2021

Influence of Geography on the Healthy Gut Microbiome and the Role of the Gut Microbiome in IBD Symptom and Disease Progression

Sayf Al-Deen Hassouneh
University of Central Florida

INFLUENCE OF GEOGRAPHY ON THE HEALTHY GUT MICROBIOME AND ROLE OF THE GUT MICROBIOME IN IBD SYMPTOM AND DISEASE PROGRESSION

by

SAYF AL-DEEN HASSOUNEH

B.S. University of South Florida, 2016

M.S. University of Central Florida, 2018

A dissertation submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy in Biomedical Sciences

in the Burnett School of Biomedical Sciences

in the College of Medicine

at the University of Central Florida

Orlando, Florida

Summer Term

2021

Major Professor: Shibu Yooseph

ABSTRACT

The human gut microbiome is believed to play an integral role in host health and disease. In a microbial community, associations between constituent members play an important role in determining the overall structure and function of the community. To understand the nature of bacterial associations at the species level in healthy human gut microbiomes, we analyzed previously published collections of whole-genome shotgun sequence data, from fecal samples obtained from four different healthy human populations. Using a Random Forest Classifier, we identified bacterial species that were prevalent in these populations and whose relative abundances could be used to accurately distinguish between the populations. Bacterial association networks were also constructed using these signature species revealed conserved bacterial associations across populations and a dominance of positive associations over negative associations, with this dominance being driven by associations between species that are closely related either taxonomically or functionally. Functional analysis using protein families suggests that much of the taxonomic variation across human populations does not foment substantial functional differences. Next, multiple external healthy controls from the same geographical regions (American population) were compared to Inflammatory Bowel Disease (IBD) samples from the American population using shotgun sequencing data. We identified 34 bacterial species that were significantly elevated in IBD samples, relative to all control groups. These species elevated in IBD appear to play important roles in the healthy control groups, but it is possible that their over-abundance has deleterious effects on the host, possibly due to many of these

bacteria being involved in mucin degradation, immune modulation, antibiotic resistance, and inflammation. We also identified differences in functional capacities between IBD and healthy controls, and linked the changes in the functional capacity to previously published clinical research and to symptoms that commonly occur in IBD, such as rectal bleeding, diarrhea, vitamin K deficiency, and inflammation.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS/ABBREVIATIONS

aLPA: Asynchronous label propagation algorithm

ASPL: Average shortest path length

CFRP: Cohort functional role profile

CLR: Centered log-ratio

EM: Expectation-maximization

GGM: Gaussian graphical model

IBD: Inflammatory bowel disease

IBD: Irritable bowel syndrome

MFP: Module functional profile

PCA: Principal component analysis

RFC: Random forest classifier

rRNA: Ribosomal ribonucleic acid

SAF: Simplified annotation format

TBP: Trillion base-pairs

WGS: Whole-genome shotgun

CHAPTER 1: INTRODUCTION

The microbiome is defined as the community of microbes that exists throughout a host, both and internally and on external surfaces, and is believed to play an important role in maintaining host health.(Methé et al. 2012) This community of microbes exists as a complex consortium whose ecological and metabolic interactions are believed to heavily influence host health, especially in host metabolism, immunological modulation and development, mucosal regeneration and homeostasis, cell signaling, and pathogen resistance (Rahaman, n.d.; Thaiss et al. 2016; Das and Nair 2019; Shreiner, Kao, and Young 2015; Kostic, Xavier, and Gevers 2014; Kho and Lal 2018; Petersen and Round 2014). The disruption of this community, commonly termed ‘dysbiosis’, has been associated with a multitude of varying diseases such as obesity, diabetes, cardiovascular disease, inflammatory bowel disease (IBD), and various cancers (Koren et al. 2011; Karlsson et al. 2012; 2013; Franzosa et al. 2019; Becker, Neurath, and Wirtz 2015; Kostic et al. 2013). However, it is difficult to discern if the disruption of the gut microbiome is a cause or an effect of the associated diseases. Furthermore, defining what a healthy, or ‘eubotic’, gut microbiome is difficult due to the large number of bacterial species found in the gut and the large intra-personal variation of the gut microbiome across human populations (Huttenhower et al. 2012; Johnson et al. 2019). Identifying what microbiota constitute a healthy microbiome is integral, as one of the primary translational goals of microbiome research is to identify what a dysbiotic microbiome is and return it to its healthy state.

The bacterial compositions of the microbiome are most commonly examined by DNA sequencing, either by targeted sequencing of a marker gene or by shotgun sequencing of whole genomes. Targeted sequencing utilizes marker genes, such as the 16S ribosomal RNA gene for bacteria, as a phylogenetic marker (George E Fox et al. 1977). While the gut microbiome is comprised of bacteria, archaea, viruses, and fungi, most studies focus on the bacterial constituents of the gut microbiome, mainly due to bacteria being the largest constituents of the microbiome (Kho and Lal 2018).

While targeted sequencing approaches are cheaper and allow for higher-throughput, the highly conserved nature of the 16S rRNA gene and the short lengths of the sequenced regions makes it difficult to distinguish bacterial species (G. E. Fox, Wisotzkey, and Jurtshuk 1992; Ranjan et al. 2016). Furthermore, bacterial relative abundance estimation is obfuscated by the presence of multiple copies of the 16S gene within many bacterial species and the intragenic variation these copies exhibit (Rastogi et al. 2009; Ibal et al. 2019). Finally, due to the targeted sequencing only focusing on one gene, it is difficult to accurately identify the functional capacity of a bacterial species. In contrast to targeted sequencing, whole genome shotgun (WGS) sequencing yields more accurate estimates of relative abundances, better taxonomic resolution, and greater ability to estimate genomic functional capacity (Laudadio et al. 2018; Ranjan et al. 2016).

Regardless of sequencing methodology, the resulting sequencing data are compositional in nature due to the fixed number of reads generated by a sequencing instrument (Gloor et al. 2017). Compositional data are parts of a whole and thus only

contain relative information (Pawlowsky-Glahn and Egozcue 2006). Due to the compositional nature of sequencing data, it can be difficult to analyze differential abundance, infer associations, or estimate correlations (Aitchison 1982; Jonathan Friedman and Alm 2012; Tsilimigras and Fodor 2016; Pearson 1896). To mitigate the issues caused by the compositional nature of the sequencing data, we utilized a centered log-ratio (CLR) transformation (Aitchison 1982). The CLR transformation allows to examine the differential abundance data and infer associations without inducing spurious correlations (Gloor et al. 2017; Tsilimigras and Fodor 2016). Furthermore, the covariance matrix of log-transformed relative abundance data provides a good approximation of the covariance matrix of the log-transformed absolute abundance data enabling us to better model the associations between bacteria (Kurtz et al. 2015).

Associations within bacterial communities are composed of the direct and indirect interactions between the constituent bacteria, and are important for understanding the dynamics underlying the community assembly. Bacterial association networks are commonly inferred using pair-wise estimation correlation such as the Pearson or Spearman correlations. However, due to the compositional nature of the sequencing data used, it is difficult to accurately infer associations, especially due to the possibility of spurious correlations arising (Pearson 1896; Jonathan Friedman and Alm 2012). Even if the sequencing is CLR transformed, pair-wise correlation methods are unable to accurately infer bacterial association networks due to their inability to identify conditional independence (Kurtz et al. 2015). One way to identify the conditional independences within the bacterial association networks is to utilize a Gaussian graphical model (GGM)

to estimate the underlying covariance structure (Wermuth and Lauritzen 1990). Furthermore, due to the sparse nature of biological networks, in which most constituents are not strongly associated, it is important to conduct a sparse estimation of the bacterial association networks (Jerome Friedman, Hastie, and Tibshirani 2008; Jonathan Friedman and Alm 2012). Here, we utilize a Gaussian Graphical Model (GGM) framework in conjunction with a graphical lasso (glasso) to construct bacterial association networks from the CLR-transformed relative abundance data (Jerome Friedman, Hastie, and Tibshirani 2008; Loftus, Hassouneh, and Yooseph 2021). These bacterial association networks are represented as an unweighted graph in which nodes denote bacterial species and an edge between two nodes denotes an association between the corresponding bacterial species.

The random forest classifier (RFC) has become an important tool for classification and feature identification in microbiome research due to its ability to utilize with non-parametric, 'noisy', and multi-dimensional data (Breiman 2001; Díaz-Uriarte and Alvarez de Andrés 2006; Loomba et al. 2017; Saulnier et al. 2011; Roguet et al. 2018; Shi et al. 2005). The RFC can incorporate bacterial relative abundance data and metadata to generate a model that account for microbiome taxonomic profiles as well as subject characteristics, such clinical and demographic characteristics, when classifying samples. Furthermore, the RFC is able to assign feature importances, numeric values indicating the relative importance of a feature for achieving a correct classification, to the input features. These feature importances are helpful for identifying features that may be informative in relation to the specified sample labels. . One shortcoming of

these feature importances, however, is their lack of statistical significance. Due to the stochastic nature of model construction using an RFC, some features may be relatively important in one instance of an RFC model, but relatively unimportant in another instance of the RFC model. To enable us to utilize RFC feature importance to distinguish potentially important features and reduce the dimensionality of our data, we formulated a framework that allowed us to add statistical significance to the feature importances.

Importantly, many studies examining the microbiome suffer from a lack of cross-cohort consistency making it difficult to generalize findings to populations rather than just the utilized study groups (Pasolli et al. 2016). One proposed remedy for this lack of cross-cohort consistency is to utilize external samples from independent cohorts, especially when comparing diseased and healthy microbiomes, and applying the same methods and techniques across all samples (Pasolli et al. 2016; Thomas et al. 2019). To this end, we include two external healthy controls when analyzing IBD samples to enable us to generalize our findings to a population group, rather than just the study participants.

By utilizing shotgun sequencing data, we are able to more accurately determine relative abundances, bacterial taxonomies, and genomic functional capacities. Furthermore, employing the GGM framework on CLR-transformed data enables to approximate the covariance structure of the absolute abundances as well as account for conditional independence between the constituent species (Wermuth and Lauritzen

1990; Aitchison 1982). Finally, due to our use of external cohorts, we can corroborate our findings and arrive at generalizable conclusions that represent the population and not only the study participants.

References

1. Aitchison, J. 1982. "The Statistical Analysis of Compositional Data." *Journal of the Royal Statistical Society. Series B (Methodological)* 44 (2): 40. http://www.jstor.org/stable/2345821.
2. Becker, Christoph, Markus F Neurath, and Stefan Wirtz. 2015. "The Intestinal Microbiota in Inflammatory Bowel Disease." *ILAR Journal* 56 (2): 192–204. https://doi.org/10.1093/ilar/ilv030.
3. Breiman, Leo. 2001. "Random Forests." *Machine Learning*. https://doi.org/10.1023/A:1010933404324.
4. Consortium, Human Microbiome Project. 2012. "Structure, Function and Diversity of the Healthy Human Microbiome." *Nature* 486 (7402): 207–14. https://doi.org/10.1038/nature11234.
5. consortium, MetaHIT, Emmanuelle Le Chatelier, Trine Nielsen, Junjie Qin, Edi Prifti, Falk Hildebrand, Gwen Falony, et al. 2013. "Richness of Human Gut Microbiome Correlates with Metabolic Markers." *Nature* 500 (7464): 541–46. https://doi.org/10.1038/nature12506.
6. Cordasco, Gennaro, and Luisa Gargano. 2010. "Community Detection via Semi-Synchronous Label Propagation Algorithms." *2010 IEEE International Workshop on Business Applications of Social Network Analysis, BASNA 2010*. https://doi.org/10.1109/BASNA.2010.5730298.
7. Das, Bhabatosh, and G Balakrish Nair. 2019. "Homeostasis and Dysbiosis of the Gut Microbiome in Health and Disease." *Journal of Biosciences* 44 (5): 117. https://doi.org/10.1007/s12038-019-9926-y.
8. David, Lawrence A., Corinne F. Maurice, Rachel N. Carmody, David B. Gootenberg, Julie E. Button, Benjamin E. Wolfe, Alisha V. Ling, et al. 2014. "Diet Rapidly and Reproducibly Alters the Human Gut Microbiome." *Nature* 505 (7484): 559–63. https://doi.org/10.1038/nature12820.
9. Díaz-Uriarte, Ramón, and Sara Alvarez de Andrés. 2006. "Gene Selection and Classification of Microarray Data Using Random Forest." *BMC Bioinformatics* 7: 1–13. https://doi.org/10.1186/1471-2105-7-3.
10. Edgar, Robert C. 2018. "Accuracy of Taxonomy Prediction for 16S RRNA and Fungal ITS Sequences." *PeerJ* 6 (4): 1–29. https://doi.org/10.7717/peerj.4652.
11. Efron, B., and R. Tibshirani. 1986. "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy." *Statistical Science* 1 (1): 54–75. https://doi.org/10.1214/ss/1177013815.
12. Eiler, Alexander, Friederike Heinrich, and Stefan Bertilsson. 2012. "Coherent Dynamics and Association Networks among Lake Bacterioplankton Taxa." *The*

ISME Journal 6 (2): 330–42. https://doi.org/10.1038/ismej.2011.113.
13. Falony, Gwen, Marie Joossens, Sara Vieira-Silva, Jun Wang, Youssef Darzi, Karoline Faust, Alexander Kurilshikov, et al. 2016. “Population-Level Analysis of Gut Microbiome Variation.” *Science* 352 (6285): 560–64. https://doi.org/10.1126/science.aad3503.
14. Fox, G. E., J. D. Wisotzkey, and P. Jurtshuk. 1992. “How Close Is Close: 16S RRNA Sequence Identity May Not Be Sufficient to Guarantee Species Identity.” *International Journal of Systematic Bacteriology* 42 (1): 166–70. https://doi.org/10.1099/00207713-42-1-166.
15. Fox, George E, Linda J Magrum, William E Balcht, Ralph S Wolfef, and Carl R Woese. 1977. “Classification of Methanogenic Bacteria by 16S Ribosomal RNA Characterization (Comparative Oligonucleotide Cataloging/Phylogeny/Molecular Evolution).” *Evolution* 74 (10): 4537–41. https://doi.org/10.1073/pnas.74.10.4537.
16. Franzosa, Eric A, Alexandra Sirota-Madi, Julian Avila-Pacheco, Nadine Fornelos, Henry J Haiser, Stefan Reinker, Tommi Vatanen, et al. 2019. “Gut Microbiome Structure and Metabolic Activity in Inflammatory Bowel Disease.” *Nature Microbiology* 4 (2): 293–305. https://doi.org/10.1038/s41564-018-0306-4.
17. Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2008. “Sparse Inverse Covariance Estimation with the Graphical Lasso.” *Biostatistics* 9 (3): 432–41. https://doi.org/10.1093/biostatistics/kxm045.
18. Friedman, Jonathan, and Eric J. Alm. 2012. “Inferring Correlation Networks from Genomic Survey Data.” Edited by Christian von Mering. *PLoS Computational Biology* 8 (9): e1002687. https://doi.org/10.1371/journal.pcbi.1002687.
19. Gevers, Dirk, Subra Kugathasan, Lee A. Denson, Yoshiki Vázquez-Baeza, Will Van Treuren, Boyu Ren, Emma Schwager, et al. 2014. “The Treatment-Naive Microbiome in New-Onset Crohn’s Disease.” *Cell Host and Microbe* 15 (3): 382–92. https://doi.org/10.1016/j.chom.2014.02.005.
20. Gloor, Gregory B., Jean M. Macklaim, Vera Pawlowsky-Glahn, and Juan J. Egozcue. 2017. “Microbiome Datasets Are Compositional: And This Is Not Optional.” *Frontiers in Microbiology* 8 (NOV): 1–6. https://doi.org/10.3389/fmicb.2017.02224.
21. Gould, Alison L, Vivian Zhang, Lisa Lamberti, Eric W Jones, Benjamin Obadia, Nikolaos Korasidis, Alex Gavryushkin, Jean M Carlson, Niko Beerenwinkel, and William B Ludington. 2018. “Microbiome Interactions Shape Host Fitness.” *Proceedings of the National Academy of Sciences* 115 (51): E11951–60. https://doi.org/10.1073/pnas.1809349115.
22. Hibbing, Michael E, Clay Fuqua, Matthew R Parsek, and S Brook Peterson. 2010. “Bacterial Competition: Surviving and Thriving in the Microbial Jungle.” *Nature Reviews Microbiology* 8 (1): 15–25. https://doi.org/10.1038/nrmicro2259.
23. Huttenhower, Curtis, Dirk Gevers, Rob Knight, Sahar Abubucker, Jonathan H. Badger, Asif T. Chinwalla, Heather H. Creasy, et al. 2012. “Structure, Function and Diversity of the Healthy Human Microbiome.” *Nature* 486 (7402): 207–14. https://doi.org/10.1038/nature11234.
24. Ibal, Jerald Conrad, Huy Quang Pham, Chang Eon Park, and Jae Ho Shin. 2019. “Information about Variations in Multiple Copies of Bacterial 16S RRNA Genes

May Aid in Species Identification." *PLoS ONE* 14 (2): 1–15. https://doi.org/10.1371/journal.pone.0212090.
25. Johnson, Abigail J., Pajau Vangay, Gabriel A. Al-Ghalith, Benjamin M. Hillmann, Tonya L. Ward, Robin R. Shields-Cutler, Austin D. Kim, et al. 2019. "Daily Sampling Reveals Personalized Diet-Microbiome Associations in Humans." *Cell Host & Microbe* 25 (6): 789-802.e5. https://doi.org/10.1016/j.chom.2019.05.005.
26. Kara, Emily L, Paul C Hanson, Yu Hen Hu, Luke Winslow, and Katherine D McMahon. 2013. "A Decade of Seasonal Dynamics and Co-Occurrences within Freshwater Bacterioplankton Communities from Eutrophic Lake Mendota, WI, USA." *The ISME Journal* 7 (3): 680–84. https://doi.org/10.1038/ismej.2012.118.
27. Karlsson, Fredrik H, Frida Fåk, Intawat Nookaew, Valentina Tremaroli, Björn Fagerberg, Dina Petranovic, Fredrik Bäckhed, and Jens Nielsen. 2012. "Symptomatic Atherosclerosis Is Associated with an Altered Gut Metagenome." *Nature Communications* 3 (1): 1245. https://doi.org/10.1038/ncomms2266.
28. Karlsson, Fredrik H, Valentina Tremaroli, Intawat Nookaew, Göran Bergström, Carl Johan Behre, Björn Fagerberg, Jens Nielsen, and Fredrik Bäckhed. 2013. "Gut Metagenome in European Women with Normal, Impaired and Diabetic Glucose Control." *Nature* 498 (7452): 99–103. https://doi.org/10.1038/nature12198.
29. Kho, Zhi Y., and Sunil K. Lal. 2018. "The Human Gut Microbiome - A Potential Controller of Wellness and Disease." *Frontiers in Microbiology* 9 (AUG): 1835. https://doi.org/10.3389/fmicb.2018.01835.
30. Koren, O, A Spor, J Felin, F Fak, J Stombaugh, V Tremaroli, C J Behre, et al. 2011. "Human Oral, Gut, and Plaque Microbiota in Patients with Atherosclerosis." *Proceedings of the National Academy of Sciences* 108 (Supplement_1): 4592–98. https://doi.org/10.1073/pnas.1011383107.
31. Kostic, Aleksandar D., Eunyoung Chun, Lauren Robertson, Jonathan N. Glickman, Carey Ann Gallini, Monia Michaud, Thomas E. Clancy, et al. 2013. "Fusobacterium Nucleatum Potentiates Intestinal Tumorigenesis and Modulates the Tumor-Immune Microenvironment." *Cell Host & Microbe* 14 (2): 207–15. https://doi.org/10.1016/j.chom.2013.07.007.
32. Kostic, Aleksandar D., Ramnik J. Xavier, and Dirk Gevers. 2014. "The Microbiome in Inflammatory Bowel Disease: Current Status and the Future Ahead." *Gastroenterology* 146 (6): 1489–99. https://doi.org/10.1053/j.gastro.2014.02.009.
33. Kurtz, Zachary D., Christian L. Müller, Emily R. Miraldi, Dan R. Littman, Martin J. Blaser, and Richard A. Bonneau. 2015. "Sparse and Compositionally Robust Inference of Microbial Ecological Networks." Edited by Christian von Mering. *PLoS Computational Biology* 11 (5): 1–25. https://doi.org/10.1371/journal.pcbi.1004226.
34. Laudadio, Ilaria, Valerio Fulci, Francesca Palone, Laura Stronati, Salvatore Cucchiara, and Claudia Carissimi. 2018. "Quantitative Assessment of Shotgun Metagenomics and 16S RDNA Amplicon Sequencing in the Study of Human Gut Microbiome." *OMICS: A Journal of Integrative Biology* 22 (4): 248–54. https://doi.org/10.1089/omi.2018.0013.

35. Layeghifard, Mehdi, David M. Hwang, and David S. Guttman. 2017. “Disentangling Interactions in the Microbiome: A Network Perspective.” *Trends in Microbiology* 25 (3): 217–28. https://doi.org/10.1016/j.tim.2016.11.008.
36. Loomba, Rohit, Victor Seguritan, Weizhong Li, Tao Long, Niels Klitgord, Archana Bhatt, Parambir Singh Dulai, et al. 2017. “Gut Microbiome-Based Metagenomic Signature for Non-Invasive Detection of Advanced Fibrosis in Human Nonalcoholic Fatty Liver Disease.” *Cell Metabolism* 25 (5): 1054-1062.e5. https://doi.org/10.1016/j.cmet.2017.04.001.
37. Lupatini, Manoeli, Afnan K A Suleiman, Rodrigo J S Jacques, Zaida I Antoniolli, AdÃ£o de Siqueira Ferreira, Eiko E Kuramae, and Luiz F W Roesch. 2014. “Network Topology Reveals High Connectance Levels and Few Key Microbial Genera within Soils.” *Frontiers in Environmental Science* 2 (April). https://doi.org/10.3389/fenvs.2014.00010.
38. Methé, Barbara A, Karen E Nelson, Mihai Pop, Heather H Creasy, Michelle G Giglio, Curtis Huttenhower, Dirk Gevers, et al. 2012. “A Framework for Human Microbiome Research.” *Nature* 486 (7402): 215–21. https://doi.org/10.1038/nature11209.
39. Pasolli, Edoardo, Duy Tin Truong, Faizan Malik, Levi Waldron, and Nicola Segata. 2016. “Machine Learning Meta-Analysis of Large Metagenomic Datasets: Tools and Biological Insights.” *PLoS Computational Biology* 12 (7): 1–26. https://doi.org/10.1371/journal.pcbi.1004977.
40. Pawlowsky-Glahn, Vera, and J. J. Egozcue. 2006. “Compositional Data and Their Analysis: An Introduction.” *Geological Society Special Publication* 264: 1–10. https://doi.org/10.1144/GSL.SP.2006.264.01.01.
41. Pearson, K. 1896. “Mathematical Contributions to the Theory of Evolution.” *Proeedings of the Royal Society* 60 (1834): 489–98. http://books.google.com/books?hl=en&lr=&id=aIU_AQAAIAAJ&oi=fnd&pg=PA1&dq=Mathematical+Contributions+to+the+Theory+of+Evolution&ots=6q0ynawAzT&sig=FdqqMWpdG0a5gRGfvPbW2BRUw8I.
42. Petersen, Charisse, and June L Round. 2014. “Defining Dysbiosis and Its Influence on Host Immunity and Disease: How Changes in Microbiota Structure Influence Health.” *Cellular Microbiology* 16 (7): 1024–33. https://doi.org/10.1111/cmi.12308.
43. Prettejohn, Brenton J, Matthew J Berryman, and Mark D McDonnell. 2011. “Methods for Generating Complex Networks with Selected Structural Properties for Simulations: A Review and Tutorial for Neuroscientists.” *Frontiers in Computational Neuroscience* 5 (August). https://doi.org/10.3389/fncom.2011.00011.
44. Rahaman, Shaik O. n.d. *Gut Microbiome and Its Impact on Health and Diseases*.
45. Ranjan, Ravi, Asha Rani, Ahmed Metwally, Halvor S. McGee, and David L. Perkins. 2016. “Analysis of the Microbiome: Advantages of Whole Genome Shotgun versus 16S Amplicon Sequencing.” *Biochemical and Biophysical Research Communications* 469 (4): 967–77. https://doi.org/10.1016/j.bbrc.2015.12.083.
46. Rastogi, Rajat, Martin Wu, Indrani Dasgupta, and George E. Fox. 2009.

"Visualization of Ribosomal RNA Operon Copy Number Distribution." *BMC Microbiology* 9: 208. https://doi.org/10.1186/1471-2180-9-208.
47. Roguet, Adélaïde, A. Murat Eren, Ryan J. Newton, and Sandra L. McLellan. 2018. "Fecal Source Identification Using Random Forest." *Microbiome* 6 (1): 1–15. https://doi.org/10.1186/s40168-018-0568-3.
48. Saulnier, Delphine M., Kevin Riehle, Toni Ann Mistretta, Maria Alejandra Diaz, Debasmita Mandal, Sabeen Raza, Erica M. Weidler, et al. 2011. "Gastrointestinal Microbiome Signatures of Pediatric Patients with Irritable Bowel Syndrome." *Gastroenterology* 141 (5): 1782–91. https://doi.org/10.1053/j.gastro.2011.06.072.
49. Saunders, Aaron M, Mads Albertsen, Jes Vollertsen, and Per H Nielsen. 2016. "The Activated Sludge Ecosystem Contains a Core Community of Abundant Organisms." *The ISME Journal* 10 (1): 11–20. https://doi.org/10.1038/ismej.2015.117.
50. Sender, Ron, Shai Fuchs, and Ron Milo. 2016. "Revised Estimates for the Number of Human and Bacteria Cells in the Body." *PLOS Biology* 14 (8): e1002533. https://doi.org/10.1371/journal.pbio.1002533.
51. Shetty, Sudarshan A, Floor Hugenholtz, Leo Lahti, Hauke Smidt, and Willem M de Vos. 2017. "Intestinal Microbiome Landscaping: Insight in Community Assemblage and Implications for Microbial Modulation Strategies." *FEMS Microbiology Reviews* 41 (2): 182–99. https://doi.org/10.1093/femsre/fuw045.
52. Shi, Tao, David Seligson, Arie S. Belldegrun, Aarno Palotie, and Steve Horvath. 2005. "Tumor Classification by Tissue Microarray Profiling: Random Forest Clustering Applied to Renal Cell Carcinoma." *Modern Pathology* 18 (4): 547–57. https://doi.org/10.1038/modpathol.3800322.
53. Shreiner, Andrew B, John Y Kao, and Vincent B Young. 2015. "The Gut Microbiome in Health and in Disease:" *Current Opinion in Gastroenterology* 31 (1): 69–75. https://doi.org/10.1097/MOG.0000000000000139.
54. Su, Weijie, Malgorzata Bogdan, Emmanuel Candès, and Emmanuel Candes. 2017. "False Discoveries Occur Early on the Lasso Path." *Annals of Statistics* 45 (5): 2133–50. https://doi.org/10.1214/16-AOS1521.
55. Thaiss, Christoph A., Niv Zmora, Maayan Levy, and Eran Elinav. 2016. "The Microbiome and Innate Immunity." *Nature* 535 (7610): 65–74. https://doi.org/10.1038/nature18847.
56. Thomas, Andrew Maltez, Paolo Manghi, Francesco Asnicar, Edoardo Pasolli, Federica Armanini, Moreno Zolfo, Francesco Beghini, et al. 2019. "Metagenomic Analysis of Colorectal Cancer Datasets Identifies Cross-Cohort Microbial Diagnostic Signatures and a Link with Choline Degradation." *Nature Medicine* 25 (4): 667–78. https://doi.org/10.1038/s41591-019-0405-7.
57. Tsilimigras, Matthew C.B., and Anthony A. Fodor. 2016. "Compositional Data Analysis of the Microbiome: Fundamentals, Tools, and Challenges." *Annals of Epidemiology* 26 (5): 330–35. https://doi.org/10.1016/j.annepidem.2016.03.002.
58. Venter, J Craig, Karin Remington, John F Heidelberg, Aaron L Halpern, Doug Rusch, Jonathan A Eisen, Dongying Wu, et al. 2004. "Environmental Genome Shotgun Sequencing of the Sargasso Sea." *Science* 304 (5667): 66. https://doi.org/10.1126/science.1093857.

59. Větrovský, Tomáš, and Petr Baldrian. 2013. “The Variability of the 16S RRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses.” Edited by Josh Neufeld. *PLoS ONE* 8 (2): e57923. https://doi.org/10.1371/journal.pone.0057923.
60. Villmones, Heidi Cecilie, Erik Skaaheim Haug, Elling Ulvestad, Nils Grude, Tore Stenstad, Adrian Halland, and Øyvind Kommedal. 2018. “Species Level Description of the Human Ileal Bacterial Microbiota.” *Scientific Reports* 8 (1): 1–9. https://doi.org/10.1038/s41598-018-23198-5.
61. Wermuth, Nanny, and Steffen Lilholt Lauritzen. 1990. “On Substantive Research Hypotheses, Conditional Independence Graphs and Graphical Chain Models.” *Journal of the Royal Statistical Society: Series B (Methodological)* 52 (1): 21–50. https://doi.org/10.1111/j.2517-6161.1990.tb01771.x.
62. Zhou, Jizhong, Ye Deng, Feng Luo, Zhili He, and Yunfeng Yang. 2011. “Phylogenetic Molecular Ecological Network of Soil Microbial Communities in Response to Elevated CO2.” Edited by David Relman. *MBio* 2 (4): e00122-11. https://doi.org/10.1128/mBio.00122-11.

CHAPTER 2: BACTERIAL ASSOCIATIONS IN THE HEALTHY HUMAN GUT MIROBIOME ACROSS POPULATIONS

Note: This section has been published in part and the citation link is: Loftus, M., Hassouneh, S. A.-D., & Yooseph, S. (2021). Bacterial associations in the healthy human gut microbiome across populations. Scientific Reports, 11(1), 1–14. https://doi.org/10.1038/s41598-021-82449-0.

Introduction

The community of microbial cells in the human gut is estimated to be comparable in magnitude to the number of human cells(Sender, Fuchs, and Milo 2016). This community, deemed the human gut microbiome, is mainly composed of bacteria, archaea, fungi, and viruses, with bacteria being the largest constituent. These bacterial cells exist in a complex consortium of ecological and metabolic interactions that ultimately influence the taxonomic and functional profile of the microbial community, as well host health. The gut microbiome of healthy individuals is believed to be mainly symbiotic and is known to play important roles in host metabolism, immunological modulation and development, cell signaling, pathogen colonization resistance, and mucosal homeostasis (Kho and Lal 2018; Kostic, Xavier, and Gevers 2014; Thaiss et al. 2016).

The continued stability of this community and its functions, i.e. homeostasis (Das and Nair 2019; Shreiner, Kao, and Young 2015), is important and its disruption, broadly

described as ‘dysbiosis’ (Petersen and Round 2014), has been associated with numerous diseases including, but not limited to: diabetes(Karlsson et al. 2013), cardiovascular disease(Koren et al. 2011; Karlsson et al. 2012), obesity(consortium et al. 2013), inflammatory bowel disease(Franzosa et al. 2019),(Becker, Neurath, and Wirtz 2015), and various cancers(Kostic et al. 2013). However, it remains unclear whether disease onset is the consequence or cause of the microbiome community disruption. Furthermore, what constitutes a healthy gut microbiome is still under investigation due to the overwhelming amount of bacterial species found in the gut, and the large variation in their carriage rates across human populations and individuals(Consortium 2012; Johnson et al. 2019). These issues are of great importance as one of the ultimate goals of microbiome research is to modulate the community from a ‘dysbiotic’ state into a healthy ‘homeostatic’ one.

Early research towards this goal chose to limit their focus to taxonomic differences between healthy and disease microbiomes(Villmones et al. 2018; Gevers et al. 2014; David et al. 2014). While these comparisons are valuable, since the bacterial community taxonomic profile generally represents the potential metabolic and transcriptional profiles that are present within the ecosystem; simply profiling the community fails to acknowledge the underlying bacterial associations and the impact they exert on both the microbial ecosystem and host health. In fact, many studies within natural systems and animal hosts have shown that the associations (positive and negative) between bacteria are an important foundation for the continued stability and proper functioning of these ecosystems(Zhou et al. 2011; Lupatini et al. 2014; Eiler, Heinrich, and Bertilsson 2012; Kara et al. 2013; Shetty et al. 2017; Gould et al. 2018).

