q, Jin Song S, Kosciolek T, Bokulich NA, Lefler J, Brislawn CJ, Humphrey G, Owens SM, Hampton-Marcell J, Berg-Lyons D, McKenzie V, Fierer N, Fuhrman JA, Clauset A, Stevens RL, Shade A, Pollard KS, Goodwin KD, Jansson JK, Gilbert JA, Knight R, Earth Microbiome Project Consortium. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551:457–463.

32. Morton JT, Marotz C, Washburne A, Silverman J, Zaramela LS, Edlund A, Zengler K, Knight R. 2019. Establishing microbial composition measurement standards with reference frames. Nat Commun 10:2719.

33. Wang M, Wang J, Carver J, Pullman BS, Cha SW, Bandeira N. 2018. Assembling the Community-Scale Discoverable Human Proteome. Cell Syst 7:412–421.e5.

34. Ludwig M, Nothias L-F, Dührkop K, Koester I, Fleischauer M, Hoffmann MA, Petras D, Vargas F, Morsy M, Aluwihare L, Dorrestein PC, Böcker S. 2019. ZODIAC: database-independent molecular formula annotation using Gibbs sampling reveals unknown small molecules. bioRxiv.

35. Simón-Manso Y, Lowenthal MS, Kilpatrick LE, Sampson ML, Telu KH, Rudnick PA, Mallard WG, Bearden DW, Schock TB, Tchekhovskoi DV, Blonder N, Yan X, Liang Y, Zheng Y, Wallace WE, Neta P, Phinney KW, Remaley AT, Stein SE. 2013. Metabolite profiling of a NIST Standard Reference Material for human plasma (SRM 1950): GC-MS,

LC-MS, NMR, and clinical laboratory analyses, libraries, and web-based resources. Anal Chem 85:11725–11731.

36. McDonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, Aksenov AA, Behsaz B, Brennan C, Chen Y, DeRight Goldasich L, Dorrestein PC, Dunn RR, Fahimipour AK, Gaffney J, Gilbert JA, Gogul G, Green JL, Hugenholtz P, Humphrey G, Huttenhower C, Jackson MA, Janssen S, Jeste DV, Jiang L, Kelley ST, Knights D, Kosciolek T, Ladau J, Leach J, Marotz C, Meleshko D, Melnik AV, Metcalf JL, Mohimani H, Montassier E, Navas-Molina J, Nguyen TT, Peddada S, Pevzner P, Pollard KS, Rahnavard G, Robbins-Pianka A, Sangwan N, Shorenstein J, Smarr L, Song SJ, Spector T, Swafford AD, Thackray VG, Thompson LR, Tripathi A, Vázquez-Baeza Y, Vrbanac A, Wischmeyer P, Wolfe E, Zhu Q, American Gut Consortium, Knight R. 2018. American Gut: an Open Platform for Citizen Science Microbiome Research. mSystems 3.

37. Martens L, Chambers M, Sturm M, Kessner D, Levander F, Shofstahl J, Tang WH, Römpp A, Neumann S, Pizarro AD, Montecchi-Palazzi L, Tasman N, Coleman M, Reisinger F, Souda P, Hermjakob H, Binz P-A, Deutsch EW. 2011. mzML--a community standard for mass spectrometry data. Mol Cell Proteomics 10:R110.000133.

38. Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J, Hoff K, Kessner D, Tasman N, Shulman N, Frewen B, Baker TA, Brusniak M-Y, Paulse C, Creasy D, Flashner L, Kani K, Moulding C, Seymour SL, Nuwaysir LM, Lefebvre B, Kuhlmann F, Roark J, Rainer P, Detlev S, Hemenway T, Huhmer A, Langridge J, Connolly B, Chadick T, Holly K, Eckels J, Deutsch EW, Moritz RL, Katz JE, Agus DB, MacCoss M, Tabb DL, Mallick P. 2012. A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol 30:918–920.