As such, it is of great importance to assess the relationships that exist between bacteria within the healthy human gut microbiome in order to better understand the ecological associations important for the structure and maintenance of the gut microbiome and its related processes. Naturally, this raises an important question: are there similarities in the structural features of bacterial association networks in human gut microbiomes across healthy populations, and if so, are there conserved associations?

Microbial associations in a community are characterized by both direct and indirect interactions between the constituents(Hibbing et al. 2010). In this paper, we depict these associations using a weighted graph (network) in which the nodes represent bacterial species and an edge between two nodes represents an association between the corresponding species, with the edge weight capturing the strength of the association. This framework enables us to model both positive and negative associations between species, and thus can help to shed light on cooperation and competition between species in the community. Once a network is constructed, an analysis of the various topological properties of the network can enable us to decipher the underlying ecological rules associated with the microbial ecosystem. These networks also provide the ability to determine the relative importance of species for ecosystem structure and function.

Microbial association networks are typically constructed from a sample-taxa count matrix generated by collecting multiple samples from the community and determining the taxa counts in each sample. With the availability of high-throughput and

low-cost DNA sequencing technologies, these counts are generated by sequencing the collected biological samples. Microbiome sequence data are generated either using a targeted approach, involving the sequencing of a taxonomic marker gene (e.g., the 16S ribosomal RNA gene)(George E Fox et al. 1977) or using a whole-genome shotgun (WGS) sequencing approach(Venter et al. 2004). However, estimates of taxa abundances using 16S rRNA sequences can be confounded by several factors including the presence of multiple copies and variants of the 16S rRNA gene in genomes, and the lack of taxonomic resolution in the selected variable region of the 16S gene(Větrovský and Baldrian 2013; Edgar 2018). Conversely, WGS data can be used to provide more accurate estimates of genome relative abundances as well as higher resolution taxonomic classification, compared to 16S rRNA data(Ranjan et al. 2016; Laudadio et al. 2018). Regardless of sequencing approach, the taxa count data generated by DNA sequencing are compositional in nature and provide only relative abundance information of the constituent taxa(Gloor et al. 2017). This poses challenges for inferring associations, and the computation of measures like correlation directly from the observed sequence counts can be misleading(Jonathan Friedman and Alm 2012). While several methods have been proposed for constructing association networks that address this challenge(Layeghifard, Hwang, and Guttman 2017), here we use a Gaussian Graphical Model (GGM) framework on Centered Log-Ratio (CLR) transformed count data to construct an association network(Aitchison 1982; Kurtz et al. 2015).

We are motivated by the observation that the covariance matrix of a multivariate Gaussian distribution used to fit log-transformed relative count data provides a good

approximation to the covariance matrix of the log-transformed absolute count data(Aitchison 1982). The GGM framework also enables the modeling of conditional dependencies of the random variables that represent taxa abundances. The adjacency matrix of the association network that we construct is the inverse covariance matrix (i.e. the precision matrix) of the underlying multivariate Gaussian distribution used in the GGM. This graph has the property that an edge exists between two nodes if and only if the corresponding entry in the precision matrix is non-zero. A zero entry in the precision matrix indicates conditional independence between the two corresponding random variables. We also incorporate sparsity in our framework using the l1-penalty norm and construct sparse association networks using the graphical lasso method (glasso)(Jerome Friedman, Hastie, and Tibshirani 2008).

In this study we investigate bacterial association networks in gut microbiomes across four healthy human populations. Previous studies analyzing bacterial association networks have mainly used 16S rRNA data, and given its lower taxonomic resolution, these studies have analyzed associations at the genus level(Falony et al. 2016; Jerome Friedman, Hastie, and Tibshirani 2008)Instead, here we use a large collection of WGS samples from multiple human populations to investigate bacterial associations at the species level. We use a machine learning algorithm to identify a set of signature species that can accurately distinguish between the different healthy populations. Using these signature species, we construct networks by employing a glasso method that incorporates a bootstrapping(Efron and Tibshirani 1986) approach to reduce the number of false positive edges inferred(Su et al. 2017). We analyze these networks to

assess the theoretical ecology, and potential importance of species within healthy human gut microbial communities.

Results

Signature Species in the Healthy Human Gut Microbiome

For each cohort, the prevalence of individual species across all samples was measured and plotted. All cohorts exhibited a skewed bi-modal distribution (**Figure 1a)**. The first peak in the distribution was centered around a prevalence of 10%, while the second peak occurred around a prevalence of 90%**.** This skewed bi-modal distribution has been previously observed in a microbial community, and organisms that were highly prevalent were deemed the 'abundant core' as they were found to account for the majority of total sample abundances(Saunders et al. 2016). The 90% prevalent species set for each cohort consisted of 127 (American), 109 (Indian), 182 (European), and 146 (Japanese) species respectively, and these species were found to account for a large majority of the total sample proportions, the median values for the cohorts were 0.93 (American), 0.93 (Indian), 0.87 (European), and 0.81 (Japanese) **(Figure 1b)**. Next, we utilized a Random Forest Classifier (RFC) to determine the effect of prevalence thresholds on the ability to distinguish between cohorts using the taxonomic profiles of the constituent samples. The RFC was able to distinguish between cohorts with an F1-score >0.85 for all prevalence thresholds (0%, 50%, 90%, 100%), but demonstrated the highest F1-score at the 90% threshold, even though less than 10% of the original species remained **(Figure 2)**. Based on this analysis, we define the set of *signature*

species to be the union of the prevalent (>90%) species sets from the four cohorts. The signature species set consisted of 202 species and was used for constructing the bacterial association network for each cohort. We explored the variability in signature species relative abundance between samples using principal components analysis (PCA) applied to the CLR-transformed data **(Figure 1c)**. PCA showed evidence for separation of samples from the Indian and American cohorts, but ultimately the PCA only explained a small amount of the total variance (PC1: 11.38%, PC2: 10.91%).

Bacterial Association Networks

Prior to its application on the cohort data, the network inference method with bootstrapping was tested on synthetic data (see supplemental) notably, most graph-types were inferred with an F1-score above 0.7 (band: 0.974, hub: 0.885, random: 711, cluster: 0.692, scale-free: 0.416) (**Figure 3a**). Furthermore, we demonstrate that as the sample-to-taxa ratio increases, F1-scores approach 1, and all groups demonstrate mean F1-scores above 0.9 (**Figure 3a**). Finally, we observe that our network inference method tends to underestimate edge weights, and on average the estimated edge weights are 53.23% of the actual edge weights (**Figure 3b**). A bacterial association network was constructed for each cohort using the CLR-transformed relative abundances of the signature species (see methods). Each network was modeled as an undirected graph consisting of nodes and edges **(Figure 4)**. At a high-level, differences in the structure of the four networks were apparent. The European, Japanese, and Indian networks exhibited a high density of edges occurring between nodes from the

phylum *Firmicutes*, whereas the American network had the largest density of edges existing between nodes from the phylum *Bacteroidetes*. Positive associations were dominant in all networks (American: 0.98, Indian: 0.97, European: 0.96, Japanese: 0.96), and negative associations involve nodes from the phylum *Firmicutes*. Network topology was studied by calculating the following network properties: average shortest path length (ASPL), transitivity, modularity, degree assortativity, and genera assortativity (see methods) **(Table 1)**. These properties were compared to random networks using Monte Carlo simulations (see supplemental). All cohort networks were deemed non-random in their topology and exhibited significantly low ASPL (all P-values < 0.05), significantly high modularity (all P-values < 0.01), significantly high transitivity (all P-values $< .001$), significantly high genera assortativity (all P-values $< .001$) and significantly high degree assortativity (all P-values < 0.01), relative to the random networks. The low ASPL within networks suggest that nodes are connected to one another through short paths within the network. The high transitivity and modularity indicate that nodes form cliques and networks exhibit compartmentalization (modules), respectively. Lastly, the high (assortative) degree assortativity and genera assortativity portrays that nodes tend to form connections to other nodes that have a similar degree and taxonomy.

Theoretical Ecology based on Bacterial Association Networks

All cohort networks were found to contain highly similar distributions of association (edge) weights, where positive associations were more frequent and greater

in magnitude than negative associations **(Figure 5a)**. Furthermore, a large percentage of associations (American: 40%, Indian: 40%, European: 40%, Japanese: 53%) were found to be shared with at least one other network and these associations were all positive **(Figure 5b)**. A conserved structure of 14 associations, composed of 20 species **(Figure 5c)**, mainly from the genus *Bacteroides,* was observed to be contained within all networks **(Figure 6).** No negative association was retained across networks. However, viewed at the higher taxonomic rank for those species involved in negative associations, we observed that across all cohort networks, members from the phylum *Firmicutes* were involved in a large percentage of the negative associations (American: 100%, Indian: 100%, European: 62.5% , Japanese: 100%), and specifically these negative associations were mainly occurring between species from the order *Clostridiales* (American: 25%, Indian: 89%, European: 56%, Japanese: 100%) **(Figure 7)**. We next explored the taxonomic relationship between species and their association type (positive or negative) **(Figure 8a),** as well as the genome functional profile dissimilarities, according to Bray-Curtis dissimilarity, between network neighbors against their association weight **(Figure 8b)**. We found that most positive associations take place between bacteria that are more taxonomically and functionally similar, while negative associations were never found between species within the same genus, or between species with low genome functional profile distance (<0.2).

Network Cliques and Module Detection

As our networks exhibited both high transitivity and modularity, we sought to investigate the cliques and modules of species contained within them. We first found all cliques of three species (1,588 unique cliques) within our networks (see methods). Of these cliques: 113 were shared in at least 1 other network, 8 were shared across three networks, and only 1 (*Bacteroides caecimuris, Bacteroides fluxus, Bacteroides thetaiotaomicron)* was found in all networks. Species from 66 genera were shown to participate in clique formation, however, species from the genus *Bacteroides* were found to be involved in the largest percentage of cliques (American: 21.0%, Indian: 4.0%, European: 4.9%, Japanese: 5.8%) within most cohort networks **(Figure 9)**. Interestingly, the cliques that contained species from *Bacteroides* were also the most retained (American: 20.9%, Indian: 8.5%, European: 8.5%, Japanese: 10.8%) across all cohorts **(Figure 10)**.

Following clique analysis, we performed module detection utilizing an asynchronous Label Propogation Algorithm (aLPA) (see supplemental) which identified a total of 49 modules (American: 10, European: 11, Indian: 14, Japanese: 14) that contained 3 or more members (Cordasco and Gargano 2010) **(Figure 11).** The quality of network partitioning by the module detection algorithm (performance) was analyzed (American: 0.96, Indian: 0.98, European: 0.94, Japanese: 0.98) showing that the majority of edges between nodes were contained within modules (see supplemental). PCA was utilized to examine the variance between Module Functional Profiles (MFP's) of the different cohort **(Figure 8c)**. This analysis revealed MFPs fell within one of four clusters, and each cohort had representation within each cluster. Taxonomic and

functional characteristics of the clusters were analyzed**.** Cluster I contained modules formed mainly by the genera *Streptococcus* and *Bifidobacterium* **(Figure 12a)**. Cluster II modules were mainly composed of species from the genera *Alistipes*, *Bacteroides*, and *Prevotella* (**Figure 12b)**. Cluster III modules were dominated by the genera *Bacteroides* **(Figure 12c).** Cluster IV modules were mainly composed of species from the genera *Blautia*, *Eubacterium*, *Lachnoclostridium*, and *Ruminococcus* **(Figure 12d)**. Functional analysis of clusters revealed unique roles in each cluster: Cluster I (increase in toxin production, protein secretion, anaerobic metabolism, nucleic acid metabolism; decrease in thiamine biosynthesis), Cluster II (increase in cellular metabolism and protein degradation; decrease in cell division and signal transduction), Cluster III (increase in chemoautotrophy, sulfur and phosphorous metabolism, DNA metabolism), and cluster IV (increase in transcription factors; decrease in roles associated with adaptation to atypical conditions) (**Figure 13**).

We next analyzed the sample functional profiles using PCA **(Figure 14)**. PCA explained a modest amount of variance (PC1: 27.82%; PC2: 5.99%) although samples between cohorts were found to overlap. When analyzing the Cohort Functional Role Profiles (CFRP's), only 11 differences, when comparing the signs (+/-), out of the 113 found roles were found, and only the European cohort exhibited more than two differences **(Figure 15)**.

Node centrality analysis

We utilized degree and betweenness centrality measurements to identify “hub” and “bottleneck” nodes, respectively, within our networks (see supplemental). These centrality measurements were selected because ‘hubs’ and ‘bottlenecks’ are nodes that could have strong influence within a network and have been utilized previously to identify important species within microbial ecosystems(Lupatini et al. 2014; Prettejohn, Berryman, and McDonnell 2011; Kara et al. 2013). Considering all cohort networks were deemed assortative in respect to their degree assortativity we did not expect to find network “hub” nodes. However, we did find that nearly all modules, within each cohort, were disassortative in their degree assortativity which hinted at “hub” nodes existing within modules **(Figure 16)**. For these reasons we chose to select the node within each module that exhibited the highest degree **(See Figure 4)**, and the top 10 nodes within each network with the highest betweenness. Across all cohorts we found variation in the species deemed module ‘hubs’ and ‘bottlenecks’ **(Figure 17a)**, although at the genus level there was a large amount of agreement **(Figure 17b).** In at least three out of the four cohorts, species from *Bacteroides, Alistipes*, *Bifidobacterium*, *Eubacterium*, and *Streptococcus* were designated as ‘hubs’, whereas species from *Bacteroides* and *Lachnoclostridium* were designated as ‘bottlenecks’.

Discussion

In this study, we used WGS data in conjunction with a network inference method that is robust to sequence data compositionality in order to analyze the associations occurring between species within the healthy human gut microbiome across different populations. The association networks were constructed utilizing the signature species.

We demonstrated that bacterial association networks, across all cohorts, do not have the same properties as random networks. However, relative to each other, the networks of the four cohorts display similar properties. Random networks are known to contain short average path lengths, low node clustering, and high modularity[44,45]. Compared to random networks each cohort network was found to exhibit significantly shorter average shortest path lengths, significantly higher transitivity (clustering), significantly higher modularity, significantly higher degree assortativity, and significantly higher genera assortativity. We posit that the similarities in network properties reflect an organization of the bacterial community that is important to underlying ecological processes. For instance, the short average path lengths within our networks could imply rapid signaling between bacterial species, potentially facilitating swift changes in community metabolism. This is supported by previous studies demonstrating that the human gut microbiome exhibits rapid alterations in bacterial metabolism and abundance in conjunction with change in host diet[19].

In addition to exhibiting similar properties, cohort networks also shared a large percentage of associations (American: 40%, Indian: 40%, European: 40%, Japanese: 53%), including a conserved set of 14 positive associations composed of 20 species.

These conserved associations may be indicative of strong partner fidelity, important ecological relationships, or potentially obligate partnerships. Furthermore, we found that taxonomically and functionally similar species tended to have positive associations. This finding was unexpected as some previous studies on microbial ecosystems, including the human gut[46–48], have shown negative interactions between bacteria (competition, predation, etc.) should be the dominant form of interaction[49], especially when those bacteria are taxonomically or functionally alike[50]. The differences between our results and the aforementioned research may be due to their use of non-transformed data and pairwise analysis as well as the use of low-resolution taxonomic sequencing data or in-vitro analysis[33]. Our findings would suggest that kin-selection[51] (positively associating with those of similar lineage in order to directly or indirectly pass on ones genes), as opposed to competitive exclusion[52] (bacteria with similar lineage or functionality are more likely to compete within a habitat), is more prevalent within the healthy gut microbiome. This observation cannot be excluded as there is precedence within microbial ecosystems for the co-occurrence of bacteria with similar genetic traits[50,53], and studies on bacterial dynamics in the gut that suggest close relatives to bacteria currently present in the gut are more likely to be recruited into the community, i.e. phylogenetic under-dispersion (nepotism) hypothesis[54].

Within all cohorts, positive associations were not only the most dominate form of association, but also the only associations that were shared across networks. This finding seems logical as within the anoxic environment of the gut, bacterial energy production is limited which would make positive associations, such as mutual cross-feeding, preferable in order to produce and utilize energy more efficiently[55]. In addition,

ecological community theory suggests that partitioning of resources in space and time drive coexistence[56], and bacteria within the human gut microbiota are known to exhibit diurnal fluctuations[57] and exist in distinct spatial organizations[58–60]. Furthermore, positive associations between species are also known to alleviate ecosystem stresses and allow for a greater diversity of organisms to coexist[61], and the healthy gut microbiome has a high level of biodiversity[62]. However, it is Important to be cognizant that a positive association between species does not rule out the presence of a negative interaction completely, as negative interactions between species can still have a net positive result if an increased survival rate is occurring, as well as to understand that these positive associations are not always indicative of cooperative activities as they could simply reflect a common preferred environmental niche[61]. In contrast to the large proportion of shared positive associations, negative associations were always unique to a specific cohort; however, as we viewed the higher-level taxonomic ranking of species involved in negative associations, we found that across all cohorts most negative associations were occurring between species from the order *Clostridiales*. Species from the order *Clostridiales* are known to be largely cellulolytic, in that they mainly hydrolyze the polysaccharide cellulose[63]. This limited nutritional niche could theoretically create competition between *Clostridiales sp*., and in any case, these associations might be important for community stability as negative associations within microbial communities are thought to be an important stabilizing force[48]. While the healthy human gut microbiome is indeed routinely described as stable[62], the low abundance of negative associations within our networks suggests that the gut microbiome would be more vulnerable to positive feedback loops between species which could result in instability[48].

We hypothesize that the high modularity found within all cohort networks could mitigate the vulnerability to positive feedback loops as high network modularity has been shown to have a stabilizing effect[45].

We used a module detection algorithm to identify groups of highly connected species within our networks. The algorithm identifies modules of species which have previously been noted to benefit by growing together (e.g. *Bifidobacterium sp.*)[64]. As we analyzed the variance between module functional profiles, using PCA, we found that modules gravitated towards one of four clusters. Although some cohorts had a greater proportion of modules within certain clusters, all cohorts had some level of representation within each cluster. Upon further analysis, we were able to find distinct functional and taxonomic differences between module clusters, but we were not able to distinguish overt functional differences between CFRP's. This implies that a general set of functions is present in each healthy population regardless of taxonomic differences. These module clusters may be indicative of niches that are retained in the healthy human gut microbiome, and the redundancy of multiple modules of a cohort falling within a cluster is potentially a further stabilizing force for the ecosystem. These findings agree with previous studies showing comparable communities and high functional redundancy across gut microbiome data sets[53,65].

Lastly, we identified species that acted as "hubs" or "bottlenecks" within the structure of cohort networks. Notably, we found *Bacteroides sp.* were designated as both "hubs" and "bottlenecks" across all networks. Interestingly, *Bacteroides sp.* were also found to be the largest constituent of bacterial cliques and these cliques were the

most retained across all cohorts. Additionally, of the 20 species from the 14 conserved associations found across networks, most were species belonging to *Bacteroides*. These findings suggest that *Bacteroides sp.* are important drivers of the ecosystem within the healthy human gut microbiome. Interestingly, previous studies have also designated *Bacteroides sp.,* such as *Bacteroides fragilis* and *Bacteroides stercosis*, as potentially important (keystone) species within the human gut microbiome[66].

It is important to consider the limitations of our study. Our samples originated from different geographical locations and utilized different preparation procedures both of which are known to introduce biases[24,67,68]. Additionally, due to the cross-sectional nature of our data we are only able to capture snapshots of the gut microbiome and are unable to examine the dynamics of the ecosystem. Furthermore, we utilized a reference-based mapping approach for taxonomic classification potentially causing our classifications to be limited by the genomes available. Finally, the constructed bacterial networks were undirected, and the study was non-mechanistic which prevents us from being able to examine the influence individual species have on one another (unidirectional ecological interactions).

In closing, we have demonstrated that bacterial communities across healthy human populations are similar in their organization and functional capacities. We have also revealed that positive associations regularly occur between taxonomically and functionally related species despite bacterial carriage differences, healthy human gut microbiomes across populations exhibit less variation (structural and functional) than previously believed. Our future research will build upon these findings to better

understand how bacterial associations change within the disease microbiome. Also, by using the prevalent species, we can minimize the 'noise' of bacterial variation across hosts, especially since low prevalence species may ultimately be transient in nature[41]. This could be advantageous as it has been suggested that the most abundant organisms are the ones that act as "ecosystem engineers"[50], and the study of these organisms would be important to understand how the microbiome responds to disturbances.

Materials and Methods

Data Acquisition

We utilized 606 WGS fecal samples (1.7 Tbp), which were obtained from four previously published human gut microbiome studies from four different healthy human populations (cohorts). Three cohort datasets were downloaded from the NCBI Sequence Read Archive (SRA): American[15] (PRJNA48479; 202 samples), Indian[69] (PRJNA397112; 106 samples), and European[70] (PRJEB2054; 120 samples). The Japanese cohort dataset was downloaded from the DDBJ Sequence Read Archive (DRA): Japanese[71] (PRJDB4176; 178 samples) **(Figure 18)**.

Data Pre-Processing

Reads from all samples were first trimmed using Trimmomatic[72] (version 0.36) and then human reads were filtered using BowTie2[73] (version 5.4.0) and the GRCh38.p12 (https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.38/) human

reference genome. After removal of human reads, 15.9 billion high-quality reads remained. **(Figure 19)**.

Read Mapping and Species-Level Taxonomic Profiling

Reads were mapped to a collection of 10,839 bacterial reference strain genomes downloaded from RefSeq[74], using Bowtie2. The read mapping information was analyzed using a probabilistic framework based on a mixture model to estimate the relative copy number of each reference genome in a sample. This framework used an Expectation-Maximization (EM) algorithm to optimize the log-likelihood function associated with the model[75]. The EM algorithm was found to be highly accurate when benchmarked using simulated WGS reads produced by WGSim (https://github.com/lh3/wgsim) **(Figure 20)**. Sub-sampling and benchmark testing of sample read mapping counts showed that a read depth of 250,000 mapped reads at a noise threshold of 1e-5 correlated well with samples mapping over 5 million mapped reads ($R^2 > 0.85$, **Figure 21)**. Any bacterial strain found in a sample below 1e-5 relative abundance was considered statistical noise and was dropped to an abundance of 0. Strains were then grouped by their species classification and their relative abundances were summed to produce species abundances.

Bacterial Genome Annotation and Functional Profiles

All bacterial reference genomes were functionally annotated in-house to create reference strain functional profiles. Before genome annotation, we utilized CheckM[76]

(v1.0.13) to ensure that these reference genomes were mostly complete **(Figure 22)**. Prodigal[77] (version 2.6.3) was used to identify genes, and generate protein sequence translations, which were then provided to InterProScan[78] (version 5.39-77.0) to find matches to protein families using the TIGRFAM[79] (version 15.0) database. The functional profile for a bacterial strain was created by identifying the total number TIGRFAM matches to the strain, and subsequently converting these counts to relative abundances. The functional profile for a bacterial species was created separately for each cohort. This was computed by first finding the average genome abundance of each strain within the cohort, weighting the strain functional profiles based on these proportions, and then aggregating the resulting strain profiles. Each species functional profile was then CLR-transformed. CLR-transformation is defined as:

$$\mathrm{clr}(x)=\left[\ln\frac{x_1}{g(x)}\ldots,\ln\frac{x_2}{g(x)}\ldots,\ln\frac{x_D}{g(x)}\right]$$

where x is the vector of species abundances within each sample, D is the total number of species. The geometric mean of vector x is defined as:

$$g(x)=\sqrt[D]{x_1\times x_2\times\ldots x_D}$$

TIGRFAM functional annotations were obtained from TIGRFAMs_ROLE_LINK and TIGRFAM_ROLE_NAMES files generated by J. Craig Venter Institute (JCVI) (ftp://ftp.jcvi.org/pub/data/TIGRFAMs/14.0_Release/).

Cohort Sample Functional Profiling

A Simplified Annotation Format (SAF) file containing the bacterial chromosomal coordinates of TIGRFAMS (features) for all reference strains was provided to FeatureCounts[80] (Subread package 2.0.0) to find the total features contained within sample reads. Counts of features were subsequently length normalized, summed, and re-normalized (by total) for each sample producing sample functional profiles. Protein families were grouped by their TIGRFAM role, and their relative abundances were aggregated and CLR-transformed to generate the cohort functional role profiles (CFRP). Roles that were a different sign (+/-) in one cohort, when compared to all other cohorts, were considered different (elevated/reduced).

Construction of Bacterial Association Networks

For each cohort, a sample-taxa matrix was constructed containing the relative abundances of the signature species in each sample. The bacterial association network for a cohort was constructed from its CLR transformed sample-taxa matrix using the GGM framework. In each case, a sparse precision matrix was computed using the R[81] huge[82] package, and this matrix formed the adjacency matrix of the association network. The tuning parameter ρ in the l_1-penalty model for sparse precision matrix estimation was chosen using the stability approach to regularization (StARS) method[83.] In order to reduce the number of false positives, the estimated sparse precision matrix Ω was processed further using a bootstrap method as follows: r bootstrap datasets,

each with n samples, were generated from the original CLR-transformed matrix by random sampling with replacement. A sparse precision matrix was estimated from each bootstrap dataset using the same previously chosen value of the tuning parameter ρ used to estimate Ω. The final precision matrix Ω' is derived from Ω as follows: (a) if Ω[i,j]=0, then Ω'[i,j] = 0. (b) if Ω[i,j] ≠ 0, then Ω'[i,j] = Ω[i,j] if the entry [i,j] is non-zero in at least f*r precision matrices estimated from the bootstrap datasets. Otherwise Ω'[i,j]=0. Thus, Ω' is at least as sparse as Ω. Partial Correlation matrix, *P,* was calculated as:

$$P_{[i,j]} = \frac{-\Omega'_{[i,j]}}{\sqrt{\Omega'_{[i,i]} \times \Omega'_{[j,j]}}}$$

The value f is a preset threshold (0 ≤ f ≤ 1). We used r = 50 (bootstrap replicates) and f = 0.8 (e.g. association must be non-zero ≥80% of the time) in our analysis. Partial correlation matrices were parsed using python and all associations below a magnitude of .01 were considered statistical noise and removed.

Network Property, Clique, and Module Analysis

For each cohort network, the following properties were computed using NetworkX[84] (version 2.4): average shortest path length (ASPL), transitivity, modularity, degree assortativity, degree centrality, betweenness centrality, and genera assortativity. The ASPL (α) is defined as:

$$\alpha = \Sigma_{s,t \in V} \frac{D[s,t]}{n(n-1)}$$

where V is the set of nodes in the graph (G), D[s,t] is the shortest path from s to t, and n is the total number of nodes in G (11,12). The transitivity (T) of a network is the fraction of all possible triangles present in the graph, and is defined as:

$$T = 3\frac{trianlges}{triads}$$

triangles are a clique (a subset of nodes within a network where each node is adjacent to all other nodes within the subset) of three nodes, and triads are the count of connected triples (three nodes xyz with edges (x,y) and (y,z) where the edge (x,z) can be present or absent)[84,85]. Modularity (Q) is defined as:

$$Q = \frac{1}{2m}\Sigma_{i,j}\Sigma_{i,j}\left(A[i,j]-\frac{k_i k_j}{2m}\right)\delta(C_i,C_j)$$

where A is the adjacency matrix of graph (G), m is the total number of edges, k_i is the degree of node i, and $\delta(C_i,C_j)$ is 1 if i and j (node pair) are in the same community or 0 if in different communities[85,86]. Assortative mixing is a predilection of nodes to form connections with other nodes that are like (assortative) or unlike (disassortative) themselves. We measured node mixing preference according to node degree (degree assortativity) and node genus classification (genera assortativity). Degree assortativity is calculated using the standard Pearson correlation coefficient:

$$r = \frac{\sum_{xy} x\, y\left(D[x,y] - a_x b_y\right)}{\sigma_a \sigma_b}$$

Where D is the joint probability distribution matrix, D[x,y] is the fraction of all edges in the graph that connects nodes with degree values x and y, a_x and b_y are the

fraction of edges that start and end at nodes with values x and y, and σ_a and σ_b are the standard deviations of the distributions a_x and b_y. The value of r can be any value between –1 (perfect disassortativity) and 1 (perfect assortativity)(14). Genera assortativity is defined as:

$$r = \frac{TrQ - \|Q^2\|}{1 - \|Q^2\|}$$

Where Q is the joint probability distribution matrix whose elements are Q[i,j] (the fraction of all edges in the graph that connects nodes of genus type i to genus type j), Tr is the trace of the matrix Q**,** and ||Q|| signifies the sum of all elements of the matrix Q[87].

Modules within each network were found utilizing the *label_propogation_communities* algorithm, based on the asynchronous label propagation algorithm (aLPA)[42] from NetworkX. To quantify the ability of the aLPA to partition the data, we utilized the *performance* function NetworkX. Performance (p) is defined as:

$$p = \frac{a + b}{c}$$

where a is the total intra-module edges, b is the total inter-module non-edges, and c is the total potential edges[89]. Monte Carlo simulations were utilized to test for statistical significance of network property differences (see supplemental). Three member cliques and modules within each network were found using NetworkX. Module functional profiles (MFP) were created by aggregating the functional profiles of species contained within each module.

Network Node Centrality Analysis

Degree centrality is defined as the degree (total edges) of a node. The node within each network module exhibiting the highest degree centrality was designated as a module “hub”. If two or more species were found to have equal degree centrality then centrality measurements of those nodes were re-computed in context of the entire network. The top ten nodes exhibiting the highest betweenness centrality within each network were designated as “bottlenecks”. To find “bottleneck” species, betweenness centrality was computed for each node. Betweenness centrality is defined as:

$$C_B(u) = \sum_{s,t \in V} \frac{\sigma(s,t|u)}{\sigma(s,t)}$$

where the betweenness centrality of a node (υ) is the sum of the fraction of all-pairs shortest paths that pass through υ, V is the set of all nodes, σ(s,t) is the number of shortest paths (s,t)-paths, and σ(s,t| υ) is the number of those paths passing through node υ other than s,t[90].

Figures

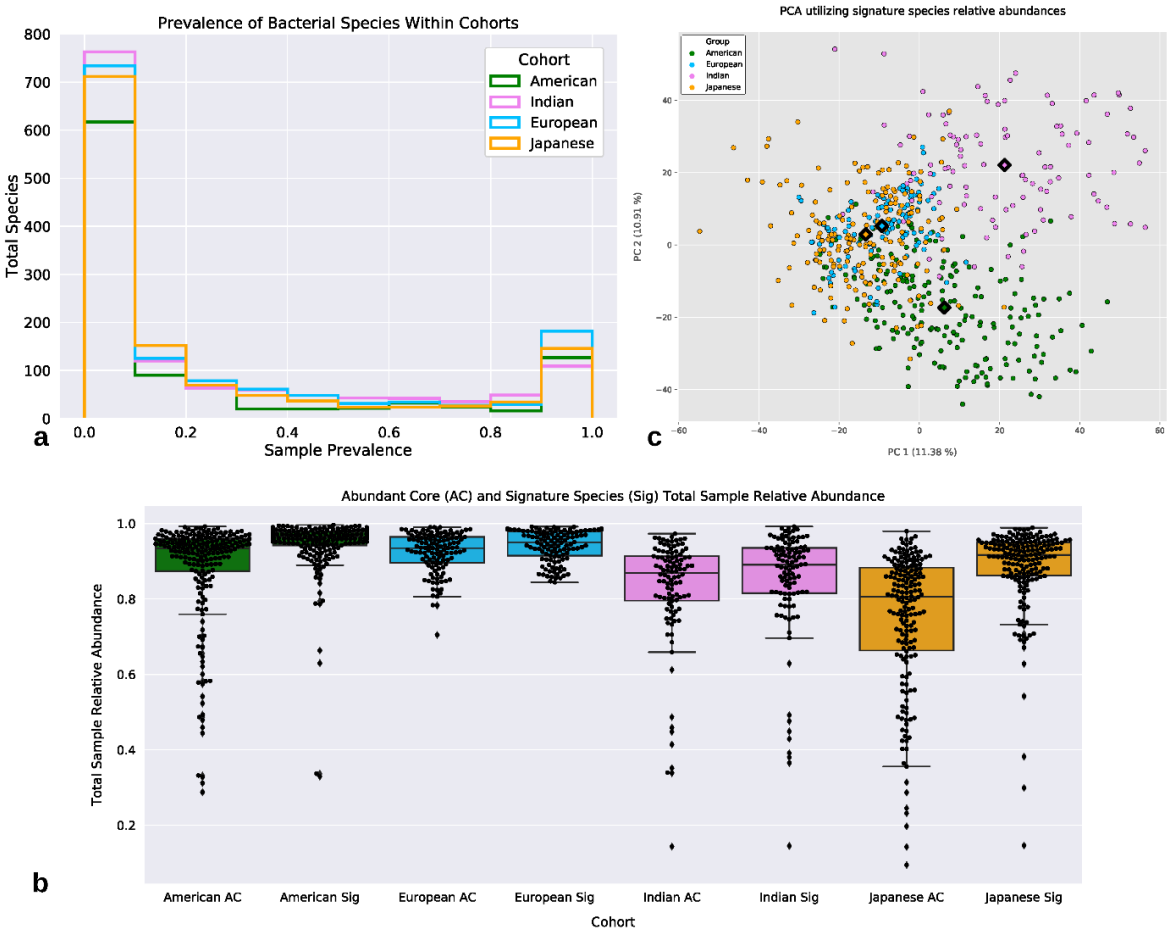
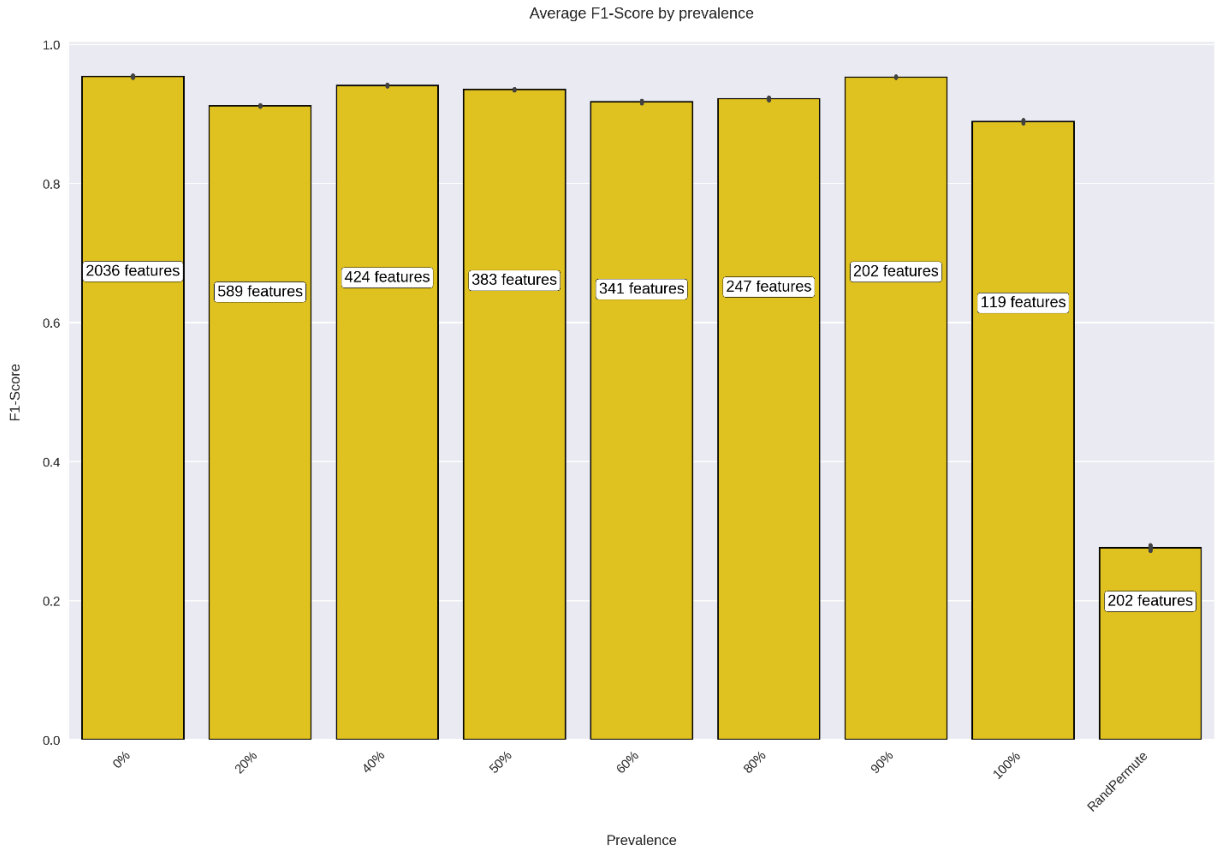


Figure 1: 'Abundant cores' and Signature Species.

a. *All cohorts exhibit a bimodal distribution for species prevalence. Species that are prevalent in 90% or more samples within a cohort is considered a member of that cohort's 'abundant core'.* ***b.*** *The proportion of total sample relative abundance each cohort's 'abundant core' species and the union of all 'abundant cores' species (i.e. Signature Species/Sig). The 'abundant core' microbiota is shown to account for the bulk of reads mapped within each sample. Each dot represents a sample from that cohort.* ***c.*** *PCA demonstrating the lack of distinct clustering of samples from different cohorts based on CLR-transformed relative abundance data of the signature species. Samples from the Indian and American cohorts appear to separate from the rest of the cohorts however, samples from the other two cohorts demonstrate little separation. The diamonds indicate cluster centroids.*

a.


Average F1-Score by prevalence
F1-Score
1.0
0.8
0.6
0.4
0.2
0.0
2036 features
589 features
424 features
383 features
341 features
247 features
202 features
119 features
202 features
0%
20%
40%
50%
60%
80%
90%
100%
RandPermute
Prevalence

b.

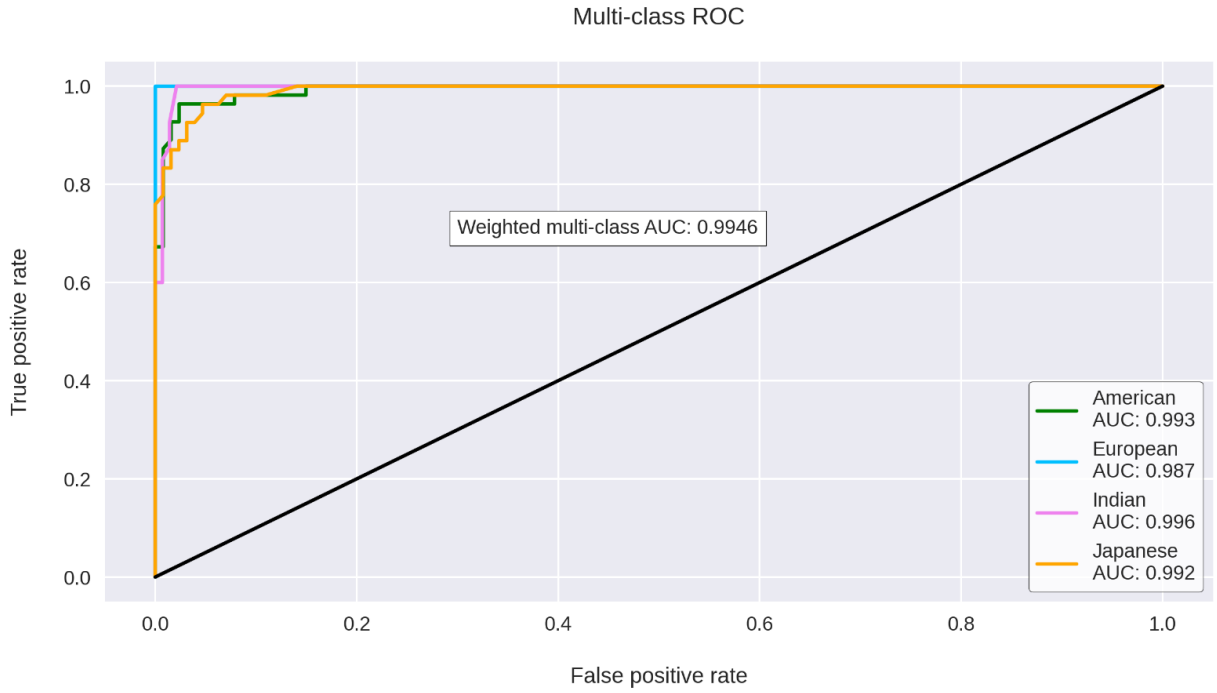
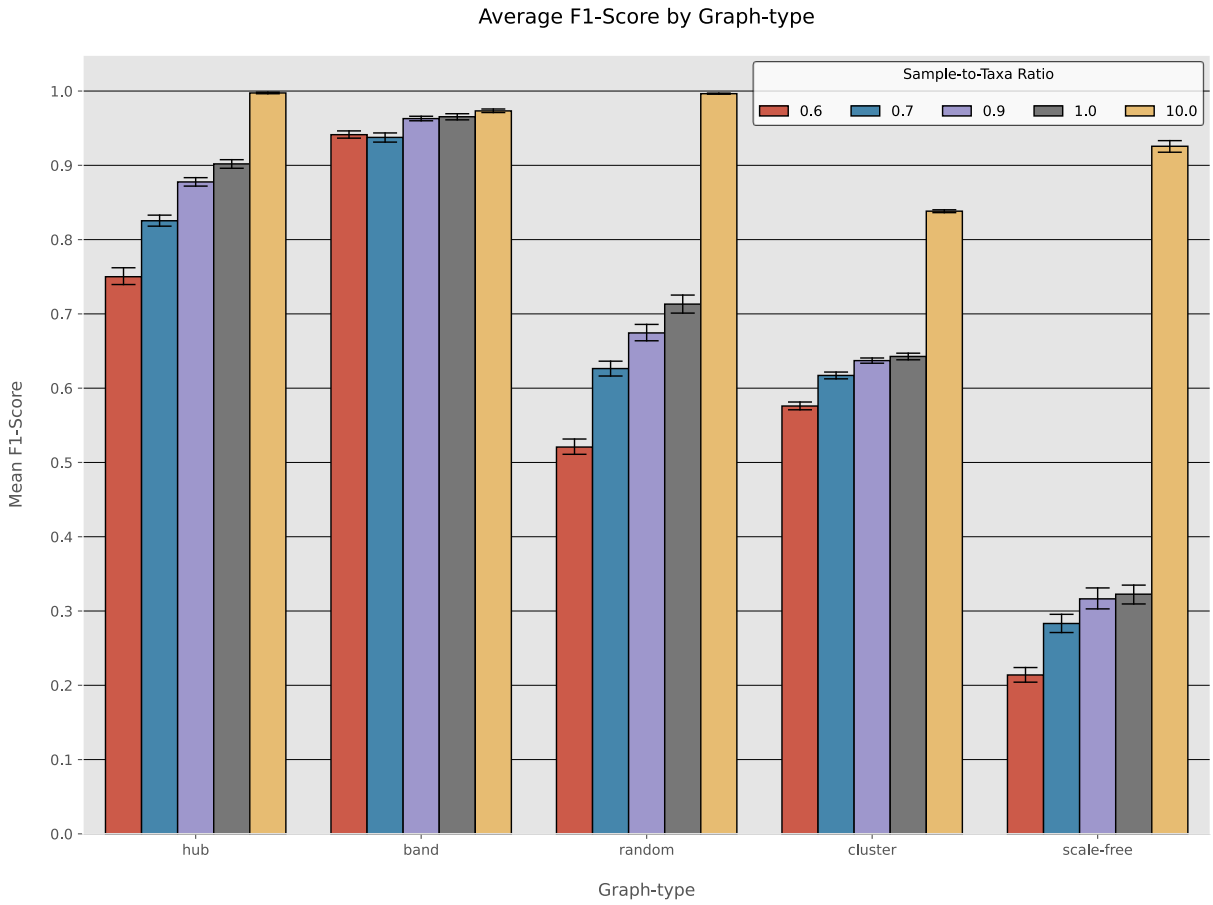


Figure 2: Effect of prevalence thresholds on RFC accuracy.

a. *The 90% bacterial prevalence threshold enables the most accurate distinction between cohorts. Bacterial species used for RFC-based classification were determined by prevalence of bacteria in the samples. The 90% prevalence threshold offers slightly better ability to distinguish between the cohorts based on their taxonomic profiles while removing over 1,800 features. The 90% prevalence threshold was then randomly permuted (RandPermute) and added to the plot as a reference. Utilizing only species that were present in 100% of samples led to diminished accuracy while removing relatively few features.* ***b.*** *Multi-class Receiver Operator Characteristic (mROC) graph was created for each cohort. Each cohort displayed a large Area Under the Curve (AUC) indicating that the model was able to accurately distinguish the different cohorts from each other using the taxonomic profiles alone. The multi-class AUC was weighted by sample size for each cohort.*

a.


Average F1-Score by Graph-type
Sample-to-Taxa Ratio
0.6
0.7
0.9
1.0
10.0
Mean F1-Score
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
hub
band
random
cluster
scale-free
Graph-type

b.

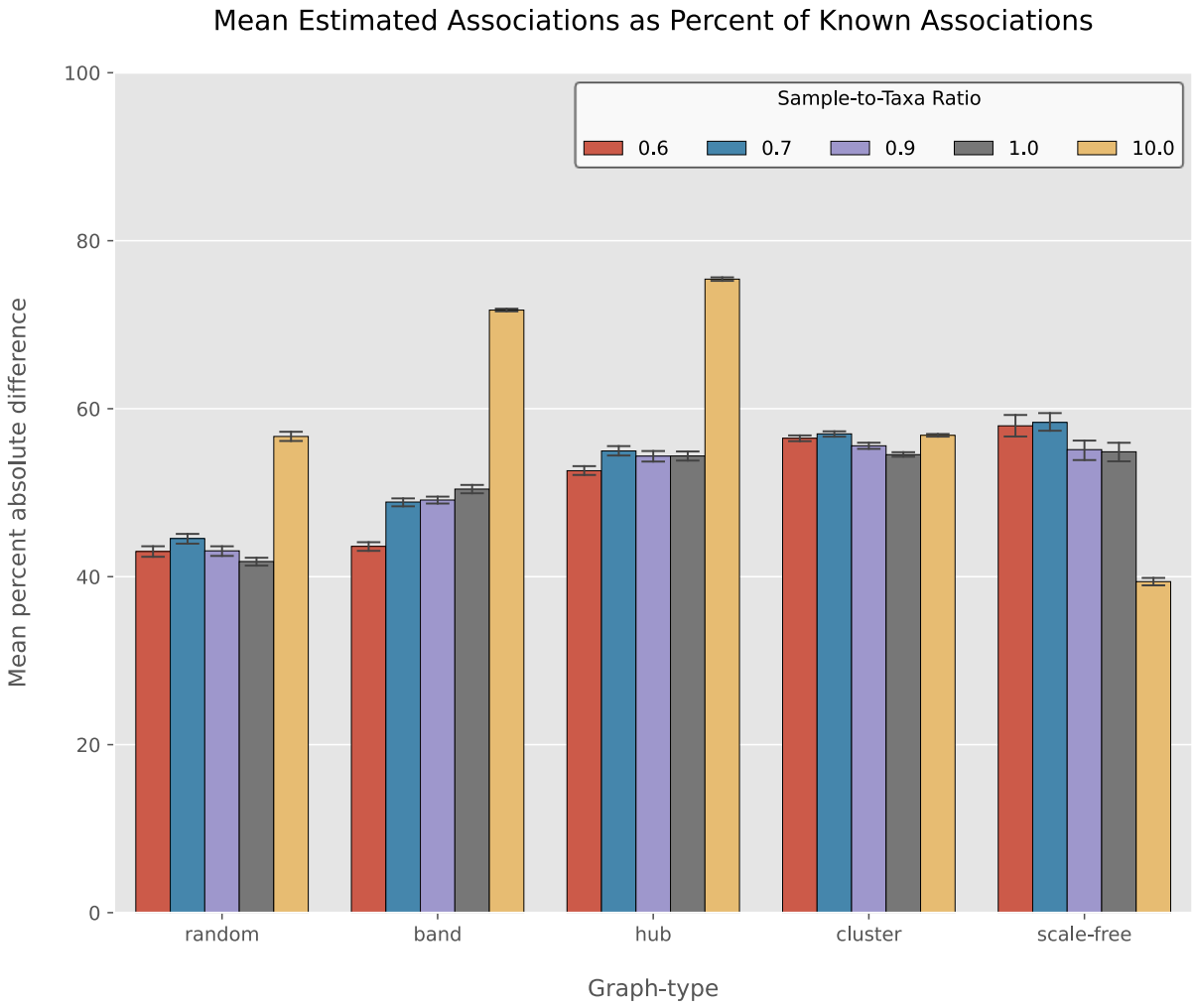


Figure 3: GGM algorithm benchmarking

Average F1-scores of the GGM algorithm for various graph-types and sample-to-taxa ratios. Synthetic data was modeled on the CLR-transformed means and sample-to-taxa ratios present in the real data sets. A sample-to-taxa ratio of 10 was added to demon-strate the effect adding additional samples has on accuracy of GGM. ***a.*** *The average F1-score for all graph-types is 0.74. The hub and band networks consistently exhibit the highest accuracy. An overt increase in accuracy is demonstrated as the sample-to-taxa ratio increases for all graph-types, with no graph-type have an F1-score <0.9 at a sam-ple-to-taxa ratio of 10.* ***b.*** *GGM consistently underestimates magnitude of associations. As sample-to-taxa ratio increases, there is an appreciable increase in the accuracy of association magnitude estimation in all, but the cluster and scale-free, graph-types.*

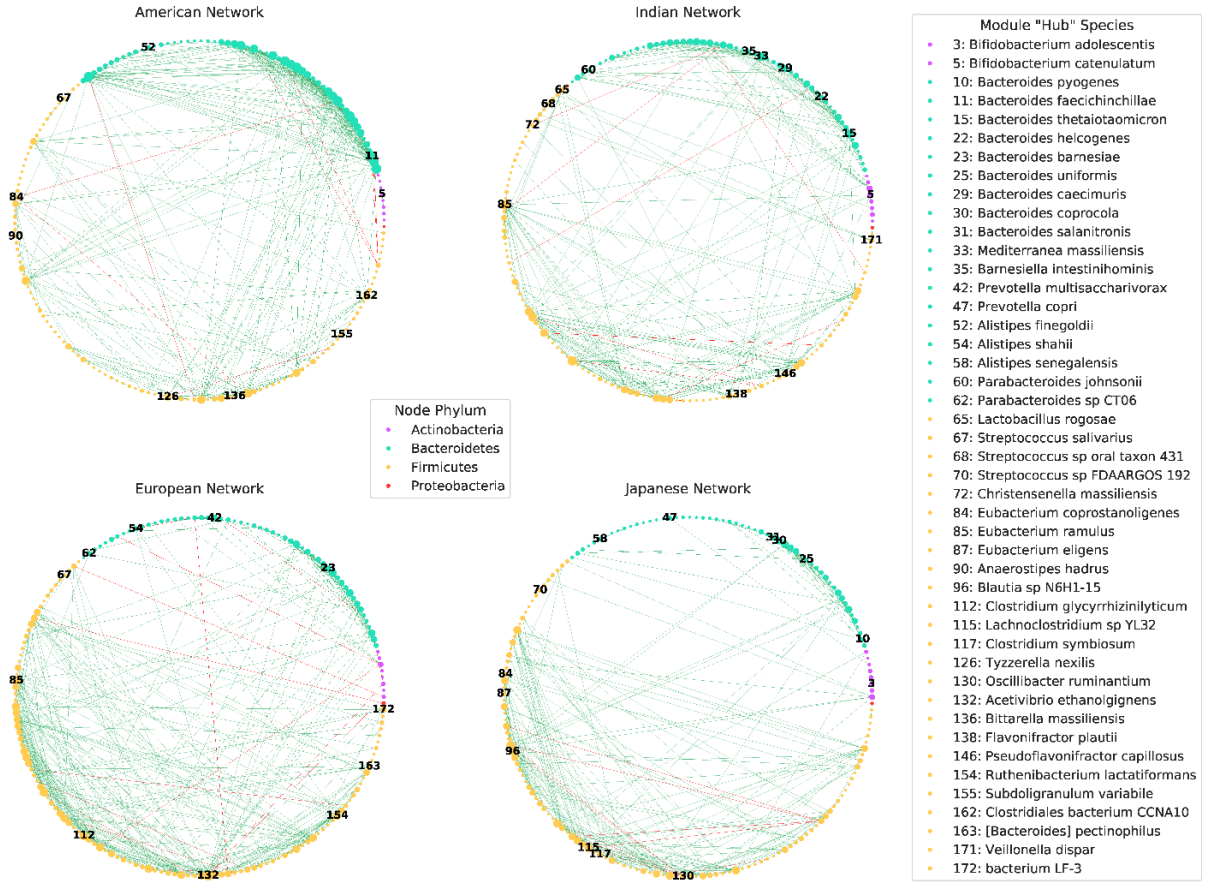


Figure 4: Species-level bacterial association networks by cohort.

Network modeling of associations between (173/202) signature species within each network. A total of 29 species were not shown as they had zero edges in all networks. Node color designates the phylum each species belongs to, node size is reflective of node degree, and edge color represents if the association is positive (green) or negative (red). Nodes are ordered counterclockwise around the circle by the alphabetical order of the concatenated string of all taxonomic levels. Nodes that are numbered correspond to species with the highest degree centrality within modules, designated as "hubs". Brackets around [Bacteroides] pectinophilus indicate that it is misclassified (i.e. placed incorrectly in a higher taxonomic rank and awaiting to be formally renamed). We utilized Blast to designate [Bacteroides] pectinophilus as belonging to the phylum Firmicutes91. For a full list of species shown and not shown within network models see supplemental.

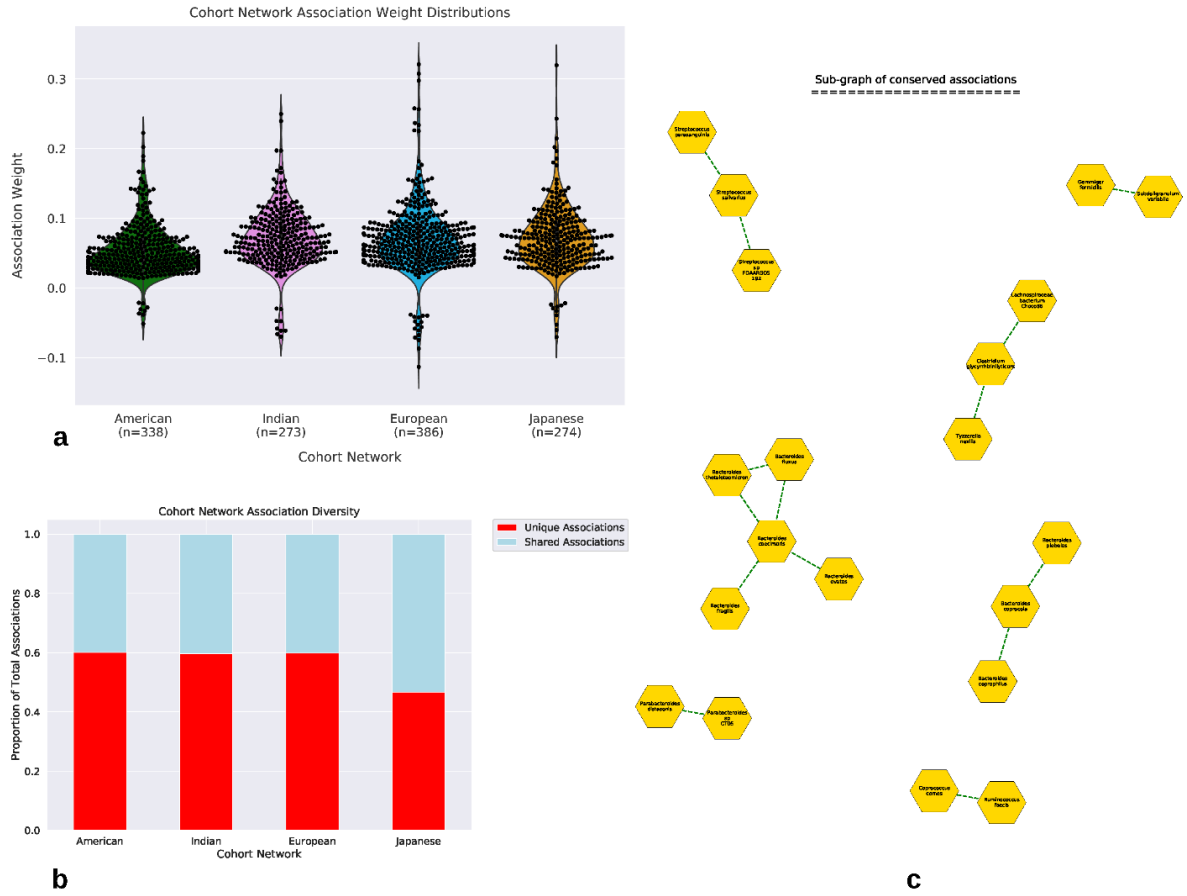


Figure 5: Cohort network association analysis.

a. The distribution of bacterial association weights within each cohort's network, dots and (n) represent total associations. b. The proportion of associations within each cohort's network that are unique (red) or shared (blue) with at least one other network. c. Sub-graph displaying only the 20 conserved nodes (species) and 14 edges (associations) retained across all cohorts.

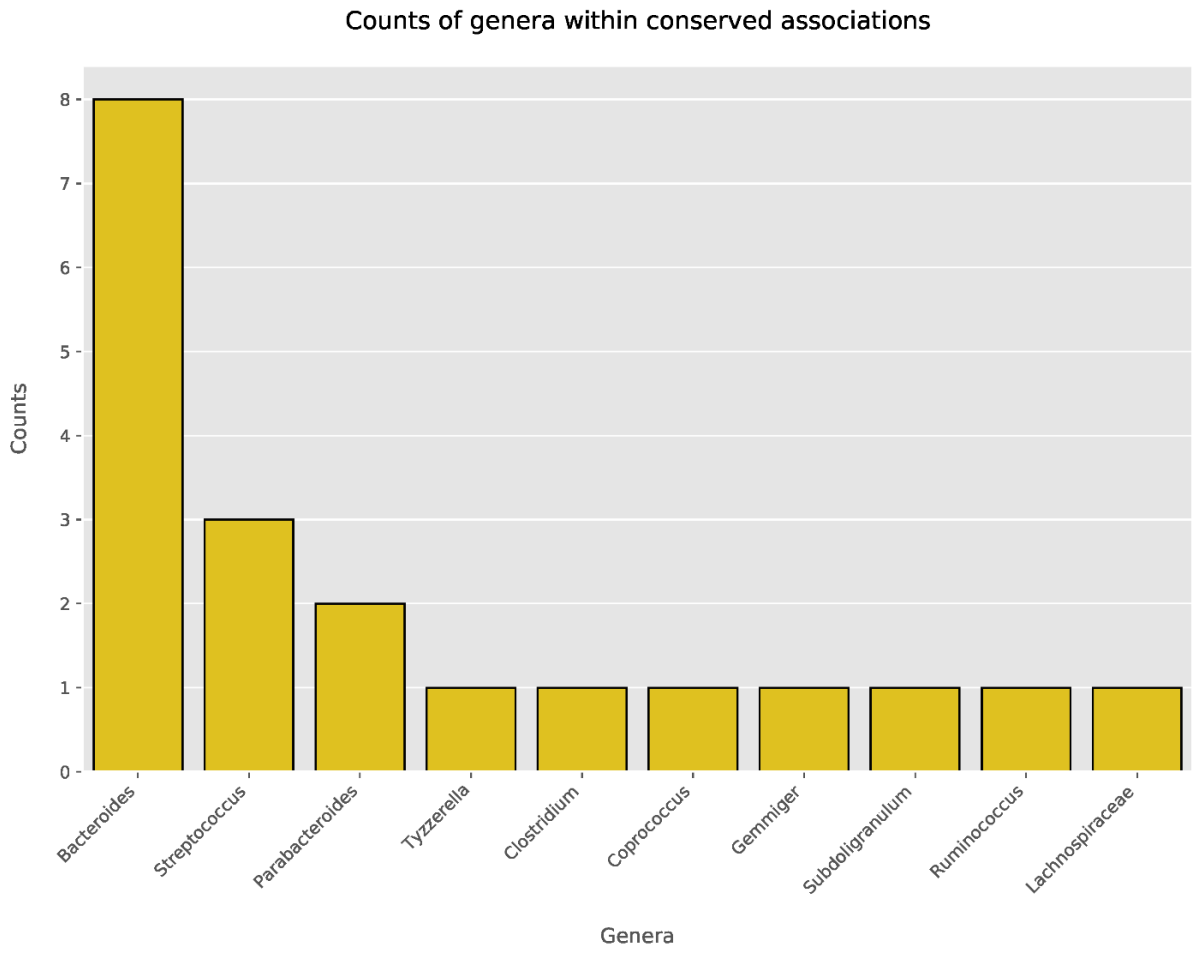


Figure 6: Conserved genera counts.

When examining the networks of all cohorts, there were 14 conserved associations comprised of 20 bacterial species. Species of the Bacteroides genus were the most abundant constituents of the bacterial associations conserved within all cohorts.

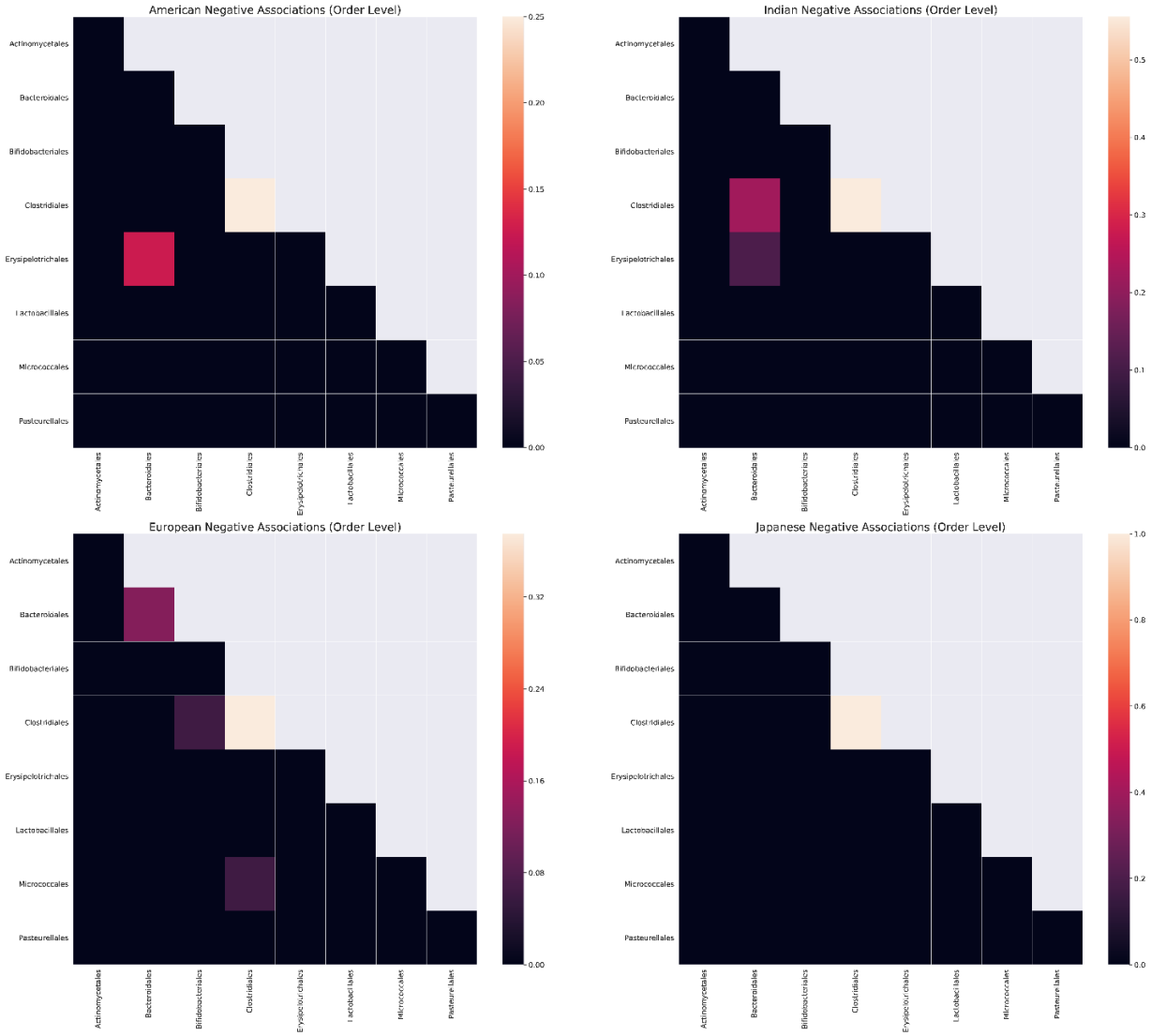


Figure 7: Cohort negative association heatmap.