Chapter 6. Conclusions

In traditional microbiology studies, it is common to study microbes in isolation by culturing them in standard laboratory conditions. In nature, however, microbes exist in diverse communities dynamically responding to environmental changes. The growing appreciation of the latter led to the emergence of the field of microbiome. Over the past two decades, microbiome research has witnessed rapid analytical, computational and theoretical advancements which extend our blinkered view of 'self' to include our microbiota — the diverse and dynamic microbial communities living on and within us (1). The human microbiome — the genetic pool of the microbiota — accounts for 99% of all the genes that constitute the human superorganism. Our microbiome coevolved with our genome to support us in a plethora of physiological functions (such as nutrient absorption from diet, protection from pathogens etc.) that are still being discovered. Association studies have shown that imbalances in the microbiome composition is linked to a wide range of illnesses such as gastrointestinal disorders, cardiovascular disease, metabolic disease as well as brain and developmental disorders.

In my doctoral research, I pursued an increased understanding of the microbial ecosystem by pairing microbial composition to its functional readout. For this, I studied microbiome data in the context of its metabolome – the hundreds to thousands of small molecules associated with the microbial community. With our starkly altered lifestyles compared to our ancestors, complex metabolic diseases such as Type 2 diabetes, chronic liver disease, heart disease, obstructive pulmonary disease and many forms of cancer are on an all-time rise. Therefore, I focused on understanding how we could reduce this pertinent health burden by leveraging our understanding of the gut ecosystem. Specifically, the primary driving question on my first few research projects was, "*Does the gut ecosystem encode information for metabolic health assessment?*"

In Chapter 2, I studied a cohort of non-alcoholic fatty liver disease (NAFLD) patients and their first-degree relatives i.e twins, parents and siblings. It has been shown that the relatives of NAFLD patients are at a higher risk of liver disease development (2). The most important clinical challenge in this context is to determine optimal strategies for screening NAFLD using accurate, non-invasive, widely available and easy-to-perform screening tests. In this work, we report a proof-of-concept that gut microbiome could be a viable and scalable source of biomarkers to accurately diagnose NAFLD in high-risk populations.

In Chapter 3 and 4, I investigated the impact of chronic obstructive sleep apnea (OSA) on the gut microbiome and metabolome. Nearly 12% of the adult population in the United States has OSA, which poses an annual cost burden of nearly $149.6 billion (3). Strikingly, 90% of the patients remain undiagnosed due to the lack of cheap and widely applicable diagnostic tests. In chapter 3, we used a mouse model system to show that intermittent hypoxia and hypercapnia (IHH), -- a hallmark of obstructive sleep apnea -- leads to marked changes in both the microbial composition and its metabolism. Therefore, stool-based tests could be a viable alternative to expensive overnight sleep studies for OSA diagnosis. To develop robust diagnostic biomarkers, understanding whether disease-associated changes in the microbiome are consistent across animal models of different genetic backgrounds, and hence potentially translatable to human populations is essential. In Chapter 4, we demonstrated the consistency of our previous findings by predicting IHH-exposure across different disease models. We also introduced a pipeline to identify robust microbiome and metabolome features that are most likely to apply in humans.

In the process of working with paired microbiome and metabolome data, I realized that the metabolomic community is in need of effective tools and standardized data analysis pipelines. I dedicated the last two years of my doctoral research towards developing new analytical solutions

for metabolomics data analysis. I developed a software to improve comparative analysis of high-dimensional chemical profiles by adapting the analytical concepts from microbial ecology. Chapter 5 introduces this tool Qemistree (pronounced *chemis-tree*) which organizes the thousands of detected molecular features into a "tree-like" hierarchy. Qemistree enables the application of phylogeny-based metrics, which have been highly advantageous for microbiome data analyses, to study chemical diversity in metabolomics data. For example, using Qemistree, we can calculate Faith's phylogenetic diversity (4) to study chemical diversity within a sample or pairwise UniFrac distances (5) to compare diversity across samples. Qemistree also integrates multiple annotation tools (molecular networking, Sirius, CSI: FingerID, ClassyFire) to boost the annotation of unknown metabolites, and has been proving useful in exploring the chemistry in complex biospecimens. Qemistree is an open-source software which is available to both the microbiome and metabolome community as a QIIME2 (6) plugin and a GNPS (7) workflow with the hope that it foments collaborative progress in the two omics fields.