Heatmaps of the proportion of total negative associations within each cohort's network that order member species were found to be involved in. Within each cohort, negative associations appear to occur mainly between species from the order Clostridiales.

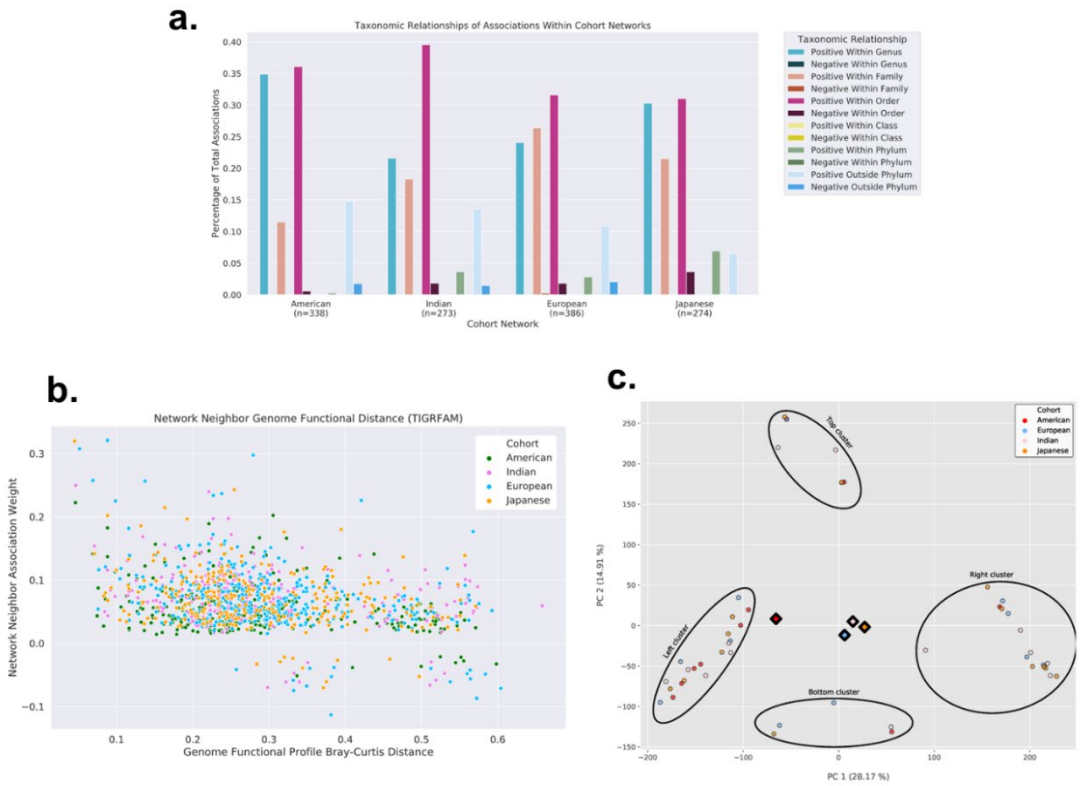


Figure 8: Taxonomic and functional relationships between species.

a. *Proportion of associations within each cohort's network that are either positive or negative at the lowest level of taxonomic relation (n = total associations). Most positive associations appear between taxonomically similar species.* ***b.*** *Association weight vs Bray-Curtis distance of genome functional profiles between network partners. Positive associations between functionally similar species are both common and greater in strength than negative associations. There appears to be a minimal distance between genome functional profiles before a negative association is demonstrated.* ***c.*** *An asynchronous LPA was used to analyze the modules composing the networks of each cohort. Four distinct clusters were found, and each cohort was represented within each cluster. The American cohort appears to be biased towards the Cluster IV, however the other cohorts do not appear overtly biased to any one cluster. Each dot represents the aggregated TIGRFAM profiles of an individual module found by aLPA and the diamonds represent the cohort centroids.*

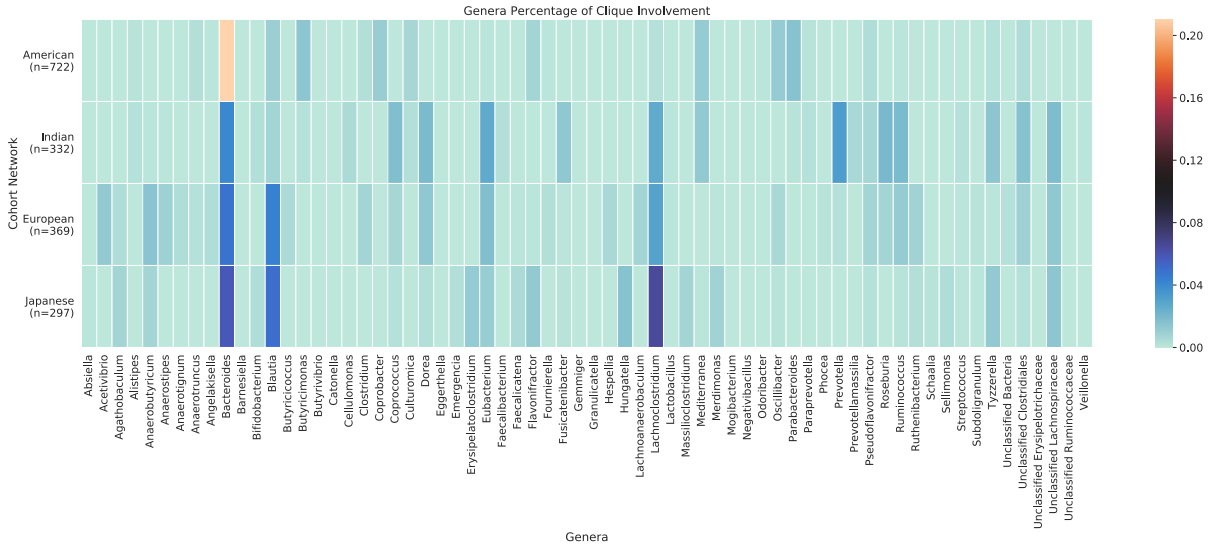


Figure 9: Genera involvement in clique formation as a percentage.

Heatmap of the proportion of total cliques found within each cohort's network that genera member species were found to be involved in (n = total 3 member cliques). Species from the genus Bacteroides tend to be found in the majority of cliques across all cohorts.

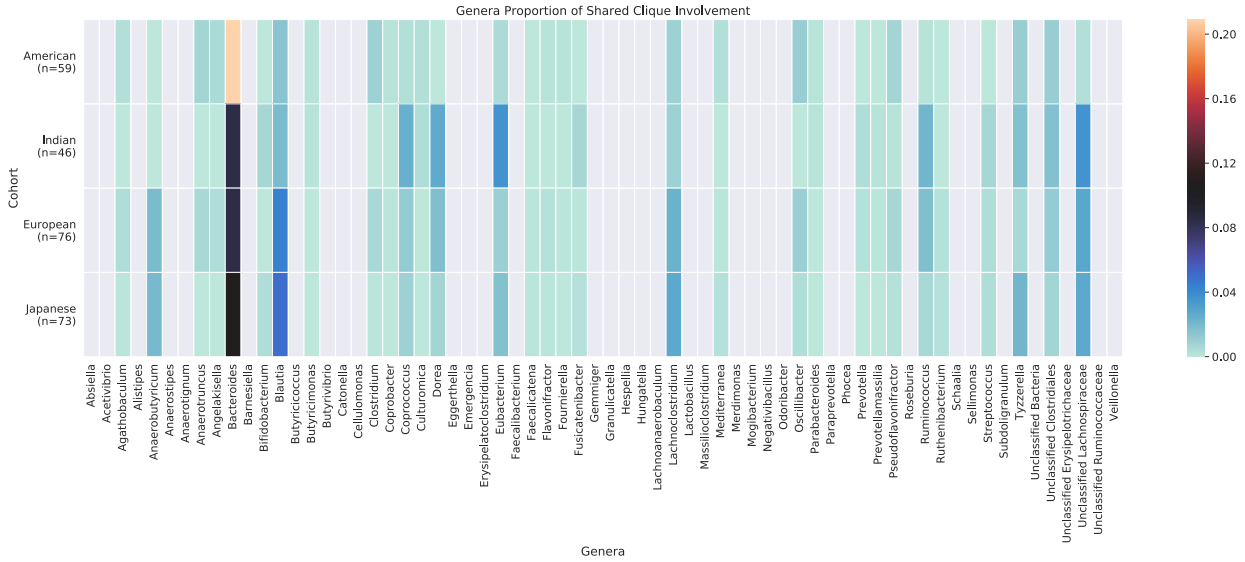


Figure 10: Proportion of genera shared in cliques.

Heatmap of cliques that were retained in at least one other network. Cliques that Bacteroides sp. are involved in are highly re-tained across networks.

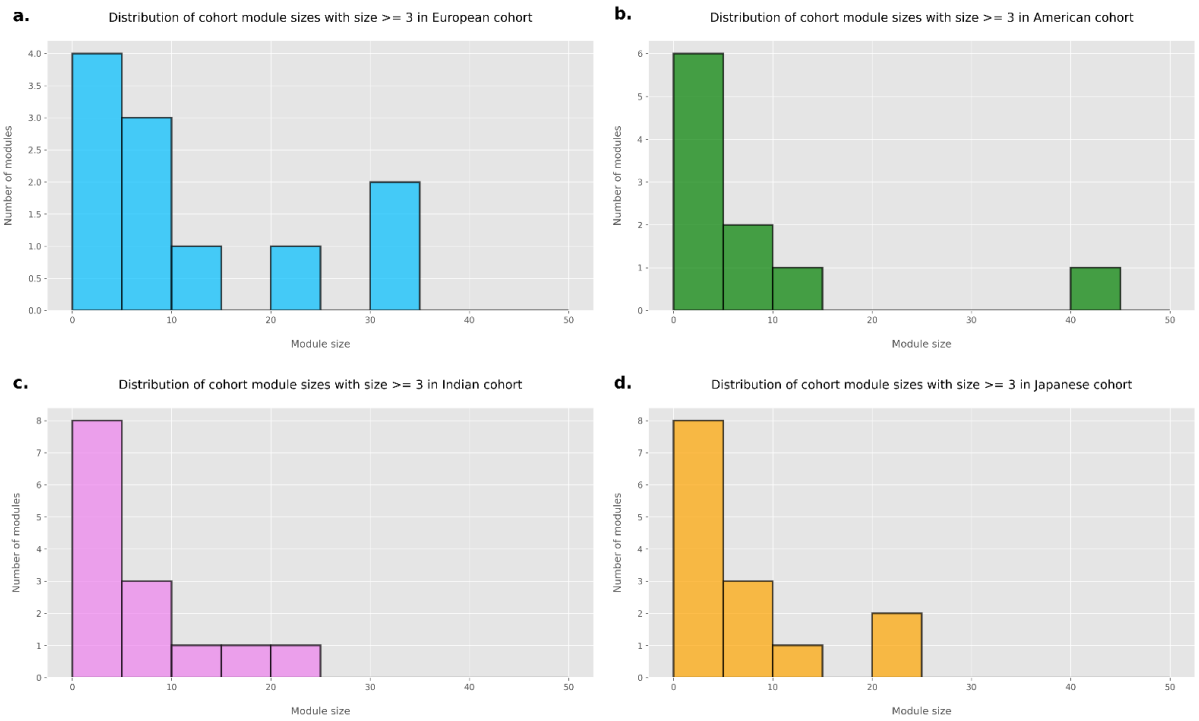


Figure 11: Distribution of module sizes.

Distribution of module sizes found by asynchronous LPA, colored by cohort. ***a.*** *Distribution of module sizes within the European cohort.* ***b.*** *Distribution of module sizes within the American cohort.* ***c.*** *Distribution of module sizes within the Indian cohort.* ***d.*** *Distribution of module sizes within the Japanese cohort.*

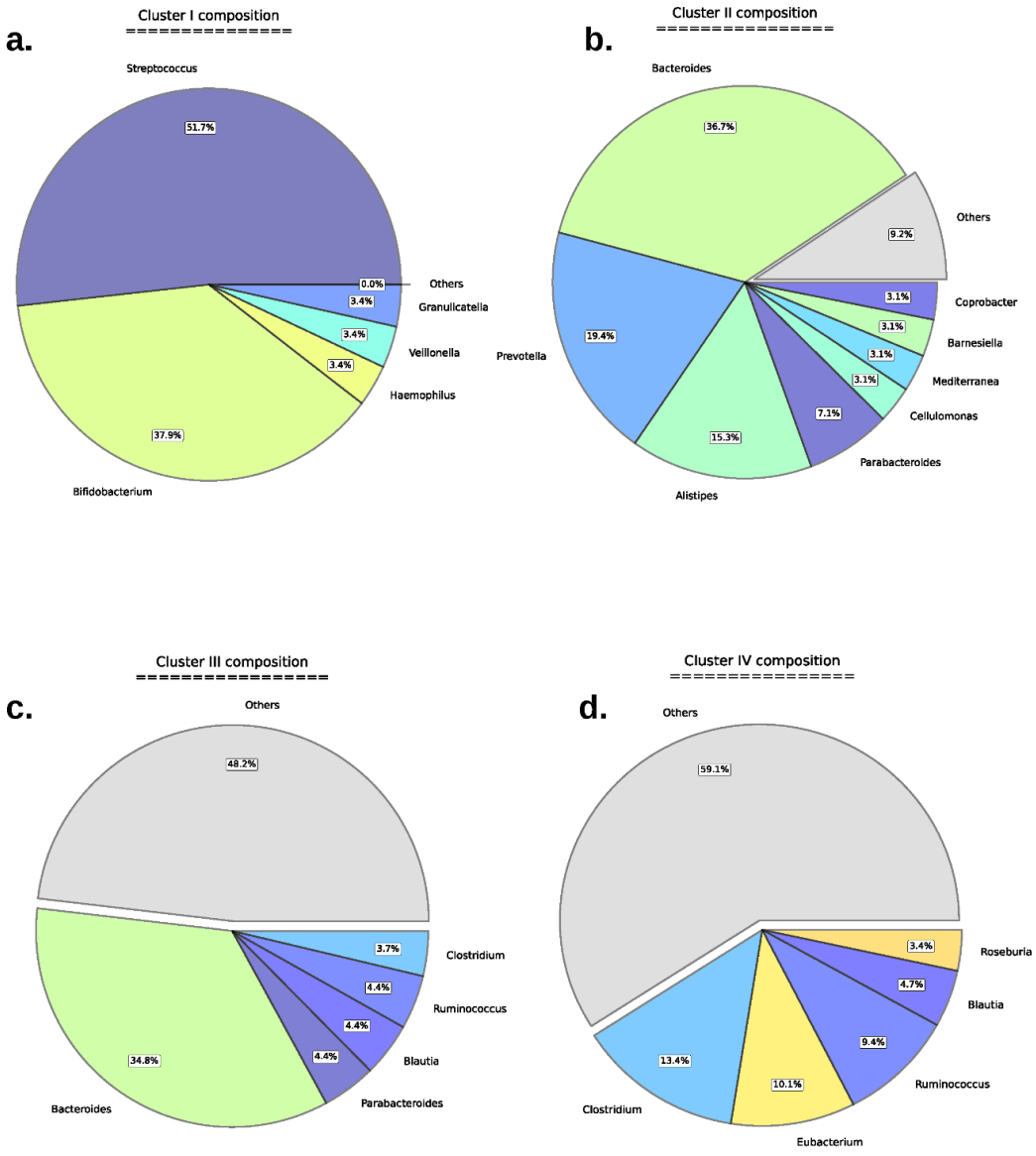


Figure 12: Pie charts of cluster taxonomy.

Pie plots demonstrating genus-level taxonomic compositions within each of the module clusters. Clusters were determined using PCA of module functional profiles for each module. ***a****. Cluster I is dominated by members of the Streptococcus and Bifidobacterium genera and no genus*

represents less than 3% relative abundance. ***b.*** *Members of the Bacteroides genus are the most abundant in the Cluster III and there are 49 genera with relative abundances below 3%.* ***c.*** *Members of the Bacteroides genus are also the most abundant in the Cluster II, however the Prevotella and Allistipes genera are also abundant and account for >70% of abundance when combined with Bacteroides. There are 6 genera with relative abundances below 3%.* ***d.*** *There are only 5 genera above 3% relative abundance and 44 genera below 3% with no one genus showing greater than 15% relative abundance. Genera with < 3% relative abundance were placed in the 'Others' category.*

Cluster I	
Function	**Status**
Degradation of polysaccharides	Elevated
Central intermediary metabolism::Other	Elevated
Toxin productions and resistance	Elevated
Aerobic metabolism	Elevated
Nucleic acid metabolism	Elevated
DNA regulation	Elevated
Peptide secretion and trafficking	Elevated
Thiamine biosysnthesis	Reduced
Cluster II	
Function	**Status**
Cellular processes::Other	Elevated
Biosynthesis and degradation of surface polysaccharides and lipopolysaccharides	Elevated
Lipoate biosynthesis	Elevated
Biosysnthesis of menaquinone and ubiquinone	Elevated
Methanogenesis	Elevated
Degradation of proteins, peptides, and glycopeptides	Elevated
One-carbon metabolism	Elevated
Transposon functions	Reduced
Regulatory functions::Other	Reduced
Anion transport and binding	Reduced
Cell division	Reduced
Protein fate::Other	Reduced
Small molecule regulation	Reduced
DNA metabolism::Other	Reduced
Signal transduction:Other	Reduced
Cluster III	
Function	**Status**
Chemoautotrophy	Elevated
Sulfur metabolism	Elevated
Amino acid and amine metabolism	Elevated
Phosphorous metabolism	Elevated
Transport and binding proteins::Unknown substrate	Elevated
DNA metabolism::Restriction/modification	Elevated
Cluster IV	
Function	**Status**
Transcription Factors	Elevated
Adaptation to atypical conditions	Reduced

Figure 13: Functional role profile differences.

Tables illustrating relative differences in functional roles within the cohorts. Roles that were different signs (+/-) in one cohort relative to all other cohorts, were deemed different. If the sign was negative after CLR transformation, the role was considered reduced and if the sign was positive the role is considered elevated ***a.*** *The different clusters ap-pear to have overt functional differences possibly indicating the importance of the existence of modules from each cluster in a cohort for the healthy functioning of the gut microbiome.*

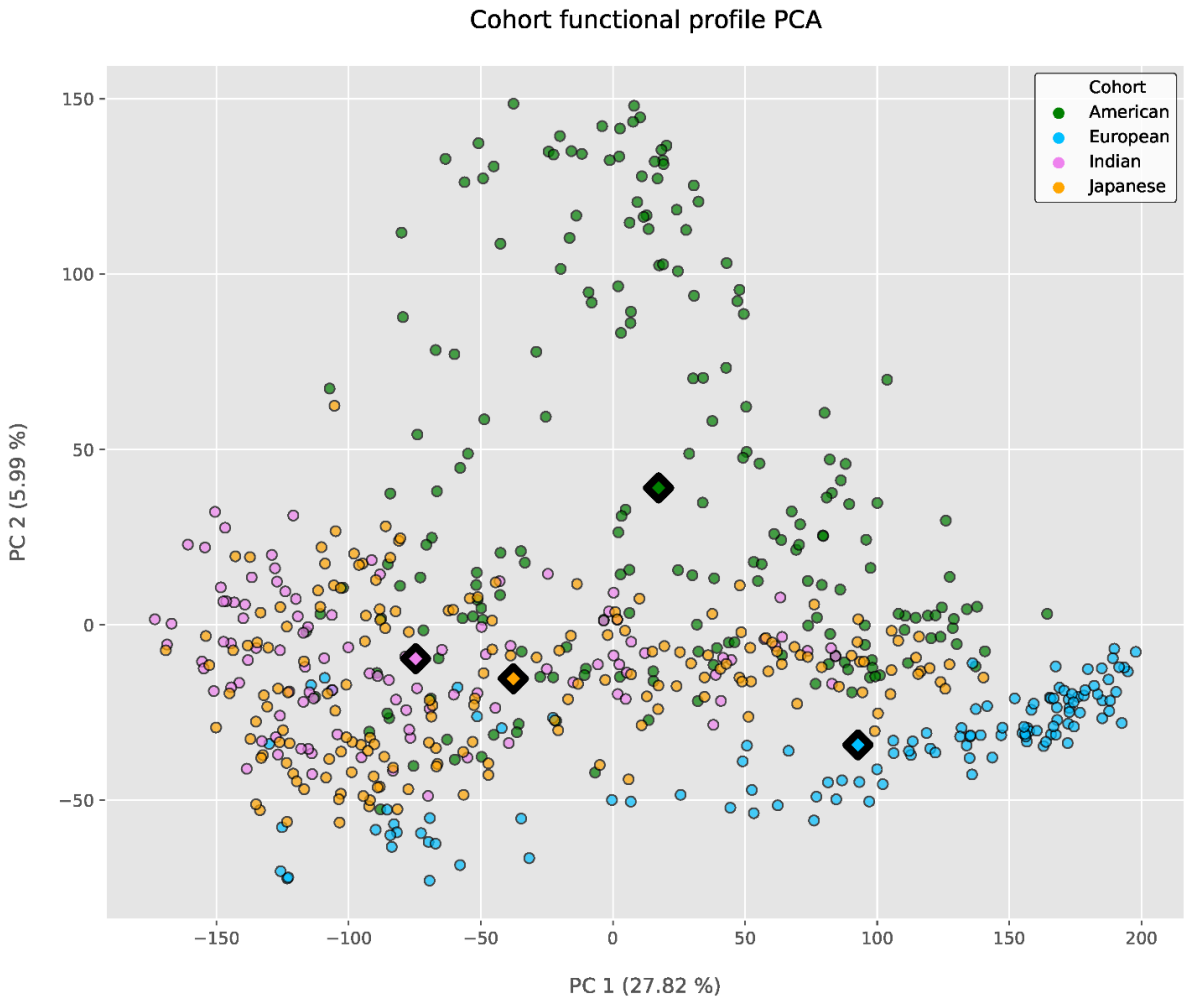


Figure 14: Cohort functional profile PCA.

PCA was performed by analyzing the aggregated cohort functional profiles of each cohort. The cohorts have a large amount of overlap and do not appear to distinctly separate.

American	
Function	**Status**
Energy metabolism::Amino acids and amines	Elevated
Mobile and extrachromosomal element functions::Transposon functions	Reduced
Indian	
Function	**Status**
Mobile and extrachromosomal element functions::Transposon functions	Elevated
European	
Function	**Status**
Protein fate::Protein and peptide secretion and trafficking	Elevated
Central intermediary metabolism::Sulfur metabolism	Reduced
Central intermediary metabolism::Other	Reduced
Biosynthesis of cofactors, prosthetic groups, and carriers::Other	Reduced
Transport and binding proteins::Amino acids, peptides and amines:	Reduced
Cellular processes::Detoxification	Reduced
Signal transduction::Two-component systems	Reduced
Japanese	
Function	**Status**
Fatty acid and phospholipid metabolism::Other	Reduced

Figure 15: Cohort functional profiles.

Few functional role differences were demonstrated between the different cohorts as only the American and European cohorts had more than one difference and only the European cohort demonstrated greater than two differences.

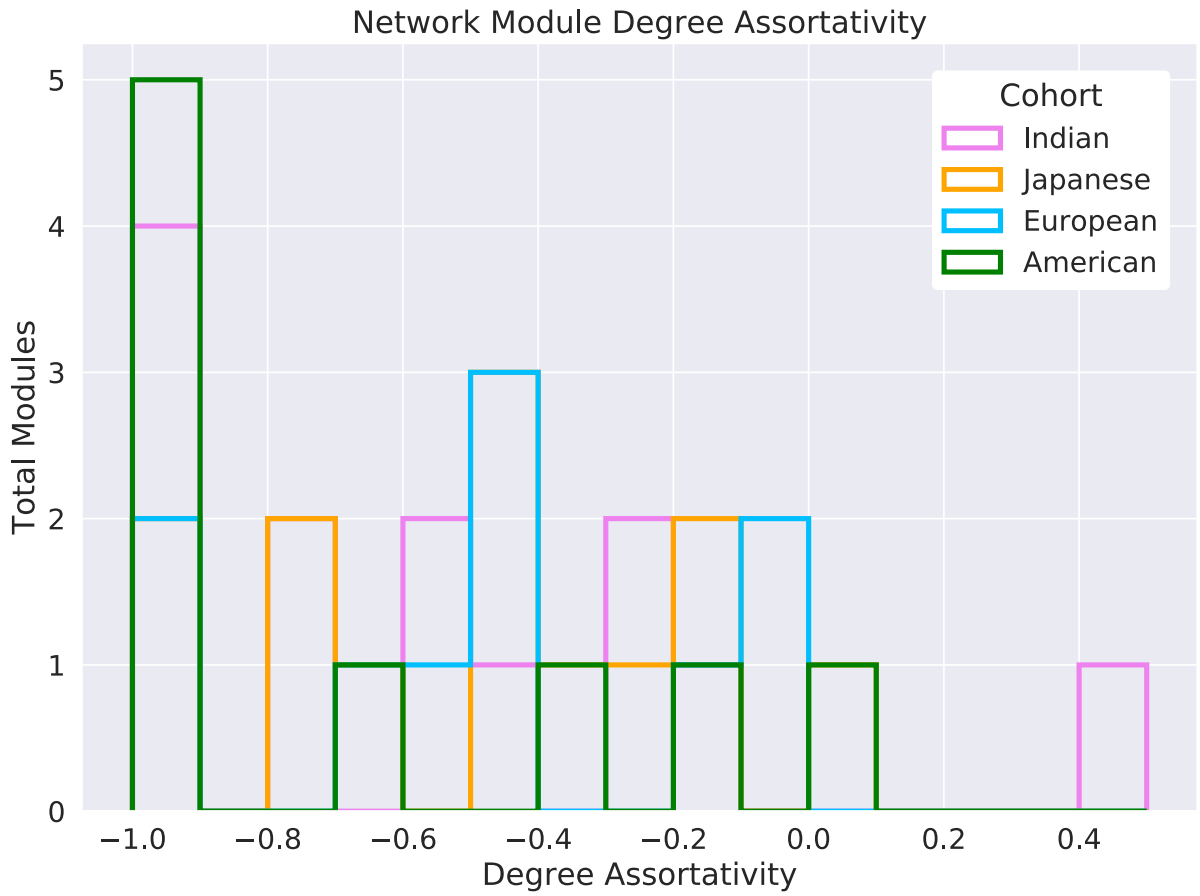
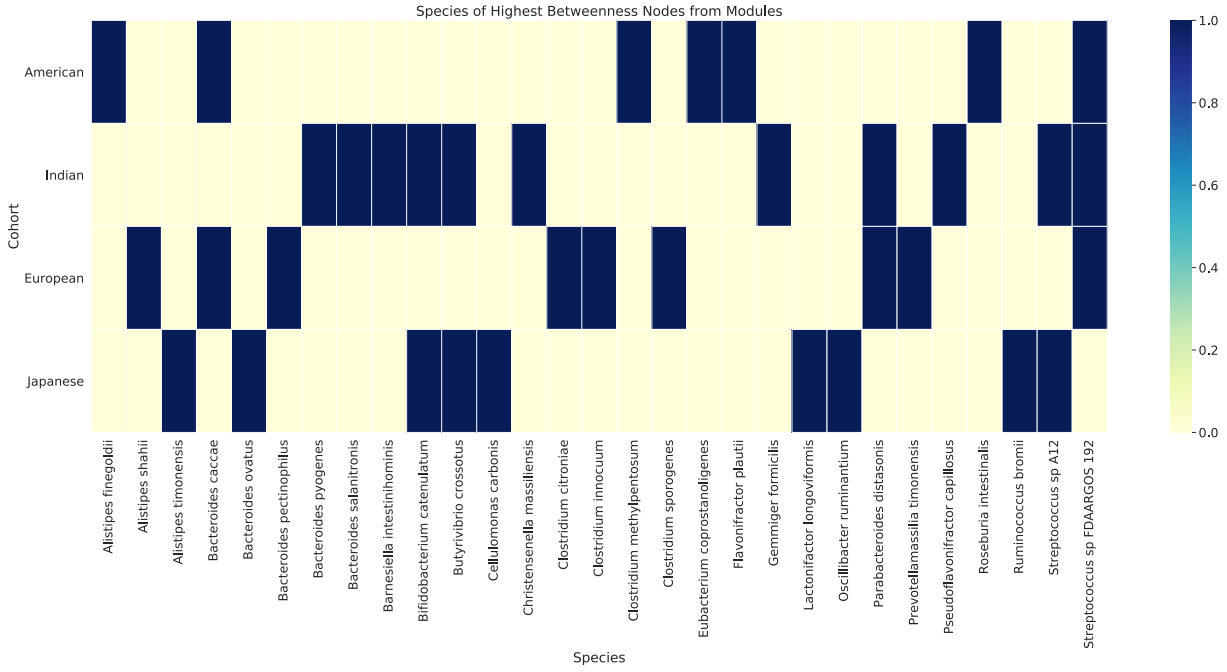


Figure 16: Degree assortativity of modules.

Distribution of the degree assortativity of modules within cohort networks. Most modules were disassortative in respect to their degree assortativity hinting at "hub" species existing within modules.

a.


Species of Highest Betweenness Nodes from Modules
Cohort
American
Indian
European
Japanese
Species
Alistipes finegoldii
Alistipes shahii
Alistipes timonensis
Bacteroides caccae
Bacteroides ovatus
Bacteroides pectinophilus
Bacteroides pyogenes
Bacteroides salanitronis
Barnesiella intestinihominis
Bifidobacterium catenulatum
Butyrivibrio crossotus
Cellulomonas carbonis
Christensenella massiliensis
Clostridium citroniae
Clostridium innocuum
Clostridium methylpentosum
Clostridium sporogenes
Eubacterium coprostanoligenes
Flavonifractor plautii
Gemmiger formicilis
Lactonifactor longoviformis
Oscillibacter ruminantium
Parabacteroides distasonis
Prevotellamassilia timonensis
Pseudoflavonifractor capillosus
Roseburia intestinalis
Ruminococcus bromii
Streptococcus sp A12
Streptococcus sp FDAARGOS 192
1.0
0.8
0.6
0.4
0.2
0.0

b.

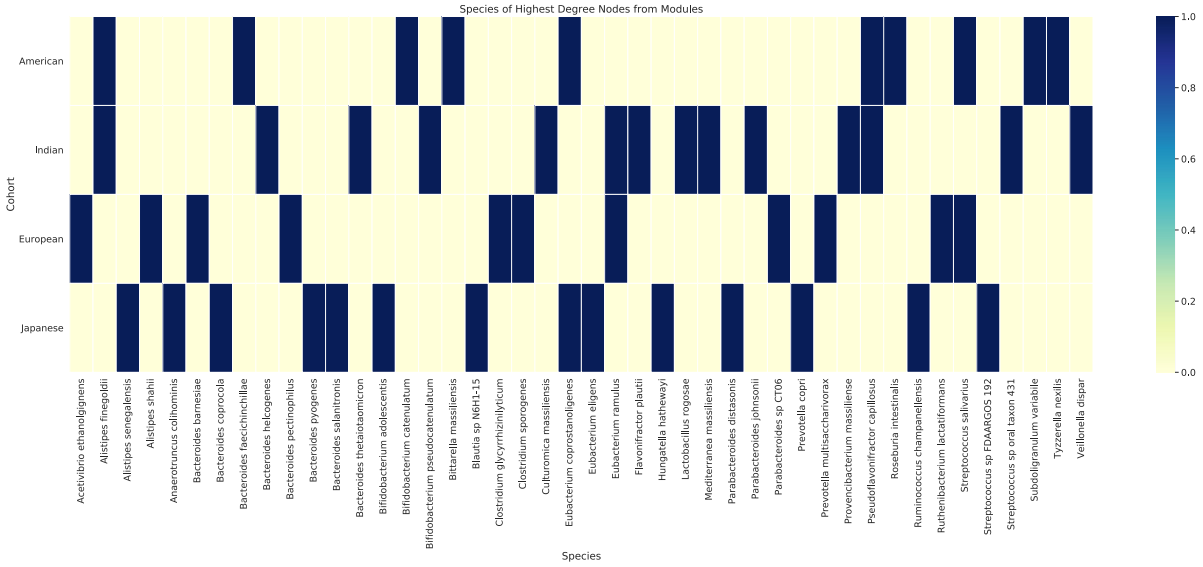


Figure 17: Cohort network "hubs" and "bottlenecks"

a. *Species with highest degree centrality are designated as "hubs."* ***b.*** *Species with highest betweenness centrality are designated as "bottlenecks."*

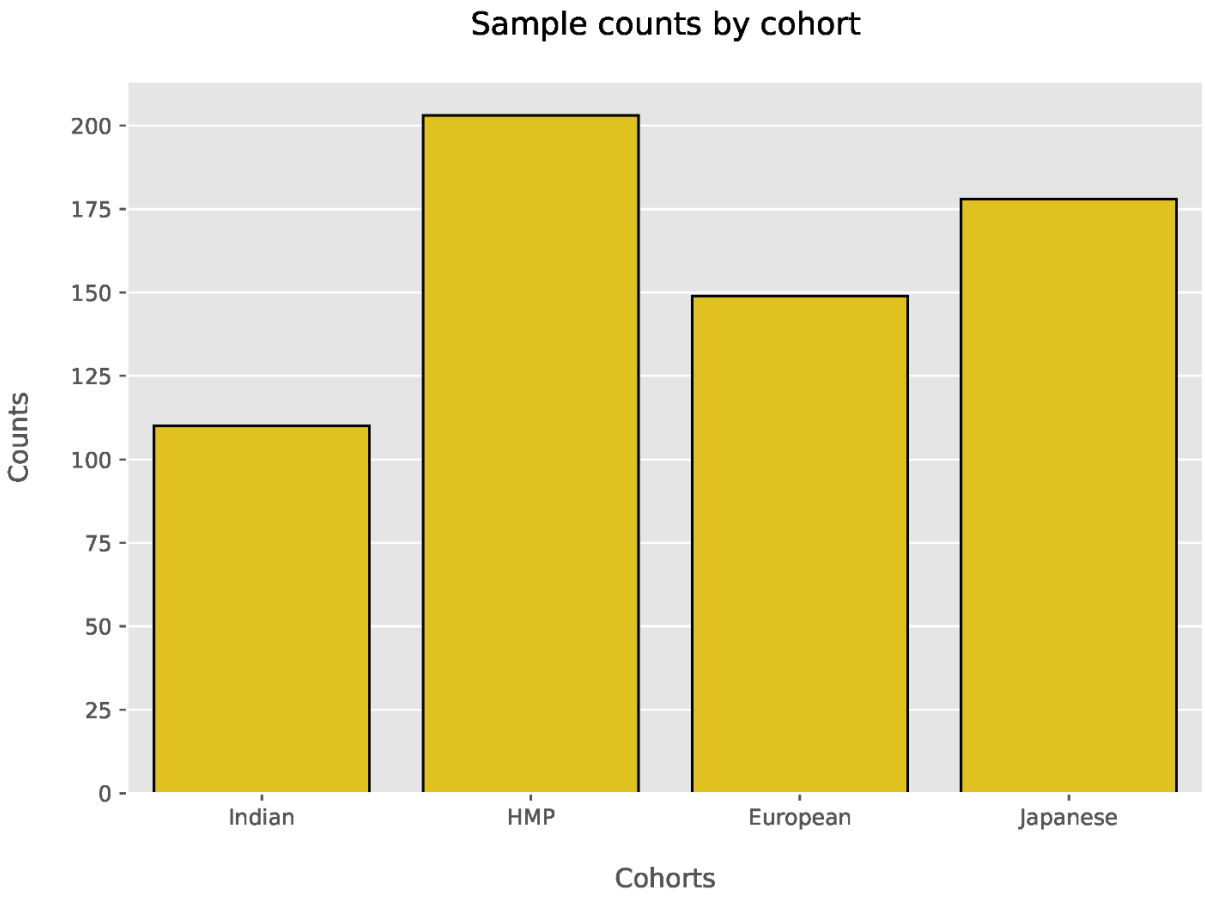


Figure 18: Sample counts in each cohort.

A bar plot representing the counts of samples from each cohort.

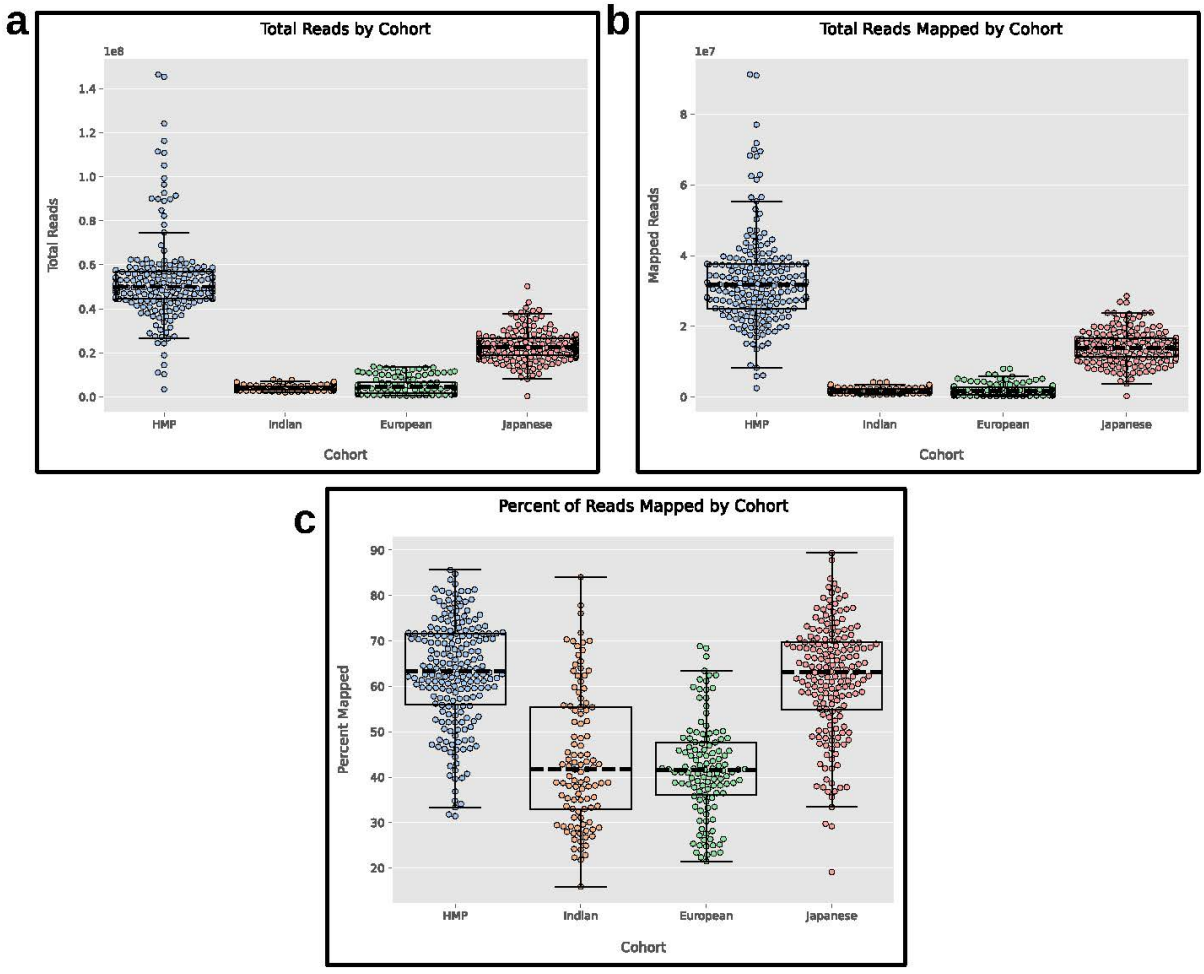


Figure 19: Read statistics by cohort.

a. *Each dot represents the total reads in an individual sample. The dashed black line in each box-plot represents the median reads of the cohort.* ***b.*** *Each dot represents the mapped reads in an individual sample. The dashed black line in each box-plot represents the median mapped reads of the cohort.* ***c.*** *Each dot represents the percent mapped reads in an individual sample. The dashed black line in each box-plot represents the median percent of reads mapped for each cohort.*

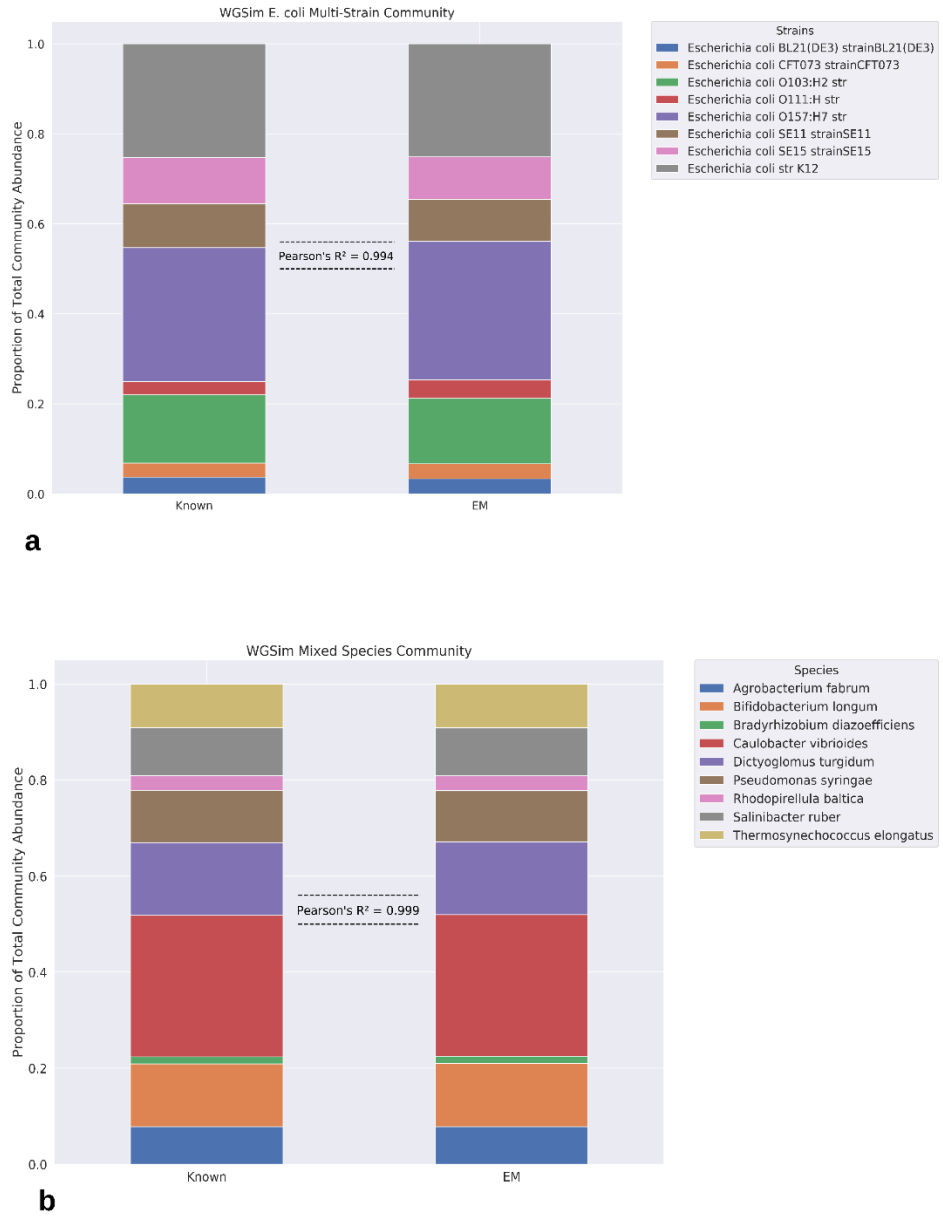


Figure 20: EM benchmarking on simulated bacterial communities.

Stacked bar graphs showing benchmarking results of our EM algorithm on estimating known genome relative abundances from simulated whole-genome shotgun sequences created with WGSim; ***a.*** *strain level results of a mixed E. coli community with Pearson's R_2 = 0.997 between known genome relative abundances and the EM genome relative abundance estimations;* ***b.*** *species level results of a mixed community with a Pearson's R_2 = 0.999 between the known genome relative abundances and the EM genome rela-tive abundance estimations.*

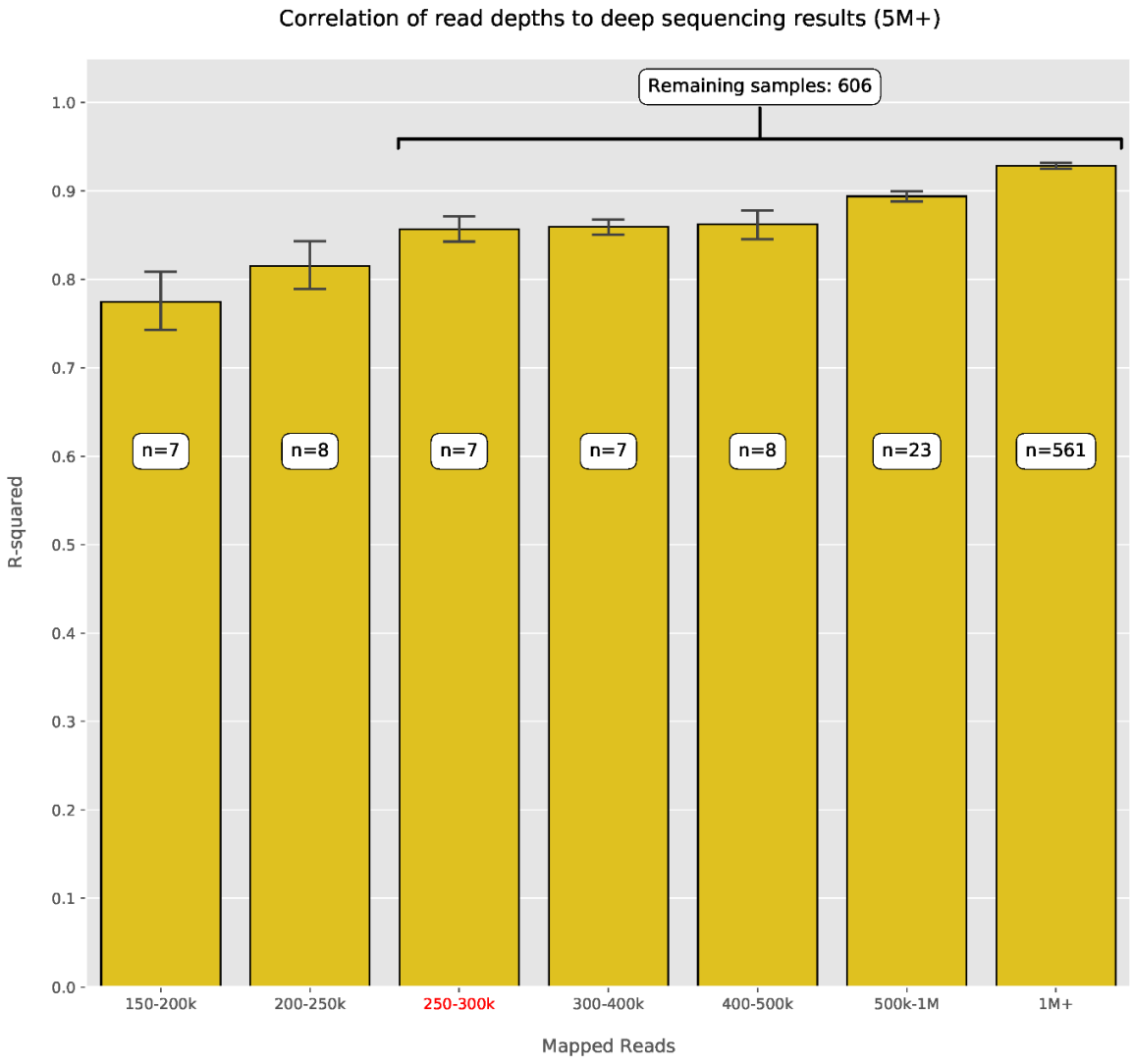


Figure 21:Read-depth benchmarking.

Correlation of varying read depths with samples at 5+ million read depth. Samples with 5+ million reads were sub-sampled to varying depths and examined using ordinary least squares linear regression. Samples with 250 000+ reads, on average, demonstrate an R^2 value greater than 0.85. The red text indicated the chosen threshold for subsequent analysis.

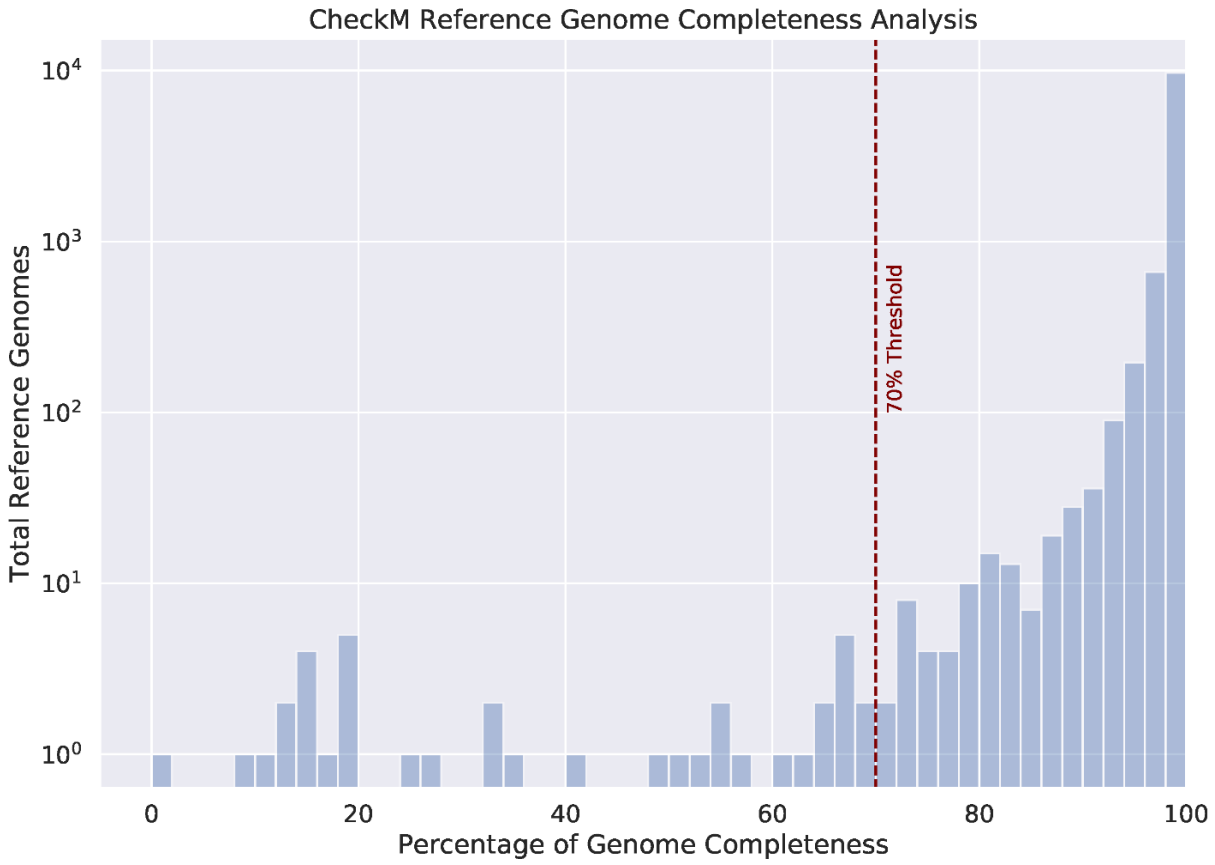


Figure 22: Reference genome completeness estimation.

All reference genomes utilized for read mapping were analyzed for their percentage of genome completeness with CheckM. In total there were 10 839 genomes of which only 38 (0.004%) that were designated as below 70% complete. One genome was marked as 0.0% complete although that was due to CheckM not having data on the lineage of that organism.

Tables

Table 1: Cohort network topological properties.

*Network topological properties calculated for each cohort's network. The plus (+) or minus (-) sign indicates that the network property was greater or lower than the average of 1,000 random networks. Stars indicate that the network property was statistically significantly different (P-value: * < 0.05, ** <0.01, ***<0.001).*

Network	Nodes	Edges	Density	ASPL	Transitivity	Modularity	Degree Assortativity	Genera Assortativity
American	202	338	0.017	1.539	0.487	0.475	0.338	0.144
Indian	202	273	0.013	1.874	0.452	0.667	0.330	0.163
European	202	386	0.019	1.369	0.353	0.681	0.158	0.196
Japanese	202	274	0.013	1.444	0.471	0.755	0.308	0.242

References

1. Sender, R., Fuchs, S. & Milo, R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLOS Biology* **14**, e1002533 (2016).

2. Kho, Z. Y. & Lal, S. K. The Human Gut Microbiome – A Potential Controller of Wellness and Disease. *Front. Microbiol.* **9**, 1835 (2018).

3. Kostic, A. D., Xavier, R. J. & Gevers, D. The Microbiome in Inflammatory Bowel Disease: Current Status and the Future Ahead. *Gastroenterology* **146**, 1489–1499 (2014).

4. Thaiss, C. A., Zmora, N., Levy, M. & Elinav, E. The microbiome and innate immunity. *Nature* **535**, 65–74 (2016).

5. Das, B. & Nair, G. B. Homeostasis and dysbiosis of the gut microbiome in health and disease. *Journal of Biosciences* **44**, 117 (2019).

6. Shreiner, A. B., Kao, J. Y. & Young, V. B. The gut microbiome in health and in disease: *Current Opinion in Gastroenterology* **31**, 69–75 (2015).

7. Petersen, C. & Round, J. L. Defining dysbiosis and its influence on host immunity and disease: How changes in microbiota structure influence health. *Cell Microbiol* **16**, 1024–1033 (2014).

8. Karlsson, F. H. *et al.* Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498**, 99–103 (2013).

9. Koren, O. *et al.* Human oral, gut, and plaque microbiota in patients with atherosclerosis. *Proceedings of the National Academy of Sciences* **108**, 4592–4598 (2011).

10. Karlsson, F. H. *et al.* Symptomatic atherosclerosis is associated with an altered gut metagenome. *Nat Commun* **3**, 1245 (2012).

11. MetaHIT consortium *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541–546 (2013).

12. Franzosa, E. A. *et al.* Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat Microbiol* **4**, 293–305 (2019).

13. Becker, C., Neurath, M. F. & Wirtz, S. The Intestinal Microbiota in Inflammatory Bowel Disease. *ILAR J* **56**, 192–204 (2015).

14. Kostic, A. D. *et al.* Fusobacterium nucleatum Potentiates Intestinal Tumorigenesis and Modulates the Tumor-Immune Microenvironment. *Cell Host & Microbe* **14**, 207–215 (2013).

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

16. Johnson, A. J. *et al.* Daily Sampling Reveals Personalized Diet-Microbiome Associations in Humans. *Cell Host & Microbe* **25**, 789-802.e5 (2019).

17. Villmones, H. C. *et al.* Species Level Description of the Human Ileal Bacterial Microbiota. *Sci Rep* **8**, 4736 (2018).

18. Gevers, D. *et al.* The Treatment-Naive Microbiome in New-Onset Crohn's Disease. *Cell Host & Microbe* **15**, 382–392 (2014).

19. David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559–563 (2014).

20. Zhou, J., Deng, Y., Luo, F., He, Z. & Yang, Y. Phylogenetic Molecular Ecological Network of Soil Microbial Communities in Response to Elevated CO2. *mBio* **2**, e00122-11 (2011).

21. Lupatini, M. *et al.* Network topology reveals high connectance levels and few key microbial genera within soils. *Front. Environ. Sci.* **2**, (2014).

22. Eiler, A., Heinrich, F. & Bertilsson, S. Coherent dynamics and association networks among lake bacterioplankton taxa. *ISME J* **6**, 330–342 (2012).

23. Kara, E. L., Hanson, P. C., Hu, Y. H., Winslow, L. & McMahon, K. D. A decade of seasonal dynamics and co-occurrences within freshwater bacterioplankton communities from eutrophic Lake Mendota, WI, USA. *ISME J* **7**, 680–684 (2013).

24. Shetty, S. A., Hugenholtz, F., Lahti, L., Smidt, H. & de Vos, W. M. Intestinal microbiome landscaping: insight in community assemblage and implications for microbial modulation strategies. *FEMS Microbiol. Rev.* **41**, 182–199 (2017).

25. Gould, A. L. *et al.* Microbiome interactions shape host fitness. *Proc Natl Acad Sci USA* **115**, E11951–E11960 (2018).

26. Hibbing, M. E., Fuqua, C., Parsek, M. R. & Peterson, S. B. Bacterial competition: surviving and thriving in the microbial jungle. *Nat Rev Microbiol* **8**, 15–25 (2010).

27. Fox, G. E., Magrum, L. J., Balch, W. E., Wolfe, R. S. & Woese, C. R. Classification of methanogenic bacteria by 16S ribosomal RNA characterization. *Proceedings of the National Academy of Sciences* **74**, 4537–4541 (1977).

28. Venter, J. C. *et al.* Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science* **304**, 66 (2004).

29. Větrovský, T. & Baldrian, P. The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. *PLoS ONE* **8**, e57923 (2013).

30. Edgar, R. C. Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. *PeerJ* **6**, e4652 (2018).

31. Ranjan, R., Rani, A., Metwally, A., McGee, H. S. & Perkins, D. L. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochem Biophys Res Commun* **469**, 967–977 (2016).

32. Laudadio, I. *et al.* Quantitative Assessment of Shotgun Metagenomics and 16S rDNA Amplicon Sequencing in the Study of Human Gut Microbiome. *OMICS: A Journal of Integrative Biology* **22**, 248–254 (2018).

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

34. Aitchison, J. The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society. Series B (Methodological)* **44**, 139–177 (1982).

35. Layeghifard, M., Hwang, D. M. & Guttman, D. S. Disentangling Interactions in the Microbiome: A Network Perspective. *Trends in Microbiology* **25**, 217–228 (2017).

36. Kurtz, Z. D. *et al.* Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLOS Computational Biology* **11**, e1004226 (2015).

37. Friedman, J., Hastie, T. & Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441 (2008).

38. Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science* **352**, 560 (2016).

39. Efron, B. & Tibshirani, R. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statist. Sci.* **1**, 54–75 (1986).

40. Su, W., Bogdan, M. & Candes, E. False Discoveries Occur Early on the Lasso Path. *arXiv:1511.01957 [cs, math, stat]* (2016).

41. Saunders, A. M., Albertsen, M., Vollertsen, J. & Nielsen, P. H. The activated sludge ecosystem contains a core community of abundant organisms. *ISME J* **10**, 11–20 (2016).

42. Cordasco, G. & Gargano, L. Community Detection via Semi-Synchronous Label Propagation Algorithms. *arXiv:1103.4550 [physics]* (2011) doi:10.1504/..045103.

43. Peura, S., Bertilsson, S., Jones, R. I. & Eiler, A. Resistant Microbial Cooccurrence Patterns Inferred by Network Topology. *Appl. Environ. Microbiol.* **81**, 2090–2097 (2015).

44. Prettejohn, B. J., Berryman, M. J. & McDonnell, M. D. Methods for Generating Complex Networks with Selected Structural Properties for Simulations: A Review and Tutorial for Neuroscientists. *Front. Comput. Neurosci.* **5**, (2011).

45. Guimerà, R., Sales-Pardo, M. & Amaral, L. A. N. Modularity from fluctuations in random graphs and complex networks. *Phys. Rev. E* **70**, 025101 (2004).

46. Trosvik, P. & de Muinck, E. J. Ecology of bacteria in the human gastrointestinal tract—identification of keystone and foundation taxa. *Microbiome* **3**, 44 (2015).

47. Verster, A. J. & Borenstein, E. Competitive lottery-based assembly of selected clades in the human gut microbiome. *Microbiome* **6**, 186 (2018).

48. Berry, D. & Widder, S. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Frontiers in Microbiology* **5**, (2014).

49. Foster, K. R. & Bell, T. Competition, Not Cooperation, Dominates Interactions among Culturable Microbial Species. *Current Biology* **22**, 1845–1850 (2012).

50. Nemergut, D. R. *et al.* Patterns and Processes of Microbial Community Assembly. *Microbiology and Molecular Biology Reviews* **77**, 342–356 (2013).

51. Hamilton, W. D. The genetical evolution of social behaviour. I. *J Theor Biol* **7**, 1–16 (1964).

52. Hardin, G. The Competitive Exclusion Principle. *Science* **131**, 1292–1297 (1960).

53. Jackson, M. A. *et al.* Detection of stable community structures within gut microbiota co-occurrence networks from different human populations. *PeerJ* **6**, e4303 (2018).

54. Darcy, J. L. *et al. A phylogenetic model for the recruitment of species into microbial communities and application to studies of the human microbiome*. http://biorxiv.org/lookup/doi/10.1101/685644 (2019) doi:10.1101/685644.

55. Pacheco, A. R., Moel, M. & Segrè, D. Costless metabolic secretions as drivers of interspecies interactions in microbial ecosystems. *Nat Commun* **10**, 103 (2019).

56. Chase, J. M. & Leibold, M. A. Spatial scale dictates the productivity–biodiversity relationship. *Nature* **416**, 427–430 (2002).

57. Zarrinpar, A., Chaix, A., Yooseph, S. & Panda, S. Diet and Feeding Pattern Affect the Diurnal Dynamics of the Gut Microbiome. *Cell Metabolism* **20**, 1006–1017 (2014).

58. Mark Welch, J. L., Hasegawa, Y., McNulty, N. P., Gordon, J. I. & Borisy, G. G. Spatial organization of a model 15-member human gut microbiota established in gnotobiotic mice. *Proc Natl Acad Sci USA* **114**, E9105–E9114 (2017).

59. Fung, T. C., Artis, D. & Sonnenberg, G. F. Anatomical localization of commensal bacteria in immune cell homeostasis and disease. *Immunol Rev* **260**, 35–49 (2014).

60. Donaldson, G. P., Lee, S. M. & Mazmanian, S. K. Gut biogeography of the bacterial microbiota. *Nat Rev Microbiol* **14**, 20–32 (2016).

61. Stachowicz, J. J. Mutualism, Facilitation, and the Structure of Ecological Communities. *BioScience* **51**, 235 (2001).

62. Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. & Knight, R. Diversity, stability and resilience of the human gut microbiota. *Nature* **489**, 220–230 (2012).

63. Lynd, L. R., Weimer, P. J., van Zyl, W. H. & Pretorius, I. S. Microbial Cellulose Utilization: Fundamentals and Biotechnology. *MICROBIOL. MOL. BIOL. REV.* **66**, 72 (2002).

64. Turroni, F. *et al.* Glycan cross-feeding activities between bifidobacteria under in vitro conditions. *Front. Microbiol.* **6**, (2015).

65. Hall, C. V. *et al.* Co-existence of Network Architectures Supporting the Human Gut Microbiome. *iScience* **22**, 380–391 (2019).