Biomedical research is largely a data-driven endeavor. With advancements in high-throughput technologies like next-generation sequencing and high-resolution mass-spectrometry, researchers are able to collect vast and complex data layers such as genomics, proteomics, transcriptomics, and metabolomics collectively known as omics data. Omics data are characterized by many shared challenges such as high dimensionality and high analytical variance which can be tackled using the same analytical solutions. For example, operations such as data normalization and scaling (to apply statistical models), projecting samples in lower dimensions (to visualize overall trends in how samples relate to one another) and testing for the differential abundance of analytes (to find biomarkers of health and disease), are routinely applied across all omics datasets. However, omics data layers are typically analyzed using tools developed independently within

each discipline. Tools like Qemistree highlight the value of mapping existing analytical solutions across omics domains to advance multi-omics data integration.

Going forward, I want to expand my understanding of other omics domains, and develop tools that enable biomedical researchers to reproducibly analyze multiple-omics datasets to improve our understanding of complex biological machineries. My doctoral training has prepared me for the next steps — I am thrilled to act as a catalyst for discoveries in biomedical research.

References

1. Davies J. 2001. In a Map for Human Life, Count the Microbes, Too. Science.

2. Caussy C, Soni M, Cui J, Bettencourt R, Schork N, Chen C-H, Ikhwan MA, Bassirian S, Cepin S, Gonzalez MP, Mendler M, Kono Y, Vodkin I, Mekeel K, Haldorson J, Hemming A, Andrews B, Salotti J, Richards L, Brenner DA, Sirlin CB, Loomba R, Familial NAFLD Cirrhosis Research Consortium. 2017. Nonalcoholic fatty liver disease with cirrhosis increases familial risk for advanced fibrosis. J Clin Invest 127:2697–2704.

3. Watson NF. 2016. Health Care Savings: The Economic Value of Diagnostic and Therapeutic Care for Obstructive Sleep Apnea. J Clin Sleep Med 12:1075–1077.

4. Faith DP. 2016. The PD Phylogenetic Diversity Framework: Linking Evolutionary History to Feature Diversity for Biodiversity Conservation. Biodiversity Conservation and Phylogenetic Systematics.

5. Lozupone CA, Knight R. 2015. The UniFrac significance test is sensitive to tree topology. BMC Bioinformatics 16:211.

6. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu Y-X, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF,

Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS 2nd, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG. 2019. Author Correction: Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37:1091.

7. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T, Porto C, Bouslimani A, Melnik AV, Meehan MJ, Liu W-T, Crüsemann M, Boudreau PD, Esquenazi E, Sandoval-Calderón M, Kersten RD, Pace LA, Quinn RA, Duncan KR, Hsu C-C, Floros DJ, Gavilan RG, Kleigrewe K, Northen T, Dutton RJ, Parrot D, Carlson EE, Aigle B, Michelsen CF, Jelsbak L, Sohlenkamp C, Pevzner P, Edlund A, McLean J, Piel J, Murphy BT, Gerwick L, Liaw C-C, Yang Y-L, Humpf H-U, Maansson M, Keyzers RA, Sims AC, Johnson AR, Sidebottom AM, Sedio BE, Klitgaard A, Larson CB, P CAB, Torres-Mendoza D, Gonzalez DJ, Silva DB, Marques LM, Demarque DP, Pociute E, O'Neill EC, Briand E, Helfrich EJN, Granatosky EA, Glukhov E, Ryffel F, Houson H, Mohimani H, Kharbush JJ, Zeng Y, Vorholt JA, Kurita KL, Charusanti P, McPhail KL, Nielsen KF, Vuong L, Elfeki M, Traxler MF, Engene N, Koyama N, Vining OB, Baric R, Silva RR, Mascuch SJ, Tomasi S, Jenkins S, Macherla V, Hoffman T, Agarwal V, Williams PG, Dai J, Neupane R, Gurr J, Rodríguez AMC, Lamsa A, Zhang C, Dorrestein K, Duggan BM, Almaliti J, Allard P-M, Phapale P, Nothias L-F, Alexandrov T, Litaudon M, Wolfender J-L, Kyle JE, Metz TO, Peryea T, Nguyen D-T, VanLeer D, Shinn P, Jadhav A, Müller R, Waters KM, Shi W, Liu X, Zhang L, Knight R, Jensen PR, Palsson BO, Pogliano K, Linington RG, Gutiérrez M, Lopes NP, Gerwick WH, Moore BS, Dorrestein PC, Bandeira N. 2016. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol 34:828–837.