66. Fisher, C. K. & Mehta, P. Identifying Keystone Species in the Human Gut Microbiome from Metagenomic Timeseries Using Sparse Linear Regression. *PLOS ONE* **9**, e102451 (2014).

67. Jones, M. B. *et al.* Library preparation methodology can influence genomic and functional predictions in human microbiome research. *Proceedings of the National Academy of Sciences* **112**, 14024–14029 (2015).

68. Lahti, L., Salojärvi, J., Salonen, A., Scheffer, M. & de Vos, W. M. Tipping elements in the human intestinal ecosystem. *Nat Commun* **5**, 4344 (2014).

69. Dhakan, D. B. *et al.* The unique composition of Indian gut microbiome, gene catalogue, and associated fecal metabolome deciphered using multi-omics approaches. *GigaScience* **8**, (2019).

70. MetaHIT Consortium *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).

71. Yachida, S. *et al.* Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat Med* **25**, 968–976 (2019).

72. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

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

74. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733–D745 (2016).

75. Xia, L. C., Cram, J. A., Chen, T., Fuhrman, J. A. & Sun, F. Accurate Genome Relative Abundance Estimation Based on Shotgun Metagenomic Reads. *PLoS ONE* **6**, e27992 (2011).

76. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* **25**, 1043–1055 (2015).

77. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).

78. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

79. Haft, D. H. TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Research* **29**, 41–43 (2001).

80. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

81. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2020).

82. Zhao, T., Liu, H., Roeder, K. & Wasserman, L. The huge Package for High-dimensional Undirected Graph Estimation in R. 6 (2016).

83. Liu, H., Roeder, K. & Wasserman, L. Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models. *arXiv:1006.3316 [stat]* (2010).

84. Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring Network Structure, Dynamics, and Function using NetworkX. 6 (2008).

85. Newman, M. E. J. Networks: An Introduction. in 168–234 (Oxford University Press, Inc., 2010).

86. Newman, M. E. J. Modularity and community structure in networks. *PNAS* **103**, 8577–8582 (2006).

87. Newman, M. E. J. Mixing patterns in networks. *Phys. Rev. E* **67**, 026126 (2003).

88. P. Erdos and A. Rényi, “On Random Graphs I,” Publicationes Mathematicae, Vol. 6, 1959, pp. 290-297.

89. Fortunato, S. Community detection in graphs. *Physics Reports* **486**, 75–174 (2010).

90. Brandes, U. A faster algorithm for betweenness centrality*. *The Journal of Mathematical Sociology* **25**, 163–177 (2001).

91. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410 (1990).

Author contributions

M.L. and S.H. initiate the study and S.Y. assisted in study design. M.L. and S.H. conducted the taxonomic and functional analyses. S.Y. created the network inference program and conducted the network inference. M.L. and S.H. conducted the bacterial network analysis. M.L. and S.H. wrote the manuscript and created the figures with input and assistance from S.Y.

CHAPTER 3: LINKING INFLAMMATORY BOWEL DISEASE SYMPTOMS TO CHANGES IN THE GUT MICROBIOME STRUCURE AND FUNCTIONS

Note: This section has been published in part and the citation link is: Hassouneh, S. A. D., Loftus, M., & Yooseph, S. (2021). Linking Inflammatory Bowel Disease Symptoms to Changes in the Gut Microbiome Structure and Function. Frontiers in Microbiology, 12, 2009. https://www.frontiersin.org/articles/10.3389/fmicb.2021.673632.

Introduction

Inflammatory bowel disease (IBD) is a heterogeneous disorder characterized by chronic inflammation of the gastrointestinal tract. The two main manifestations of IBD are Crohn's Disease (CD) and Ulcerative Colitis (UC). CD most often affects the terminal ileum but can affect any part of the gastrointestinal tract in a non-contiguous fashion, sometimes known as 'skip lesions', and often results in diarrhea, bloody stools, abdominal pain, cachexia, and fatigue (Veauthier and Hornecker 2018; Flores et al. 2015). UC most often affects the large intestine, extending from the rectum, and is characterized by contiguous inflammation and often results in rectal bleeding, bloody stools, diarrhea, cachexia, and fatigue (Flores et al. 2015; "FDA Briefing Document Gastrointestinal Drug Advisory Committee Meeting" 2018). While the etiology of IBD is not well understood, it is believed that the disorder arises due to environmental and host-related factors causing an aberrant immune response in genetically predisposed

individuals (Chiara et al. 2020; Kish et al. 2013). One such factor is believed to be the microbiome, specifically the gut microbiome (Duranti et al. 2016).

The human microbiome is the community of microbes that exists on and within the human body and has been implicated in maintaining health, as well as possibly contributing to a multitude of diseases such as IBD, Irritable Bowel Syndrome (IBS), diabetes, Parkinson's disease, and amyotrophic lateral sclerosis (Gevers, Kugathasan, Denson, Vázquez-Baeza, Van Treuren, et al. 2014; Vich Vila et al. 2018; Brown et al. 2011; Petrov et al. 2017; Wu et al. 2015; Kho and Lal 2018). The bacterial composition of the microbiome can be studied using DNA sequencing, either by targeted sequencing of a marker gene or by shotgun sequencing. Targeted sequencing involves the amplification of specific regions of bacterial genomes, such as the 16S ribosomal RNA gene, for use as a phylogenetic marker (George E Fox et al. 1977). However, due to the highly conserved nature of the 16S rRNA gene and the short lengths of the regions within the gene that are commonly targeted, the taxonomic resolution generated by these types of studies are often inadequate to distinguish bacterial species (G. E. Fox, Wisotzkey, and Jurtshuk 1992; Ranjan et al. 2016). Furthermore, estimation of bacterial relative abundances is confounded by the presence of multiple copies and intragenic variation of the 16S rRNA gene within a single bacterium (Rastogi et al. 2009; Ibal et al. 2019). In contrast, shotgun sequence data generated from the DNA extracted from a sample can be used to obtain more accurate estimates of relative abundance, higher

resolution of bacterial taxonomy, and a more accurate representation of genomic functional capacity (Ranjan et al. 2016; Laudadio et al. 2018).

Regardless of the sequencing framework used, the generated sequence data are compositional in nature enabling only an estimation of the *relative* abundances of the constituent microbial taxa (Gloor et al. 2017). This compositionality aspect makes it difficult to analyze differential abundance, infer associations, and estimate correlations (Aitchison 1982; Jonathan Friedman and Alm 2012; Tsilimigras and Fodor 2016; Pearson 1896). By utilizing a Centered Log-Ratio (CLR) transformation of the relative abundance data, we can examine the differential abundances more clearly and without inducing spurious correlations (Aitchison 1982; Jonathan Friedman and Alm 2012; Tsilimigras and Fodor 2016; Pearson 1896). Furthermore, the covariance matrix of log-transformed relative abundance data provides a good approximation of the covariance matrix of the log-transformed absolute abundance data enabling us to better model the associations between bacteria (Kurtz et al. 2015).

Associations within a bacterial community are comprised of the direct and indirect interactions between the community constituents and are important for understanding the underlying dynamics at play in a microbial community (Kurtz et al. 2015). Bacterial association networks are often constructed using pairwise correlation methods on relative abundance or count data of the bacteria found within the samples.

Due to the compositional nature of sequencing data, however, it is difficult to accurately identify correlations from counts generated from sequencing data as a result of spurious correlations that arise (Jonathan Friedman and Alm 2012). Even after CLR-transformation of the sequencing data, pairwise correlation methods are unable to account for conditional independence between bacterial species causing these methods to produce inaccurate bacterial association networks (Kurtz et al. 2015). In this paper, we used a Gaussian Graphical Model (GGM) framework in conjunction with a graphical lasso (glasso) to construct bacterial association networks from the CLR-transformed relative abundance data (Jerome Friedman, Hastie, and Tibshirani 2008; Loftus, Hassouneh, and Yooseph 2021). We represent these bacterial association networks using an unweighted graph in which nodes denote bacterial species and an edge between two nodes denotes an association between the corresponding bacterial species. Utilizing the GGM framework on the CLR-transformed data, enables us to approximate the covariance structure of the absolute abundances as well as account for conditional independence between the constituent species (Wermuth and Lauritzen 1990; Aitchison 1982). By utilizing shotgun sequence data and employing compositionally robust methodologies, we can identify potentially important differences in bacterial associations, taxonomic composition, and functional capacity between the IBD and healthy gut microbiomes that may play a role in disease and symptom progression.

Due to the Random Forest Classifier's (RFC) ability to deal with 'noisy', non-normally distributed, multi-dimensional data, it has become an important tool for identifying important features and differences in the microbiome (Breiman 2001; Díaz-Uriarte and Alvarez de Andrés 2006; Loomba et al. 2017; Saulnier et al. 2011; Roguet et al. 2018; Shi et al. 2005). These features can include bacterial relative abundances and metadata thus allowing us to generate a model that accounts for subject characteristics as well as gut microbiome taxonomic profiles. Another benefit of the RFC is its ability to assign importance to the features used for the classification. The feature importance's allow us to quantify the role a specific feature plays in making a prediction and can allow us to determine which features may be informative. One shortcoming of these feature importances, however, is their lack of statistical significance. Due to the stochastic nature of model construction using an RFC, some features may be relatively important in one instance of an RFC model, but relatively unimportant in another instance of the RFC model. To enable us to utilize RFC feature importance to distinguish potentially important features and reduce the dimensionality of our data, we formulated a framework that allowed us to add statistical significance to the feature importances.

Here, we utilized the IBD Multi-omics DataBase (IBDMDB) cohort from a previously published study to study IBD (Lloyd-Price et al. 2019). This dataset consists of shotgun sequence data generated from CD, UC, and an internal control group (henceforth also referred to as non-IBD samples). The non-IBD samples were collected

from subjects that underwent histopathologic examination (via colonoscopy) but were not diagnosed with IBD. These samples are derived from subjects presenting for routine screenings, gastrointestinal (GI) distress, or non-specific symptoms generating a heterogeneous control group. This control group design may obfuscate important differences between healthy and IBD gut microbiomes, especially if the differences may be related to presentations common between IBD and GI distress, such as diarrhea, bloating, or abdominal pain. Additionally, many studies examining the microbiome suffer from a lack of cross-cohort consistency making it difficult to generalize findings to populations rather than just the utilized study groups (Pasolli et al. 2016). One proposed remedy for this lack of cross-cohort consistency is to utilize external samples from independent cohorts, especially when comparing diseased and healthy microbiomes, and applying the same methods and techniques across all samples (Pasolli et al. 2016; Thomas et al. 2019). To enable us to generalize our findings and utilize healthy control groups in our analysis, we incorporated samples from both the Human Microbiome Project (Huttenhower et al. 2012) referred to as the Healthy-1 cohort, and from A.J. Johnson et al 2019 (Johnson et al. 2019) referred to as the Healthy-2 cohort, as external controls. The external cohorts we elected to use were shotgun sequence datasets generated from gut microbiome samples collected from healthy subjects (no overt or reported disease) and utilizing the same sequencing platform as the IBDMDB cohort (Illumina). Furthermore, due to the similarity of the results produced by the Chemagic DNA extraction kit (IBDMDB cohort) and the Mo Bio PowerSoil DNA extraction kit (Healthy-1 and Healthy-2 cohorts), we concluded that these cohorts could serve as external controls without the addition of a significant amount of technical bias

(Multinu et al. 2018). Also, due to the use of replicates within the Healthy-2 cohort and the IBDMDB cohort, we were able to examine temporal variation within subjects diagnosed with IBD relative to the non-IBD group (internal control) and the Healthy-2 group (external control). By incorporating these two independent healthy cohorts, we can compare the IBD samples to healthy samples and mitigate the possible issues inherent in the design of the IBDMDB internal control group (non-IBD group) as well as arrive at more robust and generalizable conclusions from our analysis.

To understand the effects of changes in the microbiome, we cannot solely focus on the presence, absence, or differential abundances that are found. We also need to examine how these bacteria interact with each other to understand the impacts they have on shaping the microbiome. It is also integral that we examine the functional consequences of these differences and associations to build a more complete picture of the changes the gut microbiome underwent and the possible effects these changes may foment (Heintz-Buschart and Wilmes 2018). By examining the taxonomy, the bacterial interactions, and the functional changes of the gut microbiome, our study aims to identify bacterial species that may play a role in the onset or exacerbation of IBD or IBD-related symptoms. By utilizing two external healthy controls, we are also able to corroborate our conclusions when comparing IBD and healthy samples and generalize our findings more confidently to the population. Additionally, we utilized a machine learning framework and a prevalence threshold to identify potentially important bacterial species. We also compared the functional capacity of the gut microbiome of IBD

samples to non-IBD and control samples, and identified important potential functional differences that may play a role in symptoms IBD patients typically experience.

Materials and Methods

Data Acquisition

Shotgun sequence data generated from 574 fecal samples were obtained from three previously published studies of the human gut microbiome (United States populations). Of these, two cohorts were downloaded from NCBI's Sequence Read Archive (SRA): Human Microbiome Project (SRA: PRJNA48479; 203 samples) and the IBD Multi-omics Database (SRA: PRJNA398089; 257 samples). The A.J. Johnson et al cohort was downloaded from the European Nucleotide Archive (ENA) (ENA: PRJEB29065; 114 samples). We were able to access metadata for sex and age/age-group for all cohorts.

Data pre-processing

Reads from the whole genome sequencing data were trimmed using Trimmomatic (version 0.36) and then reads corresponding to the human genome were filtered out using BowTie2 (version 5.4.0) and the GRCh38.p12 (www.ncbi.nlm.nih.gov/assembly/GCF_000001405.38) human reference genome (Bolger, Lohse, and Usadel 2014; Langmead and Salzberg 2012).

Read mapping and taxonomic identification

Reads were mapped to 10,839 bacterial reference strain genomes obtained from the NCBI RefSeq database using BowTie2 (O'Leary et al. 2016). Bacterial genome relative abundances were estimated using a probabilistic framework based on a mixture model. The framework utilized an Expectation-Maximization (EM) algorithm to perform soft assignment of the reads to the reference genomes and was found to be highly accurate (Xia et al. 2011; Loftus, Hassouneh, and Yooseph 2021). We have previously demonstrated that samples with less than 250,000 mapped reads display diminished accuracy for taxonomic profiling, consequently all samples that contained less than 250,000 mapped reads threshold were not used for downstream analysis (Loftus, Hassouneh, and Yooseph 2021). When calculating abundances, any strain that had a relative abundance below 1×10^{-5} was considered statistical noise. The relative abundance data was then transformed using the CLR transformation and the CLR-transformed data was used for all downstream analyses except for the alpha-diversity analysis. The CLR transformation is defined as:

$$CLR(x) = \left[ln\frac{x_1}{G(x)}, ln\frac{x_2}{G(x)} \ldots ln\frac{x_D}{G(x)}\right]$$

where *x* is the vector of relative abundances within a sample, *D* is the total number of species present within the sample, and *G(x)* is the geometric mean of *x*. The geometric mean is defined as:

$$G(x) = \sqrt[D]{x_1 \times x_2 \times \ldots x_D}.$$

Sample inclusion criteria

IBDMDB inclusion criteria

To reduce potential confounders within the internal control group (non-IBD samples), we instituted a set of inclusion criteria for the non-IBD group: no colonoscopy within the last two weeks, no history of bowel surgery, no immunosuppressants use, no antibiotic use, no IBS, and no diarrhea in the past two weeks. Due to the adverse associations between these variables and the gut microbiome that have been noted in the literature, we excluded any samples from subjects that violated these criteria (Schubert et al. 2014; Halfvarson et al. 2017; Bhat et al. 2017; Dethlefsen et al. 2008; Vich Vila et al. 2018; Nagata et al. 2019). We also did not utilize any samples collected prior to week 26 of the study to ensure that subjects had ample time to overcome any gastrointestinal distress they have been experiencing at the time of study initiation. To limit any potential bias from an over-representation of a subject within the cohort, no more than five randomly chosen samples were retained from any one subject for any of the sample groups in the IBDMDB cohort (CD, UC, non-IBD) resulting in a mean number of replicates of 2.5 and a median of 2.

Healthy-1 cohort inclusion criteria

Samples for the healthy-1 cohort were derived from (Huttenhower et al. 2012) and were generated as part of the Human Microbiome Project. All 203 samples utilized were derived from unique individuals and demonstrated over 250,000 mapped reads so all samples were included in the analysis.

Healthy-2 cohort inclusion criteria

Samples for the healthy-2 cohort were derived from (Johnson et al. 2019) and were generated as part of a longitudinal analysis of fecal shotgun metagenomes in healthy subjects. The study by Johnson et al aimed to examine gut microbiome responses to a changes diet. Subject were randomly given fatty acid supplementation on days 10-17 of the study. To ensure that our analysis reflected healthy samples on habitual diets, only samples taken prior to day 10 of the study were used. Furthermore, subjects were sampled daily for 17 days but not all subjects consistently had more than five samples with greater than 250,000 (minimum threshold for inclusion) mapped reads so to limit the number of replicates from a single subject a maximum of five randomly chosen samples were retained from any one subject resulting in a mean number of replicates of 3.3 and a median of 3.

Diversity analysis

Alpha diversity was analyzed using the Shannon entropy. The Shannon entropy, H, is defined as:

$$H = -\sum_{i=1}^{D} p_i \, log_2(p_i)$$

where *D* is the number of species in the sample and p_i is the proportion of species *i* in the sample (Shannon 1948). The non-transformed relative abundances were used for the Shannon entropy calculations.

Intrapersonal and interpersonal dissimilarity

The Bray-Curtis dissimilarity (BCD) between replicates within a subject was used to quantify intrapersonal variation within each cohort with replicates (IBDMDB and Healthy-2 cohorts). The BCD between subjects within diagnosis groups (interpersonal dissimilarity) was also examined to observe the variability of the gut microbiome within the diagnosis groups. The BCD between two samples, *i* and *j,* was calculated as

$$BCD_{i,j} = 1 - \frac{2C_{ij}}{S_i + S_j}$$

where C_{ij} is the sum of the relative abundances of the species with the lowest combined relative abundance within samples *i and j.* S_i and S_j are the sums of the relative abundances found in sample *i* and sample *j*, respectively. The intrapersonal dissimilarity was calculated by generating pairwise BCD's for samples from the same subject. The interpersonal dissimilarity was calculated by generating pairwise BCD's between samples from different subjects.

Prevalent species

To reduce the dimensionality of our data, we utilized only bacterial species that were present in at least 90% of samples within each diagnosis group (IBD, non-IBD, Healthy-1, and Healthy-2) for our downstream analysis (Loftus, Hassouneh, and Yooseph 2021). The union of the bacterial species present at a prevalence greater than or equal to 90% in each diagnosis group was then used for the classification of the signature species.

Classification of signature species

A modified Random Forest Classifier (RFC) framework was used to identify bacterial species for downstream analysis (Breiman 2001). The RFC was used to classify samples by the sample groups (IBD, non-IBD, and Healthy). The Healthy-1 and Healthy-2 cohort were combined for the RFC analysis to enable us to identify bacterial species importances by health status, rather than by cohort. A random noise column was added into the data prior to RFC analysis. The noise column was generated by creating a normal distribution resembling the CLR-transformed data of the genome relative abundances and randomly sampling from the distribution. The data was then label encoded due to the presence of categorical data. This process was performed 100 times, where a new random noise column would be generated each time, and the feature importances of every feature (bacterial species, metadata, and the random feature) were stored for all runs. A Mann-Whitney U test was then performed on the importances of all features with a mean feature importance higher than the random

feature to determine if the importances of these features were significantly different from the feature importances of the random column. The Benjamini-Hochberg procedure for controlling false discovery rate was utilized to account for the multiple-testing and only features with a q-values less than 0.05 were considered significantly different from the random column (Benjamini and Hochberg 1995). This framework allows us to identify the bacterial species and metadata whose feature importances were significantly higher than the random noise. The bacterial species that were significantly more important than the random noise column are referred to as the 'signature' species due to their ability to provide a non-random signal during classification. The RFC was implemented in Python 3.8 using Sci-kit Learn 0.23.1 (Van Rossum, Guido and Drake 2009; Varoquaux et al. 2015).

Differential abundance analysis

Differential abundance analysis was conducted by performing a Mann-Whitney U test and the Benjamini-Hochberg multi-test correction on the CLR-transformed relative abundance profiles. The IBD group was compared to the non-IBD group, the Healthy-1 group, and the Healthy-2 group individually. Bacterial species that were significantly differentially abundant in IBD relative to every other individual group were designated as differentially abundant.

Bacterial association network construction

The signature species were used to create a sample-taxa matrix of CLR-transformed relative abundances in each sample. The GGM framework, as previously described, was used to generate the bacterial association networks using the above sample-taxa matrices for each cohort (Loftus, Hassouneh, and Yooseph 2021). In brief, the HUGE package in R was used to compute a sparse precision matrix. The stability approach to regularization selection (StARS) method was used to determine the tuning parameter in the l_1-penalty model for sparse precision matrix estimation. To reduce false positives, the final precision matrix, Ω, underwent bootstrap testing. If Ω[i,j] ≠ 0, then Ω'[i,j] = Ω[i,j] if$[i,j] \neq 0$ in f*r or greater precision matrices estimated from bootstraping. Otherwise, Ω'[i,j]=0. The value *r* = 50 (bootstrap replicates) and *f* = 0.8 (threshold between 0 and 1 indicating proportion of edges that must be non-zero). Networks were visualized and analyzed using Python 3.8 and NetworkX 2.4 (Hagberg, Schult, and Swart 2008).

Eigenvector centrality

Eigenvector centrality (EVC) measures the influence a node has in a network by accounting for the connections of the node in question as well as the connections of its neighbors (Bonacich 1972; Ruhnau 2000). The EVC, *x*, for a given node, *i*, is defined as:

$$x_i = \sum_j A_{ij}\, x_j$$

where *A* is the adjacency matrix and *j* is a neighboring node of *i*.

Bacterial genome functional annotation

Prodigal (version 2.6.3) was used to identify genes and generate protein sequence translations (Hyatt et al. 2010). The protein sequence translations were provided to InterProScan (version 5.39-77.0) to identify protein families using the TIGRFAM (versions 15.0) protein family database (Hunter et al. 2009; Haft 2001). TIGRFAM counts were generated for each reference genome. Bacterial species that were greater than 90% prevalent within a diagnosis group (IBD, non-IBD, Healthy-1, and Healthy-2) were used for functional annotation to reduce the effects of potentially transient species when analyzing the genomic functional capacity of the microbiomes (Saunders et al. 2016; Ursell et al. 2012). Then the TIGRFAM counts were weighted based on CLR-transformed genome relative abundance, and summed by total for each cohort. Differential abundances of TIGRFAM profiles were therefore calculated by using the CLR-transformed relative abundances of the TIGRFAMs within each cohort. The TIGRFAM CLR-transformed relative abundances were then tested using a Mann-Whitney U test.

Statistical analysis and graph creation

Statistical analysis and graph creation was performed using Python 3.8 (Van Rossum, Guido and Drake 2009).

Results

A total of 569 shotgun sequence datasets from 3 previously published studies (IBDMDB, Healthy-1, and Healthy-2) of the human gut microbiome were utilized in this study. The IBDMDB cohort consisted of CD, UC, and non-IBD samples. To minimize potential confounders in the IBDMDB group, samples from individuals that reported recent colonoscopy, antibiotic or immunosuppressant use, IBS, or recent GI symptoms were excluded from the control (non-IBD) group. For each dataset, the sequence reads were quality trimmed and human reads were identified and filtered. The remaining reads were mapped to a comprehensive collection of 10,839 bacterial strain reference genomes from NCBI RefSeq and genome relative abundances were calculated using a probabilistic framework (Xia et al. 2011; Loftus, Hassouneh, and Yooseph 2021). The alpha-diversity was then calculated on the relative abundances using Shannon entropy. To reduce the dimensionality of our data, we focused our analysis on bacterial species that were prevalent in at least 90% of the samples. Next, the relative abundance vector for each sample was CLR transformed and used for all downstream analysis. A random forest classifier (RFC) framework (Breiman 2001) was then used to classify the samples by their diagnosis groups using the taxonomic profiles as well as the metadata available for all cohorts (sex, age, unique subject ID to account for replicates). For the RFC analysis, the Healthy-1 and Healthy-2 cohorts were grouped under one label (Healthy) to create a single healthy control group to compare to the IBD and non-IBD sample groups thus allowing us to identify important features that distinguish between diagnosis groups rather than cohort in a more robust manner (Pasolli et al. 2016; Thomas et al. 2019). The RFC was then trained on the taxonomic profiles as well as the metadata

available for all cohorts. While RFC's provide feature importances based on the features' contribution to classification of the given label, there is no statistical significance attached to these importances. To assess statistical significance of the features a random noise column was generated and added to the data (see methods). The species that were ranked as significantly more important than the random noise column were designated as the 'signature species' and used for all downstream analyses. A Mann-Whitney U test and Benjamini-Hochberg (BH) multi-test correction was used to compare the differential abundance of the signature species within IBD to all other groups individually.

Bacterial species that were significantly differentially abundant in IBD, relative to every other sample group, were designated as differentially abundant. Next, a GGM framework (see methods) was used to construct the bacterial association networks from the relative abundance information of each sample group. Finally, the genomic functional capacity within each sample group was determined by using the TIGRFAM protein family database. The TIGRFAM counts for each signature species were weighted by the relative abundance of the species within each sample group and then CLR-transformed. A Mann-Whitney U test and BH multi-test correction was then used to compare the differential abundance of the TIGRFAM functions within IBD to the other groups to determine differences in functional capacity.

Alpha-diversity analysis

The non-IBD group displayed a similar alpha-diversity to the UC and CD groups, however the external healthy cohorts displayed significantly higher alpha-diversities than all other groups (**Figure 23**). When examining the effect of cohort read-depth on alpha-diversity, we did not observe any significant correlation between read-depth and alpha-diversity (**Figure 24**). Notably, the Healthy-2 cohort displayed lower read-depth on average, relative to the IBDMDB cohort, but displayed significantly higher alpha-diversity.

Intrapersonal dissimilarity

When examining intrapersonal dissimilarity, it was noted that samples from the same subject were significantly more similar to each other than they were to samples from other subjects (**Figure 25a**). This trend was constant for every diagnosis group that could be tested (Healthy-1 cohort did not utilize replicates) and was statistically significant every time. Furthermore, it was observed that IBD samples demonstrated the highest levels of intrapersonal dissimilarity and were significantly higher than both non-IBD samples and Healthy-2 samples. Interestingly, the intrapersonal dissimilarity of non-IBD samples fell between the IBD and the Healthy-2 samples.

Interpersonal dissimilarity

To quantify how different the gut microbiomes of samples within a specific diagnosis group are, we examined the interpersonal dissimilarity. Once again, the IBD

samples exhibited the highest levels of dissimilarity when examining the interpersonal dissimilarity (**Figure 25b**). IBD sample interpersonal dissimilarities were significantly higher than the Healthy-1 and Healthy-2 samples but were not significantly different than the non-IBD samples. It was also noted that the non-IBD samples displayed significantly higher interpersonal dissimilarity, relative to the Healthy-1 and Healthy-2 cohorts.

Taxonomic analysis

When attempting to classify all different diagnoses (CD, UC, non-IBD, and healthy) using the RFC, it was noted that CD and UC samples were often misclassified as one another (CD as UC or *vice versa*) which contributed to the modest RFC classification accuracy (weighted average F1-score: 0.79) (**Figure 26a**). After combining the CD and UC diagnoses into the IBD sample group, the RFC was able to distinguish between the various cohorts with higher average accuracy (weighted average F1-score: 0.87) (**Figure 27**). Notably, the non-IBD group was difficult to distinguish, and these misclassifications were split between IBD and healthy controls implying that the non-IBD group had a heterogeneous composition in which some samples resembled healthy samples and others resembled IBD samples (**Figure 26b**). The RFC model identified 122 important features with the 'age' feature demonstrating the greatest feature importance. The 'unique subject ID' feature was also an important feature but was ranked 99/122 according to feature importance. The remaining 120 important features were bacterial species. The CLR-transformed relative abundances of these 120 species

were then compared between IBD and non-IBD (internal control) resulting in 55 significantly differentially abundance species. Out of these 55 species, 42 were significantly differentially abundant in IBD relative to all three control groups (non-IBD, Healthy-1 Healthy-2) with a q-value < 0.05 and greater than a two-fold difference (**Figure 28**). Of those 42 species, 34 were elevated in IBD and 8 species were elevated in the internal and external controls. All 42 of the above species were also found to be differentially abundant when utilizing the union of the 90% prevalent species for the differential abundance analysis. Out of the 34 species elevated in IBD, only the *Clostridium* (5 species) and *Blautia* (4 species) genera displayed more than 2 species elevated (**Figure 29**).

Bacterial association networks

Bacterial species elevated in IBD had non-zero degree in all bacterial association networks (**Figure 30**). While these nodes were elevated in IBD, they still maintained a higher than average number of associations within all networks (**Figure 31**). It was observed that while the nodes elevated in IBD display higher than average degree, the majority of nodes within each network were actually composed of species that were not significantly different between IBD and the control groups (IBD: 52.5%, non-IBD: 52.6%, Healthy-1: 65.6%, Healthy-2: 53.7%) (**Figure 32**). When examining the most important species within the network, defined as the species with the ten highest Eigenvector centralities, a measure of relative importance or influence of nodes, within a network, all but two of the ten species were found in the top-10 important species non-IBD or

healthy networks (**Table 2**) (Newman 2006). While there was a large amount of overlap, there were also 56 associations that are unique to the IBD network (**Figure 33**). The vast majority of these associations (85.71%) involved species that were elevated in IBD.

Differences in functional capacity

Analysis of the genomic functional capacities of the different cohorts demonstrated 6 significant differences with greater than two-fold fold change between the IBD cohort and all other cohorts (**Figure 34**). IBD samples displayed elevated relative abundance of protein families involved in sporulation and germination, synthesis and degradation of polysaccharides, signal transduction, regulatory protein interactions, and molybdopterin biosynthesis. The IBD samples also displayed reduced relative abundance of protein families involved in menaquinone and ubiquinone synthesis. Out of the 34 bacterial species elevated in IBD, 13 were previously found to be associated with IBD, CRC, IBS, obesity, or rectal bleeding and 8 of the 13 species were found to have multiple roles (**Appendix A**). A particular interest within this group of 13 bacteria were the species that have been studied *in vitro* or *in vivo* and found to potentially play a role in IBD such as *Ruminococcus gnavus, Flavonifractor plautii, Clostridium symbiosum,* and *Anaerostipes hadrus.* Out of the 21 remaining species, 16 were novel potential markers for IBD, 1 was previously found to be reduced in UC, and 4 were previously found to be elevated in healthy samples.

Discussion

This study identified numerous differences in taxonomic profiles, bacterial association networks, and genomic functional capacity between the IBD gut microbiome and the control gut microbiomes. Furthermore, our findings were corroborated by multiple external cohorts, and were generated using techniques and analyses that account for the compositionality of sequencing data. To our knowledge, this is the first study to utilize multiple external cohorts from a similar geographic region to corroborate comparisons between the internal control group and the diseased group in an analysis of the gut microbiome while also utilizing a compositionally robust methodology. Additionally, we demonstrated that bacterial species whose relative abundance is elevated in IBD are also present in the healthy microbiomes and maintain an important position in the healthy and IBD bacterial association networks implying that these species play an important role in the gut microbiome. However, these elevated bacteria are also often implicated in mucin degradation, immune system modulation, antibiotic resistance, and modulation of inflammation and their over-abundance may dysregulate these important processes possibly contributing to IBD pathogenesis and IBD-related symptoms.

We found that the IBD samples had alpha-diversities similar to internal controls (non-IBD), but significantly lower than external healthy controls. While It has previously been noted that IBD samples have lower alpha-diversity than healthy controls, we believe this may be due to the convenience selection of internal controls (Gevers,

Kugathasan, Denson, Vázquez-Baeza, Van Treuren, et al. 2014; Frank et al. 2011; Sheehan, Moran, and Shanahan 2015). As reported in Lloyd-Price et. al. 2019, the internal controls (non-IBD) consisted of “patients [who] were approached for potential recruitment upon presentation for routine age-related colorectal cancer screening, work up of other gastrointestinal (GI) symptoms, or suspected IBD, either with positive imaging (for example, colonic wall thickening or ileal inflammation) or symptoms of chronic diarrhoea or rectal bleeding”. However, due to ~75% of internal control samples being derived from subjects below the age of 45 (the earliest recommended age for colorectal cancer screening without personal or family history of colon cancer), it is presumed that the majority of these subjects presented with GI distress (Lloyd-Price et al. 2019) (**Figure 35**).

When examining the replicates present in the IBDMDB and Healthy-2 cohorts, it was noted that subjects diagnosed with IBD demonstrated increased temporal variability, as measured by the intrapersonal dissimilarity, when compared to non-IBD samples and Healthy-2 samples. This has been previously demonstrated when comparing CD and UC to non-IBD controls and has been posited to be caused by the inflammation and decreased intestinal transit time experienced by IBD patients as well as the medications and lifestyle changes employed to manage IBD (Clooney et al. 2020). It was also noted that the IBD and non-IBD samples displayed greater subject-to-subject variability relative to Healthy-1 and Healthy-2 samples. The relatively elevated temporal stability and subject-to-subject variability indicates that the gut microbiome of

our IBD samples displayed increased heterogeneity, relative to healthy controls. This has also been previously demonstrated in pediatric IBD patients and is believed to be caused by a depletion of core microbes, possibly due to inflammation and IBD therapies (Schirmer et al. 2018).

Much like the original publication utilizing the IBDMDB cohort (Lloyd-Price et. al. 2019), differentiating between the taxonomic profiles of IBD from non-IBD samples was difficult. In our study, using the RFC to classify IBD and non-IBD samples yielded many misclassifications in which non-IBD samples were consistently classified as IBD. The non-IBD samples were also misclassified as healthy. This split of RFC misclassifications for non-IBD samples indicates that the non-IBD group consists of a heterogeneous group that resembles both the IBD group, such as the subjects presenting with GI distress, and the healthy groups, such as the subjects presenting for routine screenings. It was also noted that the RFC utilizing the taxonomic profiles misclassified CD samples as UC samples and *vice versa*. This has also been previously demonstrated in other studies utilizing shotgun sequence data and is indicative of the high similarity demonstrated between the taxonomic profiles of the CD and UC gut microbiomes (Moustafa et al. 2018).

The RFC was able to distinguish between the external healthy cohorts and the IBD samples consistently and accurately, most likely due to these cohorts being composed of samples with no reported or overt disease. Our modified RFC framework

also allowed us to distinguish bacterial species that had a higher ranking than the random feature, based on the RFC feature importance's. These species were then used for differential abundance analysis, and network construction. While there was difficulty distinguishing the non-IBD sample taxonomic profiles from the IBD and healthy sample taxonomic profiles utilizing the RFC, we were able to distinguish 55 bacterial species that were significantly differentially abundant between the IBD and non-IBD groups. Of these 55 species, 42 were differentially abundant with a greater than 2-fold change in the external cohorts as well.

The bacterial association networks revealed that while some bacteria were found to be elevated in IBD, they were still present in non-zero degree in non-IBD and healthy networks. As a matter of fact, the species elevated in IBD displayed higher than average degree in all networks except for the Healthy-1 network. Furthermore, when examining the most important nodes (top-10 eigenvector centrality) within the IBD network, 8 out of the 10 species were also found in the top-10 eigenvector centrality (EVC) nodes of the healthy networks but all 10 of the top EVC species were found to have relative abundances that are elevated in IBD samples. The presence and importance of species that are elevated in IBD appears to be ubiquitous throughout all networks implying that while these species have an increased relative abundance in IBD, they still play integral roles within the non-IBD and healthy microbiomes, and that it is their over-abundance and not mere presence that plays an important role in IBD. Interestingly, while bacteria with elevated relative abundances in IBD were present and

appeared to play an important role in the non-IBD and healthy networks, they also demonstrated many associations unique to the IBD network illustrating that some bacterial species can associate with different bacteria due to factors other than just the presence of the bacteria. This implies that other factors, such as host genetics, host diet, intestinal environment, or medications may lead to the unique associations (Pérez-Gutiérrez et al. 2013; Ohland and Jobin 2015).

It was also noted that the majority of species within each network were not differentially abundant between IBD and the control groups (IBD: 52.5%, non-IBD: 52.6%, Healthy-1: 65.6%, Healthy-2: 53.8%). This is an interesting finding demonstrating that the majority of gut microbiome network constituents are similar in relative abundance between healthy and IBD gut microbiomes. Furthermore, we observed that these non-differentially abundant bacteria accounted for greater than 60% of the relative abundances in all groups (IBD: 62.6%, non-IBD: 70.5%, Healthy-1: 74.6%, Healthy-2: 64%). Most bacterial association networks and most of the gut microbiome were composed of bacteria that are not significantly differentially abundant between the IBD and control gut microbiomes indicating that the differences in the IBD gut microbiome are not wide-spread and appear to be limited to a set of bacterial species with significantly higher relative abundance. Interestingly, it was also observed that the majority of negative associations found in all networks were associated with species displaying elevated relative abundance in IBD samples (IBD network: 100%, non-IBD: 100%, Healthy-1: no negative edges, Healthy-2: 81.6%). This finding indicates

that the bacterial species elevated in IBD may play an important role in maintaining stability, possibly by preventing positive feedback loops, but due to their overabundance in IBD they may contribute to reducing the diversity of the gut microbiome in IBD samples (Coyte, Schluter, and Foster 2015).

When analyzing the protein family relative abundances in each cohort, we were not able to identify any statistically significant differences in functional roles between the IBD and non-IBD group. However, we were able to find 6 significantly different functional roles between the IBD group and each of the external control cohorts. Notably, the protein family role most elevated in IBD, relative to external healthy controls, was associated with functions related to sporulation and germination. While sporulation in the context of GI disease is most often associated with *Clostridium difficile,* many members, especially pathogens, of the *Clostridia* genus have been found to utilize sporulation which is in-line with our data demonstrating that the *Clostridium* genus is the most commonly elevated genus in IBD *(Hookman and Barkin 2009; Shen et al. 2019).* Our analysis also demonstrated that protein families involved in polysaccharide metabolism were elevated in IBD. This may be due to the increase in relative abundance of some bacteria that inhabit the intestinal mucosa and degrade mucin to derive glycans as an energy source, such as *Ruminococcus gnavus* and *Clostridium symbiosum (Bernalier-Donadille 2010; Hall et al. 2017; Desai et al. 2016*). It was also found that protein families involved in molybdopterin synthesis were significantly elevated in IBD. Molybdopterin is an important co-factor for nitrate

reductase, which reduces nitrate to nitrite (Moreno-Vivián et al. 1999). Previous research has identified nitrite as an important molecule in the regulation of mucosal blood flow, intestinal motility, and mucus membrane thickness, however it believed that an over-abundance of nitrite can have deleterious effects on commensal bacteria and has been shown to be associated with IBD as well as with increased bleeding (Lidder and Webb 2013; Tiso and Schechter 2015; Park et al. 2013). This may indicate that an increase in nitrate reduction (leading to increased nitric oxide levels) can contribute to negative selection against commensal bacteria as well as contribute to increased propensity of intestinal bleeding in IBD. Nitric oxide, the main metabolite of nitrite, is also believed to be able to increase intestinal motility and lead to diarrhea (Kukuruzovic et al. 2003).

We also observed that protein families involved in the synthesis of quinones (menaquinone and ubiquinone) were reduced in IBD. Quinones are believed to be important growth factors for gut microbiota, especially for bacteria seen as commensals (Fenn et al. 2017). Humans are also unable to synthesize menaquinone (Vitamin K) and thus must ingest it or have it produced by commensal bacteria indicating that a reduction in vitamin K synthesis by the gut microbiota may lead to a reduction of vitamin K levels in IBD (Walther et al. 2013). In fact, IBD research has long noted that IBD patients present with lower vitamin K levels (Schoon et al. 2001; Krasinski et al. 1985). Due to the important role of vitamin K in blood clotting and calcium binding, this reduction on vitamin K has been used to explain common co-occurrences and

symptoms of IBD such as osteoporosis and bleeding (Schoon et al. 2001; Agnello et al. 2014). Quinone synthesis appears to play an important role in maintaining host health and its reduction may contribute to the increased intestinal and rectal bleeding common in IBD.

Finally, we were able to identify specific bacterial species that are elevated in IBD and play important roles in fomenting inflammation, degrading mucin, and antibiotic resistance. *R. gnavus* and *C. symbiosum* are mucin-degrading bacteria that are found in healthy gut microbiomes but are shown to be elevated in IBD gut microbiomes (Crost et al. 2016). These bacteria may play an important role in preventing the over-secretion of mucus in healthy gut microbiomes, but their over-abundance may cause the mucus layers in the intestine to become too thin. We also identified *Flavonifractor plautii* as a species that was elevated in IBD. *F. plautii* has been found to degrade flavinoids, an important anti-inflammatory mediator in humans and mice (Musumeci et al. 2020). The over-abundance of *F. plautii* can lead to low levels of flavonoids which has been shown to lead to increased inflammation, particularly in the gut microbiome (Gupta et al. 2019). *R. gnavus, C. symbiosum,* and *F. plautii* are key examples of bacterial species that are present, and potentially important, in healthy microbiomes but appear to exhibit deleterious effects on host health when they become over-abundant.

While we attempted to mitigate as many confounders under our control as possible, there are still limitations to be cognizant of within our study. One very important limitation stems from the relatively low number of subjects present in the datasets we utilized. We previously demonstrated that as the sample-to-taxa ratio increases, our network inference framework generates better predictions (Loftus, Hassouneh, and Yooseph 2021), however, due to the low number of unique individuals it was necessary to construct the networks using the replicates as individual samples. While we have demonstrated that the intrapersonal variation is lower than the interpersonal variation, we do not believe that this has a negative effect on the accuracy of the networks inferred. Due to the assumption that the samples in a group are all generated using the same underlying covariance structure, it is reasonable to include subject sample replicates for network inference. Another potential limitation was a bias towards samples from younger subjects in the IBDMDB cohort. Approximately half of (46.6%) IBDMDB samples were derived from subjects below the age of 18 (**Figure 35a**) and the youngest subject was 6 years of age. In contrast, no subjects in the Healthy-2 cohort were below the age of 18 (**Figure 36**). While we did not have access to the metadata (other than sex) of the Healthy-1 cohort, it was previously published that all subjects fell between the ages of 18-40 (Methé et al. 2012). The feature 'age' also displayed the greatest feature importance during classification according to our RFC framework, indicating that there was a non-trivial difference in the ages between the diagnosis groups. It has been previously observed that the taxonomic profiles of individuals begin to resemble adult configurations by 3 years of age, indicating that the bias is unlikely to contribute to major differences in the taxonomic profiles and may just

be indicative of the younger age of subjects in the IBDMDB study (Yatsunenko et al. 2012). However, the same study did note that while interpersonal variation greatly decreased after 3 years of age, it was still significantly higher in subjects between the ages of 3-17, relative to adults (18+ years of age), which may explain some of the difference in interpersonal variability observed between IBD and non-IBD samples, relative to the Healthy-1 and Healthy-2 samples. Finally, it was noted that there was a greater proportion of female subjects in the Healthy-2 cohort relative to the Healthy-1 and IBDMDB cohorts (**Figure 37**). This does not appear to impact the classification results however, as the RFC did not find the features 'sex' to be more important than random noise.

By utilizing two external control cohorts, we were able to identify and corroborate 34 bacterial species whose relative abundance is significantly elevated in IBD. These species appear to play important roles in all bacterial association networks (IBD, non-IBD, and external healthy controls) implying that while an elevation of their relative abundance is associated with IBD, they are also important to the function of healthy gut microbiomes. Furthermore, we identified important differences in functional capacities between IBD and the healthy controls that may contribute to the onset or exacerbation of IBD-related symptoms such as diarrhea, intestinal bleeding, mucin degradation, and intestinal inflammation. Finally, we were able to corroborate many of the bacterial species we identified as elevated in IBD using previously published research and identified 17 novel bacterial species that may play an important role in IBD. To the best

of our knowledge, we are the first to corroborate our analysis of the IBD gut microbiome by using external cohorts from the same geographic region (US) allowing us to generalize our findings to the population rather than only our study groups. Furthermore, we were able to illustrate important potential mechanistic links between the bacterial species elevated in the IBD gut microbiome and IBD-related symptoms. Finally, we identified differences in the genomic functional capacity of the IBD microbiome that bridges previous findings in IBD and IBD-related symptoms with the gut microbiome.

Figures

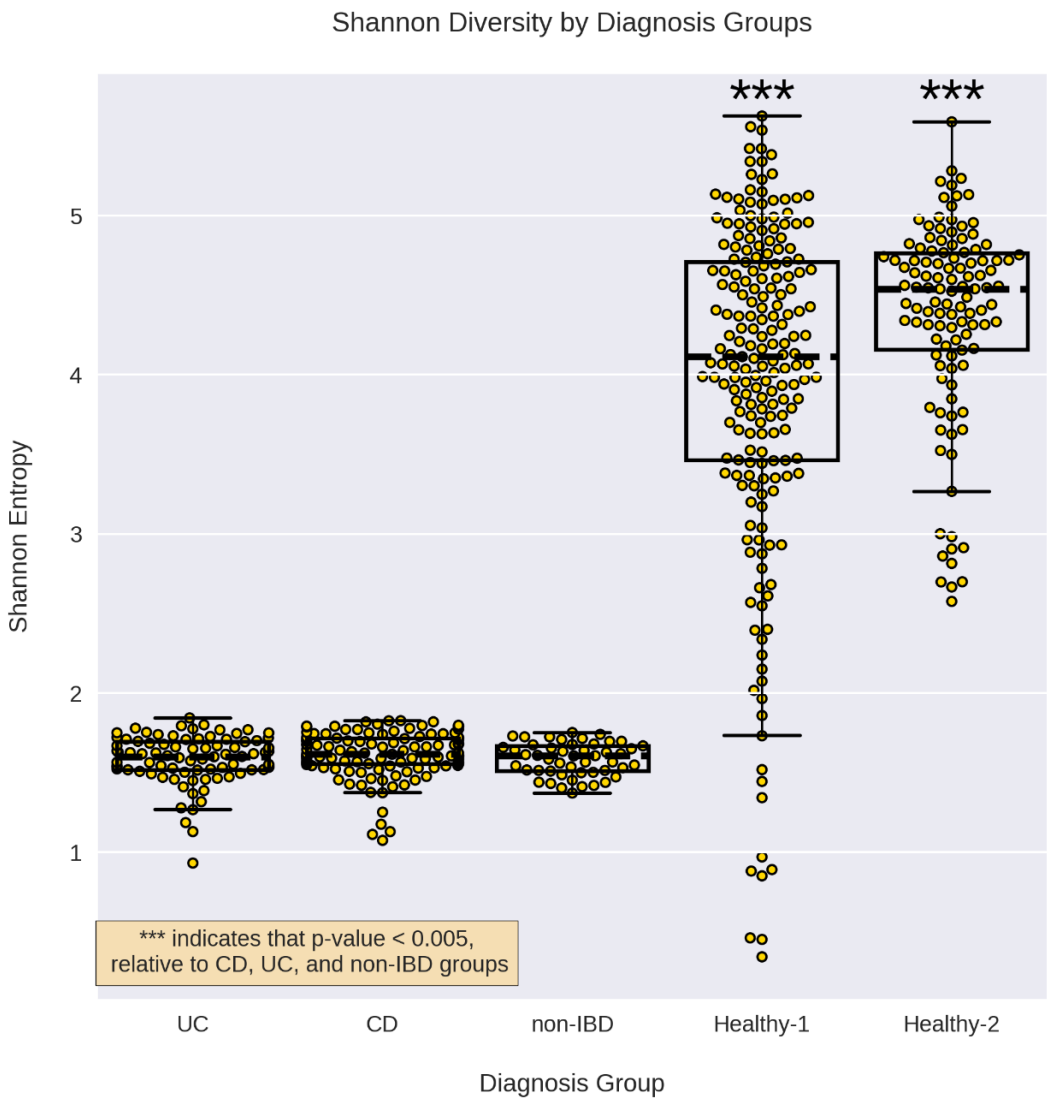


Figure 23: Alpha-diversity of sample groups.

*Alpha diversity for each sample group by was calculated using Shannon entropy. The alpha-diversity for CD, UC, and non-IBD were not significantly different from each other but all three were significantly lower than the healthy cohorts. *** indicates a p-value < 0.0005 compared to CD, UC, and non-IBD, using a Mann-Whitney U test.*

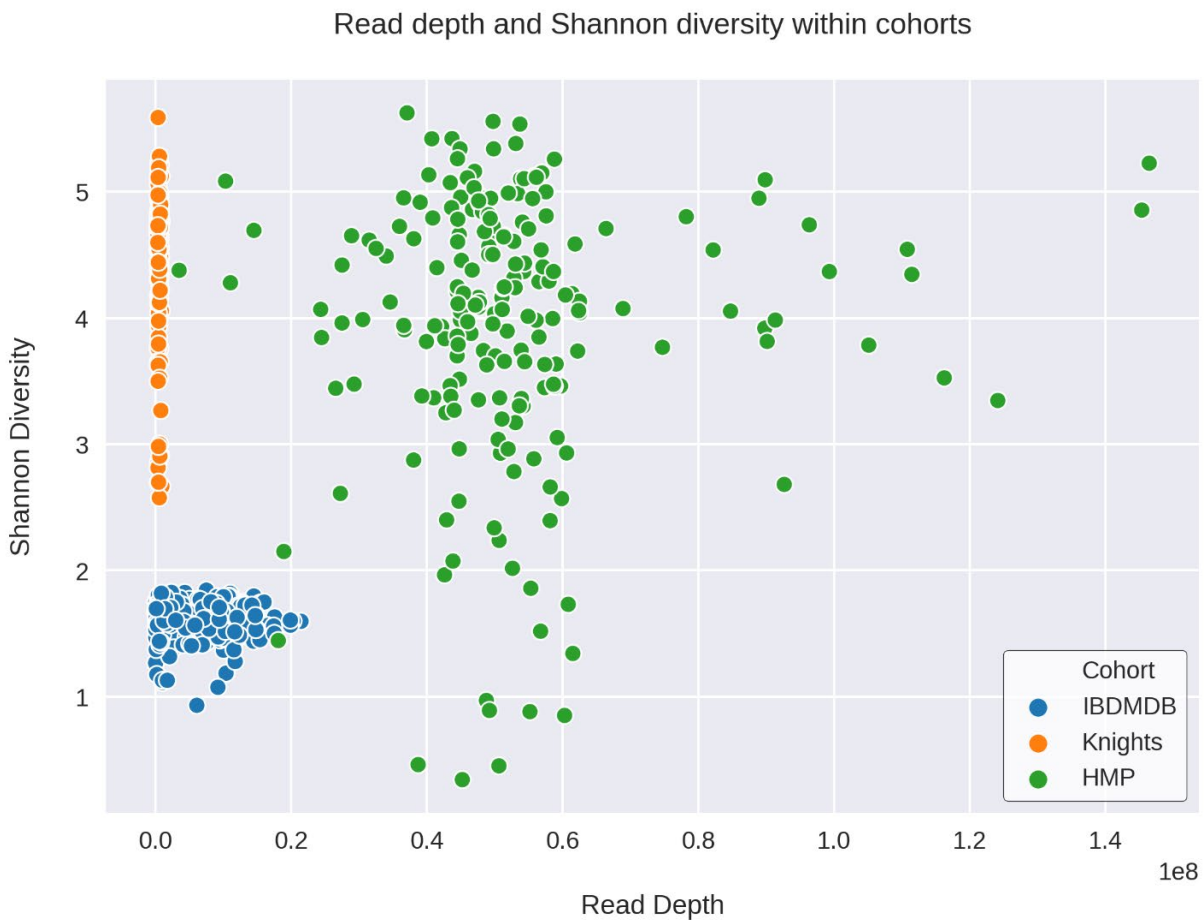


Figure 24: Effect of read depth on Shannon diversity.

To illustrate that the alpha-diversity value differences were not due to read depth,

the read depth was plotted against the alpha-diversity (Shannon entropy). Read depth did not appear to have an effect on alpha-diversity and notable the Healthy-2 cohort had lower read depth but displayed greater alpha-diversity.

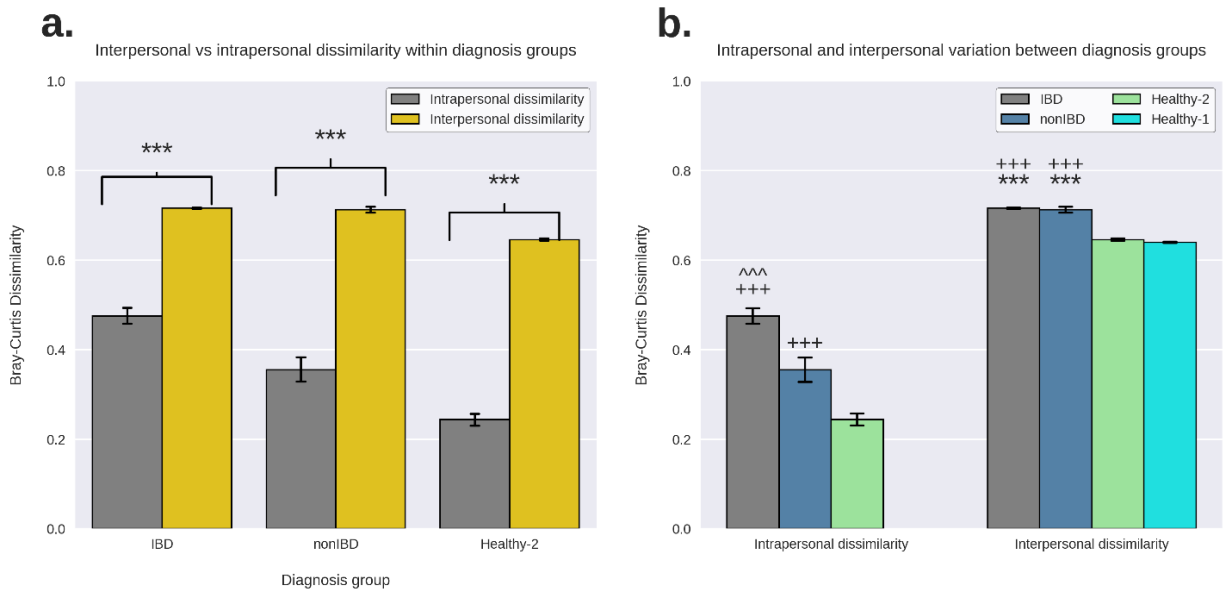


Figure 25: Intrapersonal and interpersonal variation.

*Bray-Curtis dissimilarities between replicate samples from the same subject were used to quantify intrapersonal variation while the Bray-Curtis dissimilarities between samples from different subjects were used to quantify interpersonal variation. a. It was observed that replicate samples from the same subject were significantly more similar to each other than to samples from other subjects. *** indicates a p-value < 0.0005. b. IBD samples demonstrated elevated intrapersonal variation relative to non-IBD and Healthy-2 samples and non-IBD samples demonstrate elevated intrapersonal variation relative to Healthy-2 samples. IBD and non-IBD samples both demonstrated elevated interpersonal variation relative to the Healthy-1 and Healthy-2 samples. *** indicates p-value < 0.0005, relative to Healthy-1. +++ indicates p-value < 0.0005, relative to Healthy-2. ^^^ indicates a p-value < 0.0005, relative to non-IBD.*

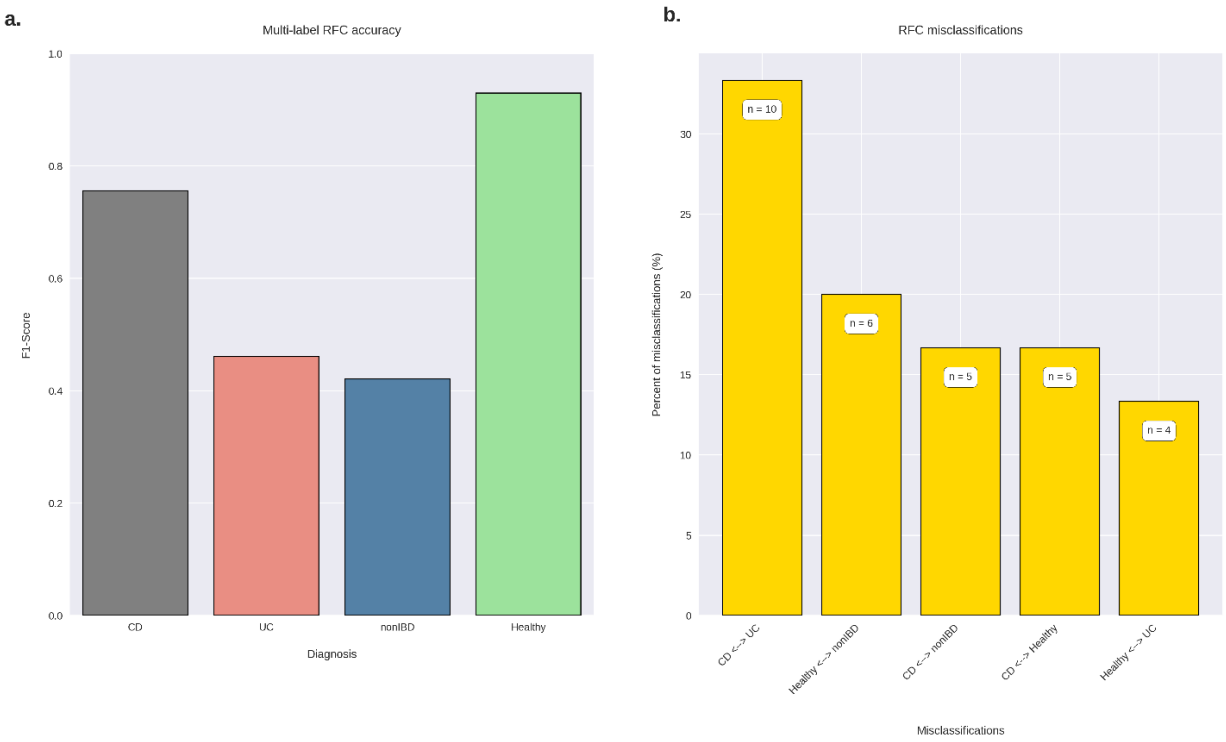


Figure 26: Classification accuracy and misclassification before combining CD and UC into one IBD group.

An RFC was trained on the taxonomic profiling data and metadata (age, sex, unique subject ID) for all groups without combining the CD and UC groups into one group. a. The RFC demonstrated poor classification accuracy when attempting to distinguish all groups (CD, UC, non-IBD, and Healthy), especially when attempting to classify the CD, UC, and non-IBD groups. b. The largest amount of misclassifications occurred due to the RFC classifying CD samples as UC and vice versa implying that the taxonomic profiles of both groups were very similar.

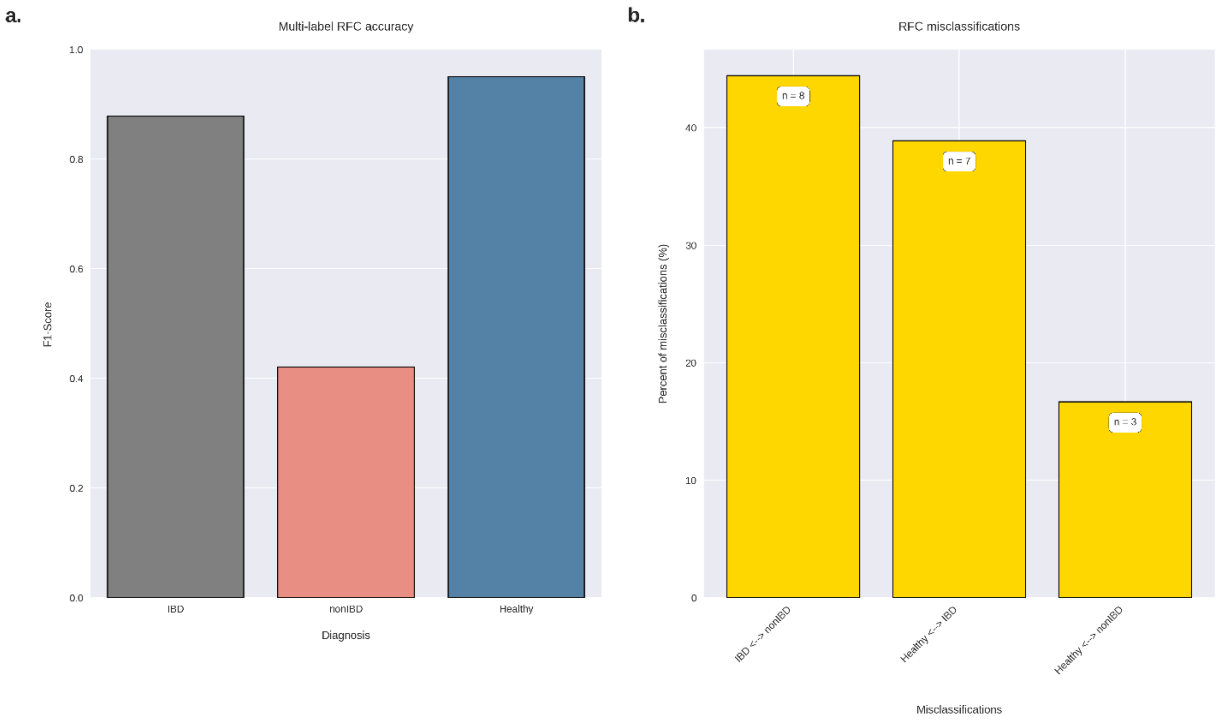


Figure 27: RFC classification accuracy by diagnosis label after grouping CD and UC as IBD. An RFC was trained on the taxonomic profiling data and metadata (age, sex, unique subject ID) for all groups after combining the CD and UC groups into one group, known as the IBD group. a. The RFC demonstrated better classification accuracy (weighted average of 0.87) compared to previous RFC without combining CD and UC samples under the IBD umbrella (weighted average of 0.79), however, non-IBD samples were still difficult to classify. b. The non-IBD samples were still consistently misclassified and were split (with a bias towards being classified as IBD) between being classified as IBD and Healthy samples.

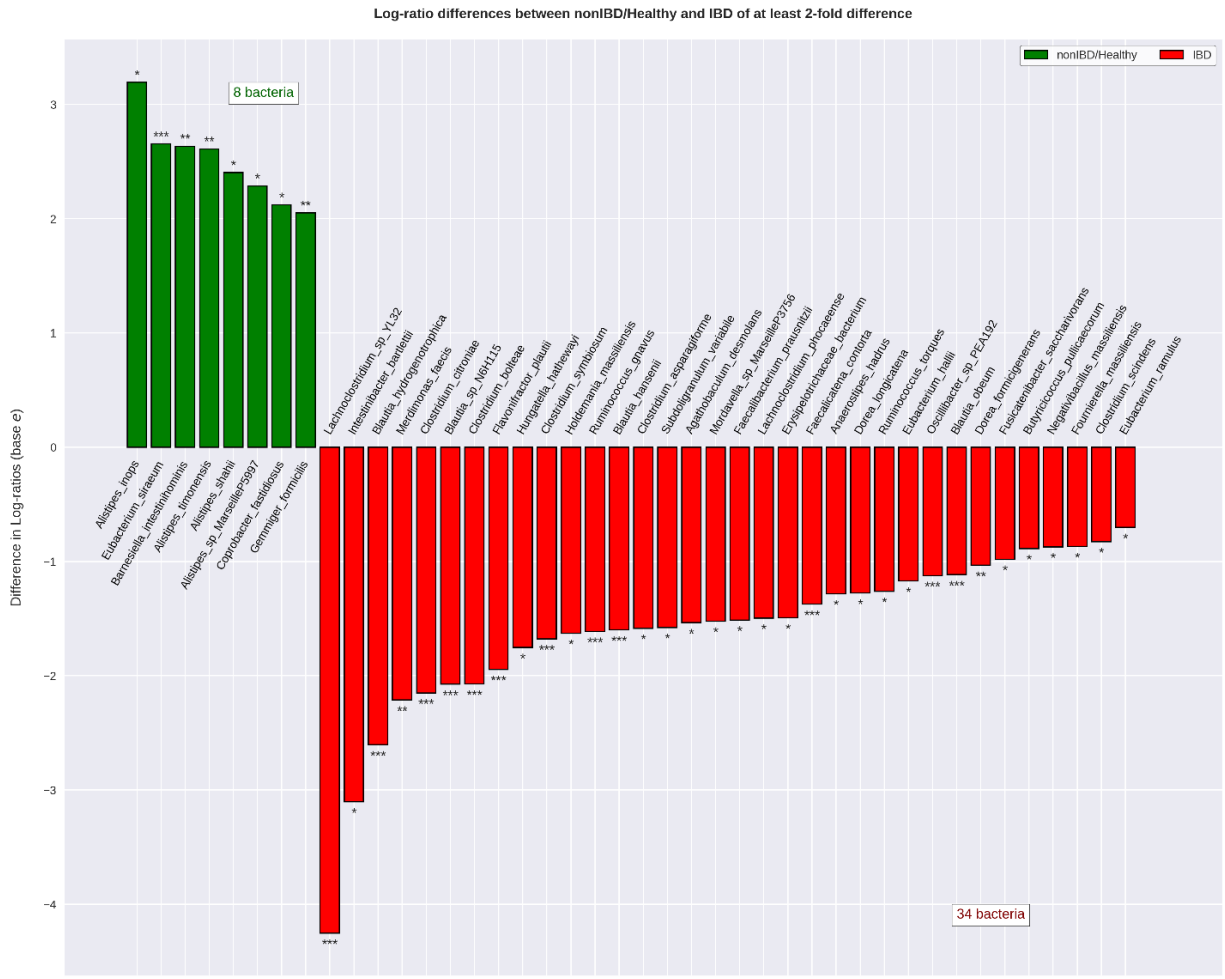


Figure 28: Differential abundance of bacterial species when comparing IBD to non-IBD and healthy groups.

*Relative abundance values were CLR-transformed and differential abundance was calculated for IBD compared to non-IBD, IBD compared to Healthy-1, and IBD compared to Healthy-2. The species that were significantly differentially abundant in IBD relative to every single other group were considered to be significantly differentially abundant resulting in 42 significantly differentially abundant bacterial species. Of these 42 species, 34 were found to be significantly more abundant in IBD relative to every other group and 8 bacterial species were found to be significantly less abundant in IBD, relative to every other group. The transformed relative abundances were then averaged and displayed under one label (non-IBD/Healthy) for ease of visualization. * indicates a q-value < 0.05. ** indicates a q-value < 0.01. *** indicates a q-value < 0.001.*

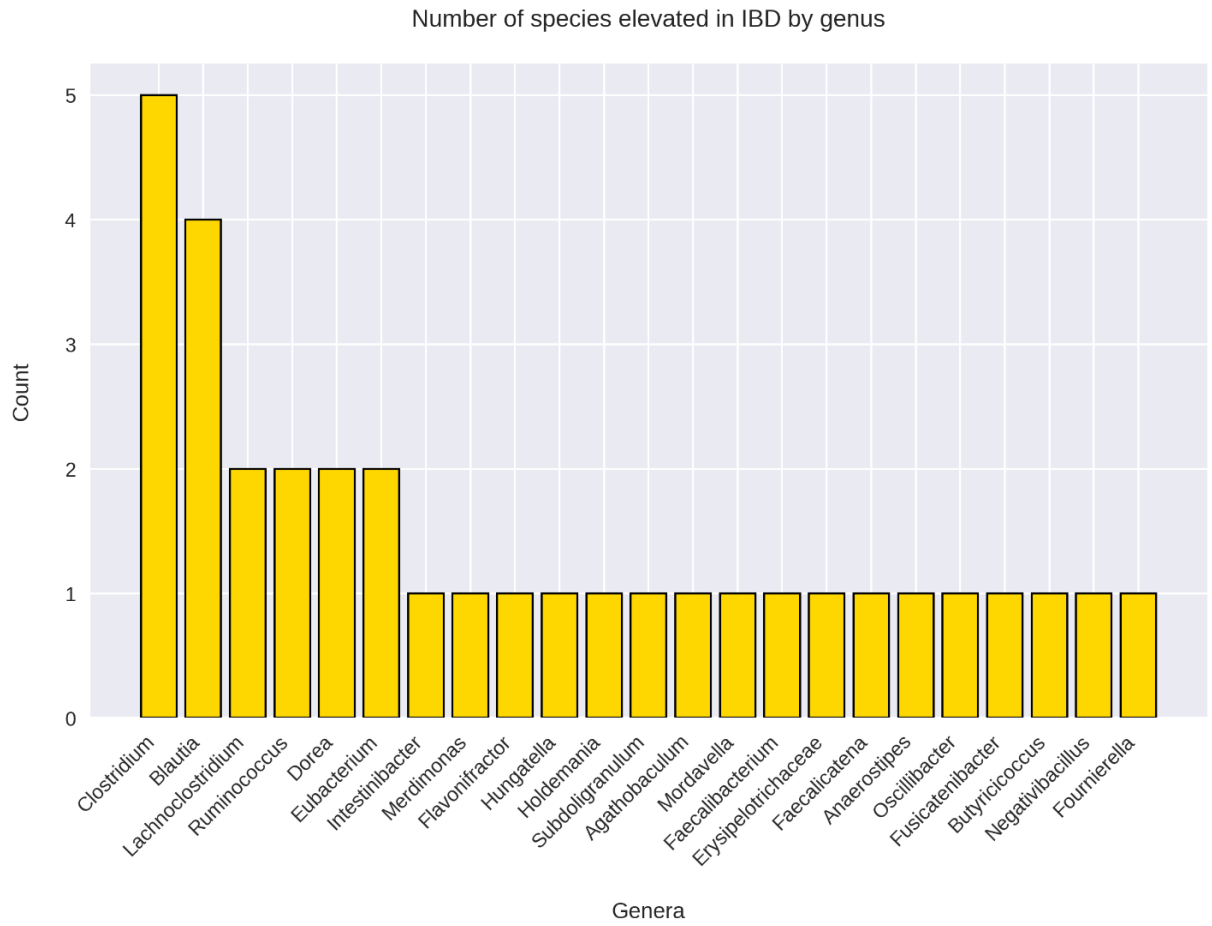


Figure 29: Genera counts of bacteria elevated in IBD.

The genera of the bacteria that were elevated in IBD, according to the differential abundance analysis, were counted. Clostridium and Blautia were the most commonly elevated genera in IBD samples. Lachnoclostridium, Ruminococcus, Dorea, and Eubacterium were the only other genera to have more than one member elevated in IBD.

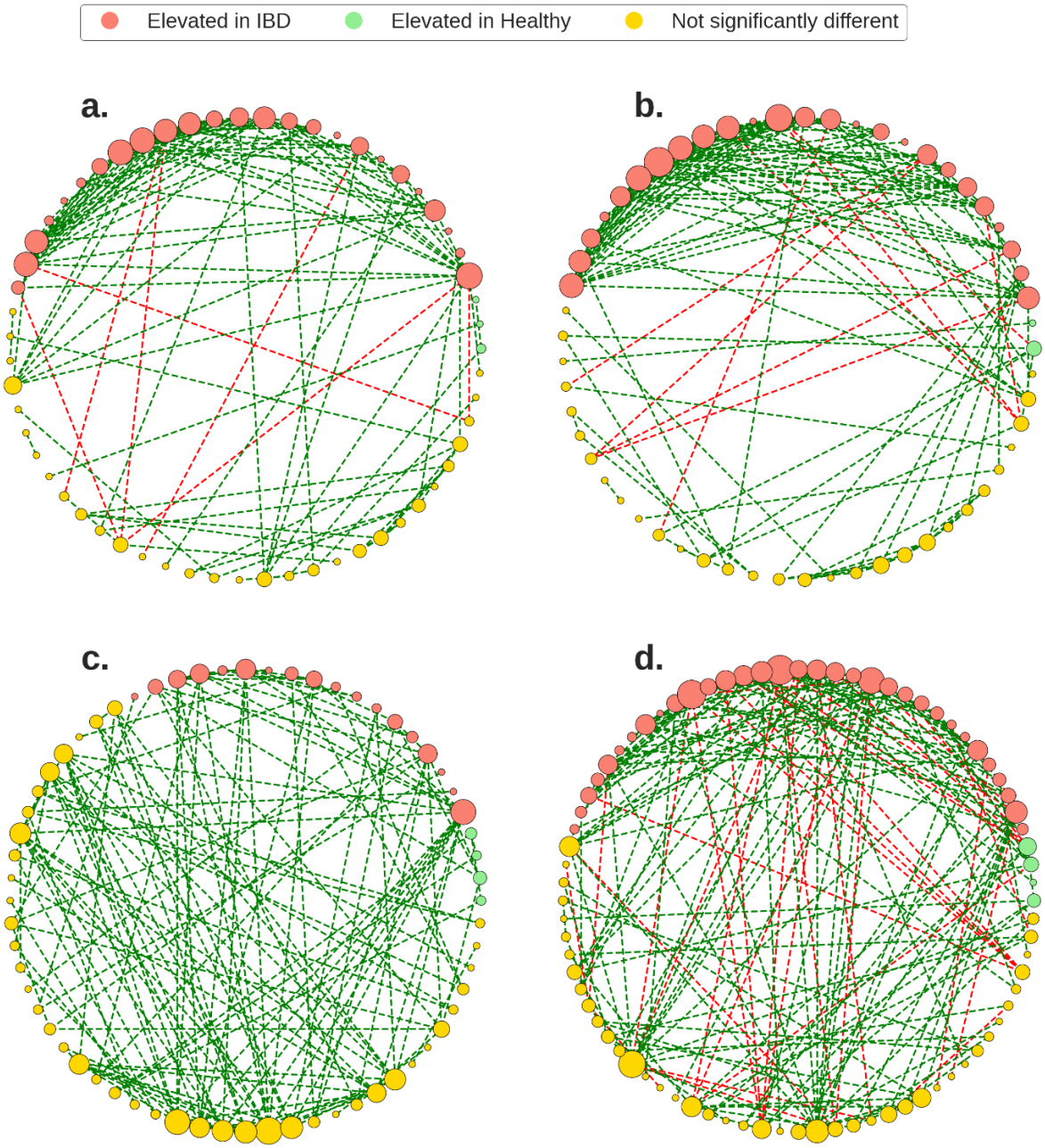


Figure 30: Gut microbiome bacterial association networks.

The GGM framework was used to generate bacterial association networks from CLR-transformed relative abundances. The bacterial species (nodes) were colored based on the differential abundance analysis and the node sizes were based degree of each node within the network. Bacterial species that were elevated in IBD were still present, and in high degree, in the non-IBD and healthy networks. It was also noted that the most common constituents of the bacterial association networks were bacteria that were not significantly differentially abundant in IBD, relative to the healthy and non-IBD groups. Finally, the IBD networks demonstrated

more negative edges when compared to the non-IBD and healthy groups. a: IBD network, b: non-IBD network, c: Healthy-1 network, d: Healthy-2 network.

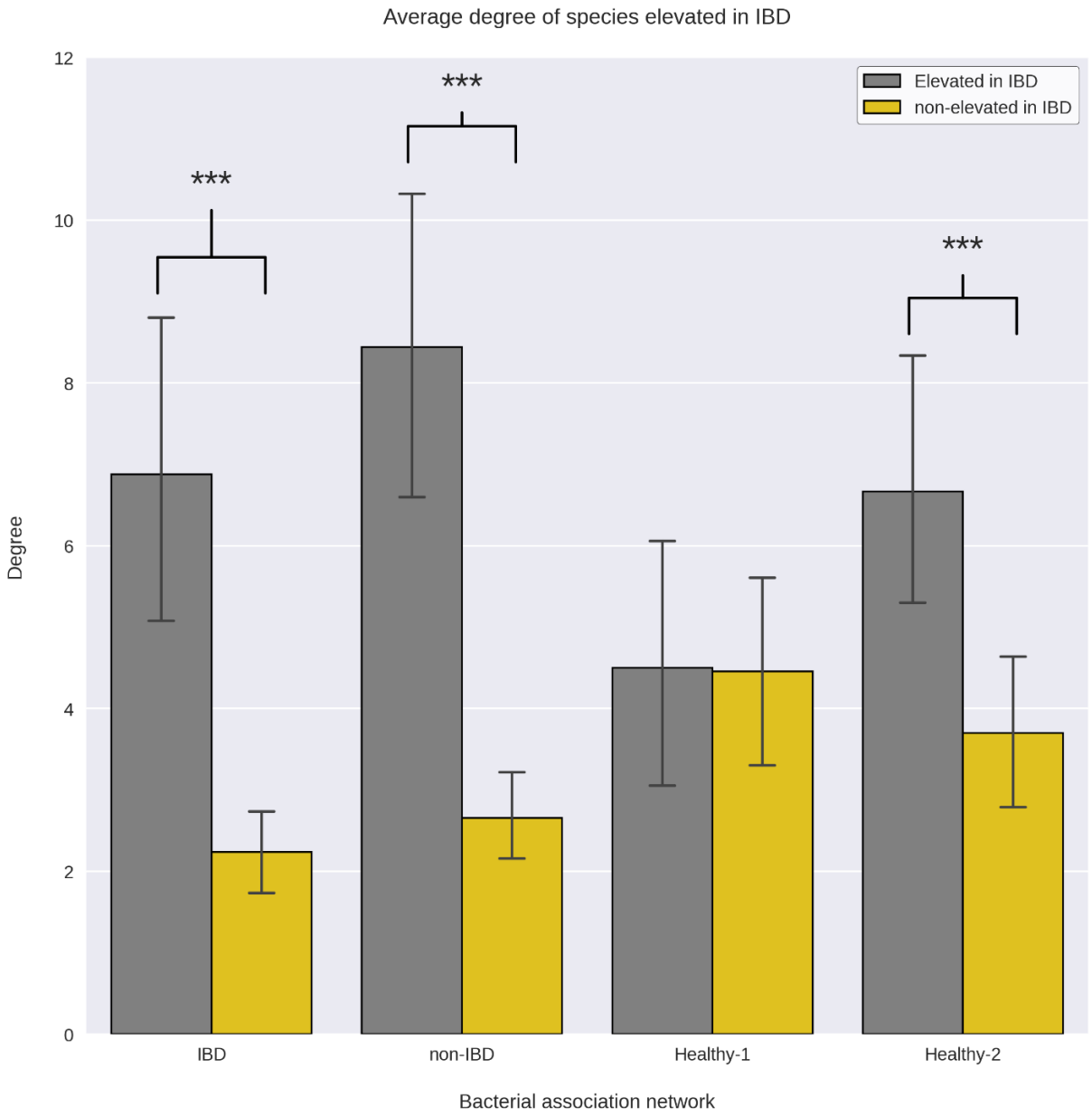


Figure 31: Average degree of bacterial species that are elevated in IBD within each network.

The average degree of the bacterial species that were elevated in IBD was calculated for each group and compared to the average degree of species not elevated in IBD (elevated in Healthy or not significantly different). On average, the species elevated in IBD displayed a higher number of connections (degree) within the bacterial association networks of all diagnosis groups and was significantly higher in 3 out of the 4 networks.

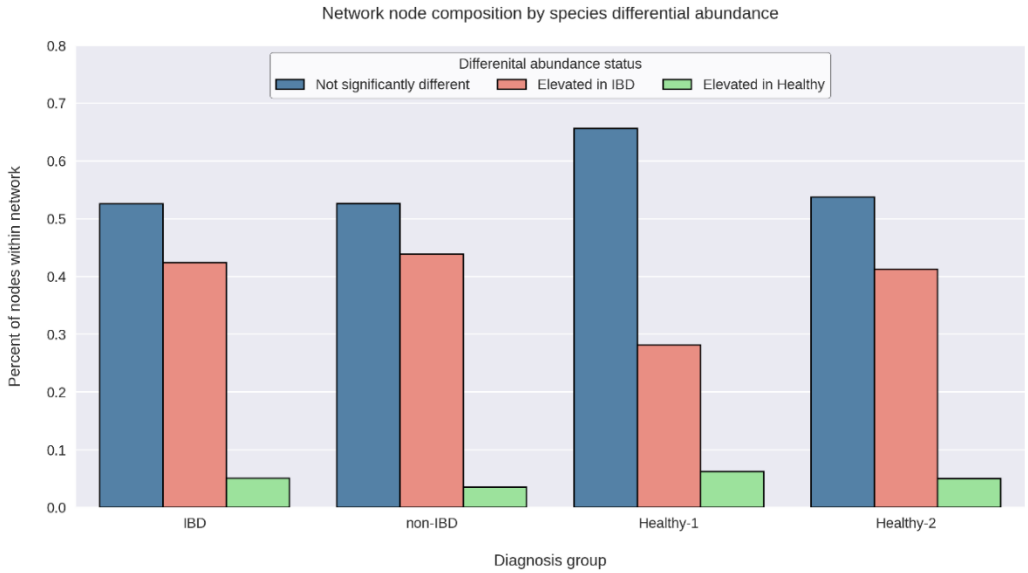


Figure 32: Network node compositions.

The Majority of nodes for each network are composed of species that are not differentially abundant between IBD and the other diagnosis groups.

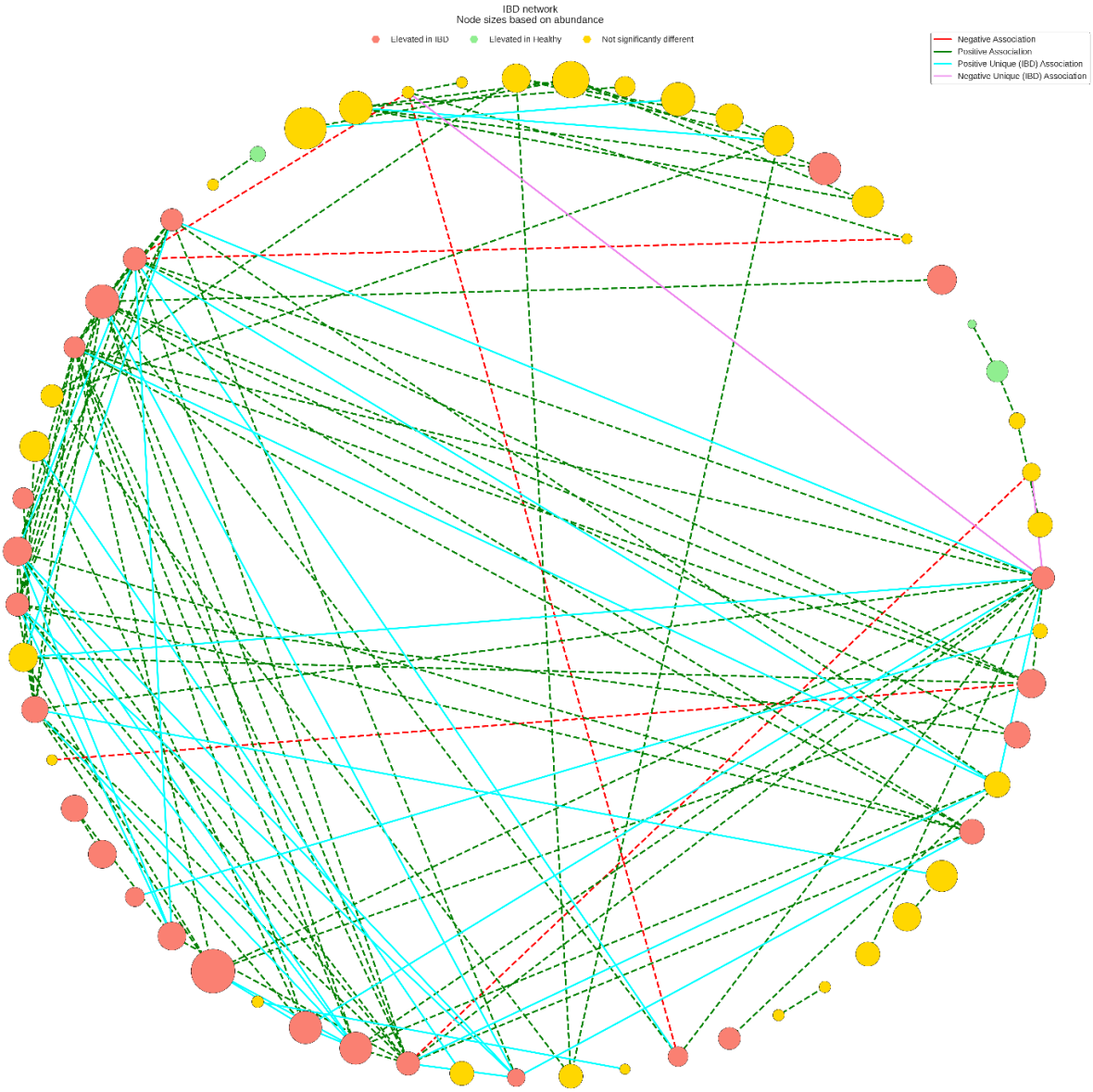


Figure 33: Unique associations within the IBD bacterial association network.

When comparing the structure of the IBD bacterial association network to all other networks, 56 associations were found that were unique to IBD networks only. The majority (85.7%) of these associations involved bacteria elevated in IBD. Even though the bacteria elevated in IBD are also present in the control networks, and in high degree, they do appear to demonstrate different associations in the IBD network.

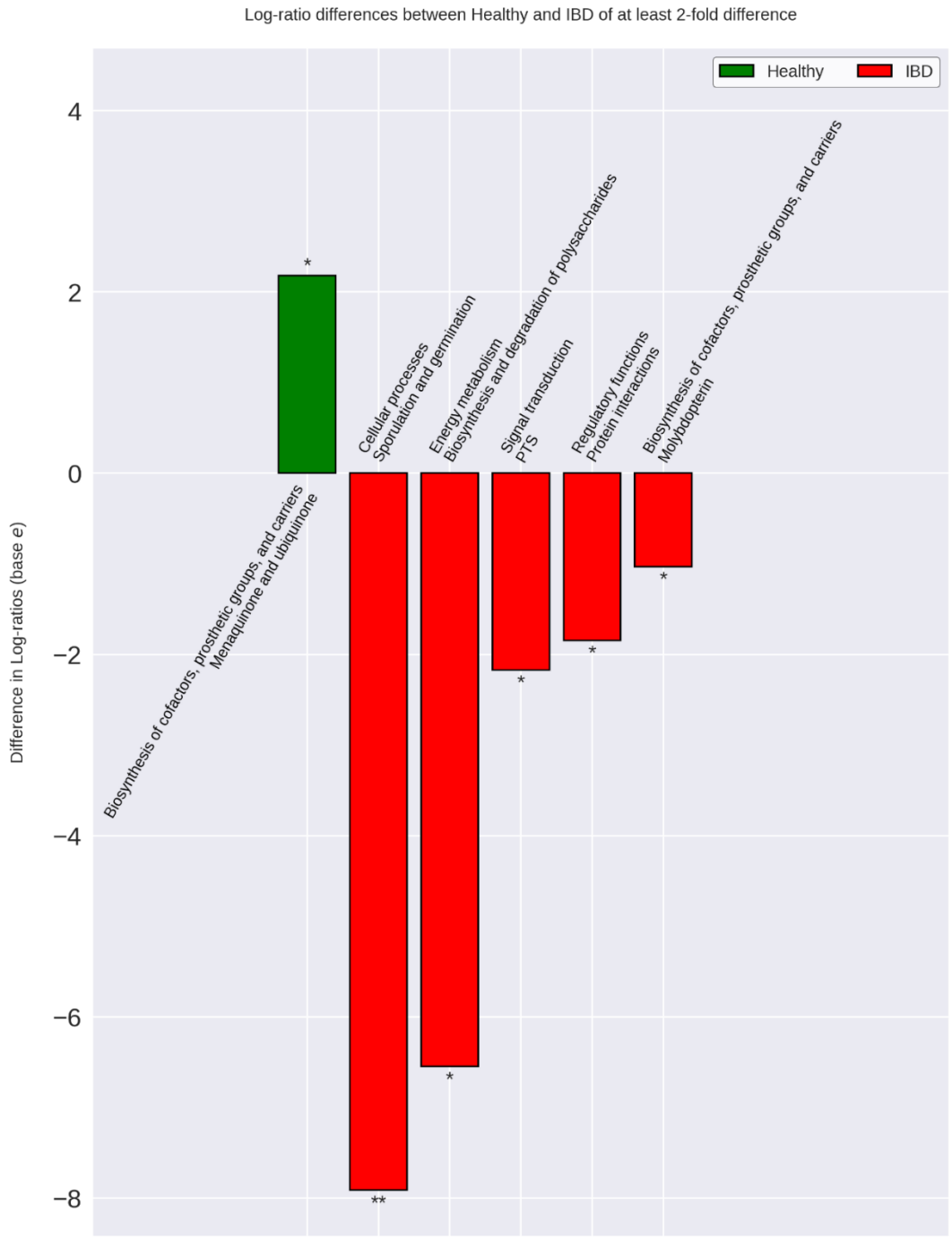


Figure 34: Differences in IBD gut microbiome functional capacity.

Genomic functional capacity was determined by using the TIGRFAM protein family database. The counts for each TIGRFAM within a bacterial species were weighted by the relative abundance of the bacterial species within each group. The CLR-transformed relative abundance of the TIGRFAM's within IBD were then compared to the non-IBD, Healthy-1 cohort, and Healthy-2 cohort individually. The differentially abundant TIGRFAM's were then summed based

*on their roles, according to the TIGRFAM database. There were no differences between the IBD and non-IBD gut microbiome functional capacities. There were 6 significantly differentially abundant protein family roles when comparing IBD to the Healthy-1 cohort that were also found in the Healthy-2 cohort. These differences are implicated in important processes that may contribute to IBD-related symptoms such as diarrhea, intestinal bleeding, and increased intestinal permeability. The relative abundances of the Healthy-1 and Healthy-2 cohort TIGRFAM roles were averaged for ease of visualization. * indicates a p-value < 0.05. ** indicates a p-value < 0.01. *** indicates a p-value < 0.001.*

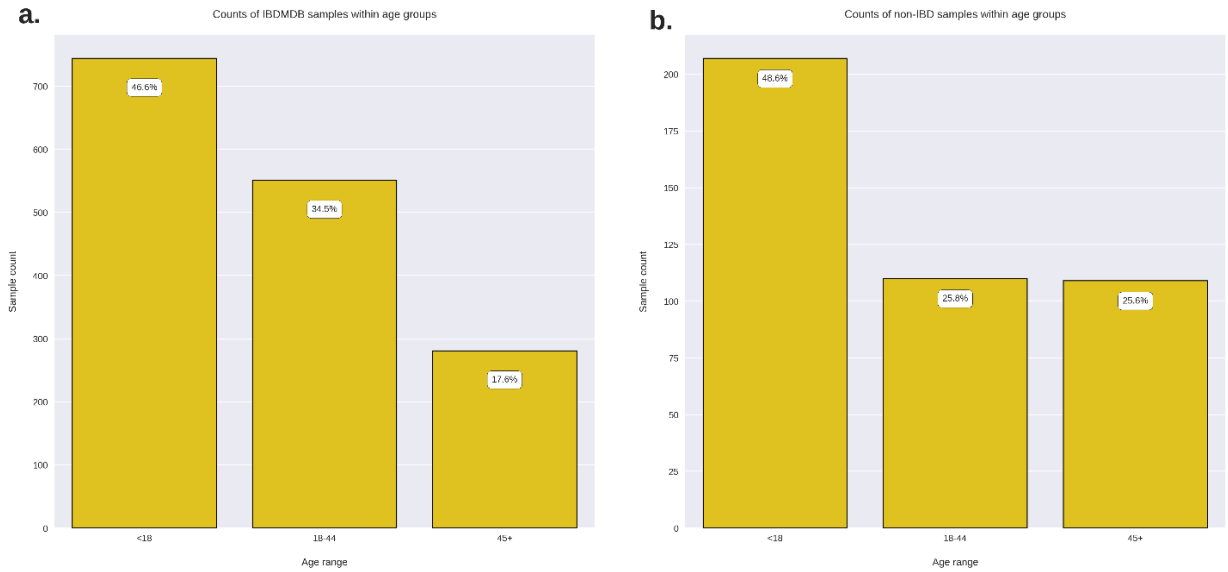


Figure 35: Age groups within the IBDMDB cohort and non-IBD samples.

The counts of age ranges for subjects from the IBDMDB cohort were plotted. a.

Almost half (46.6%) of samples in the IBDMDB cohort were from subjects below the age 18. b. The majority of non-IBD samples were derived from subjects that fell below the age of 45, the recommended age for colorectal cancer screening. Due to the description of subject recruitment from the original publication (Lloyd-Price et. al. 2019), it is presumed that the majority of the control group were comprised of samples that presented for GI distress or suspected IBD.

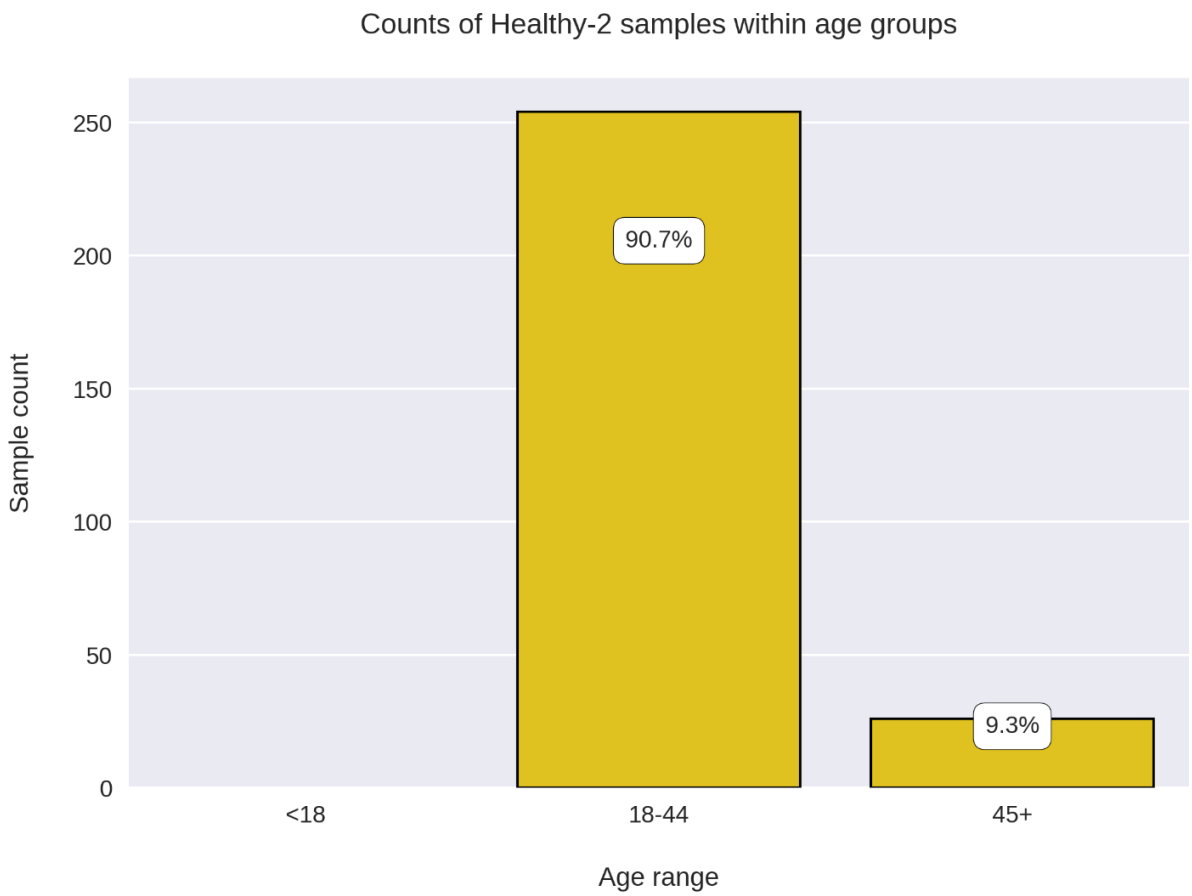


Figure 36: Age ranges of Healthy-2 samples.
The age range for subjects of the Healthy-2 cohort were plotted. The vast majority of subjects are between 18 and 44 and no samples are below 18 years of age.

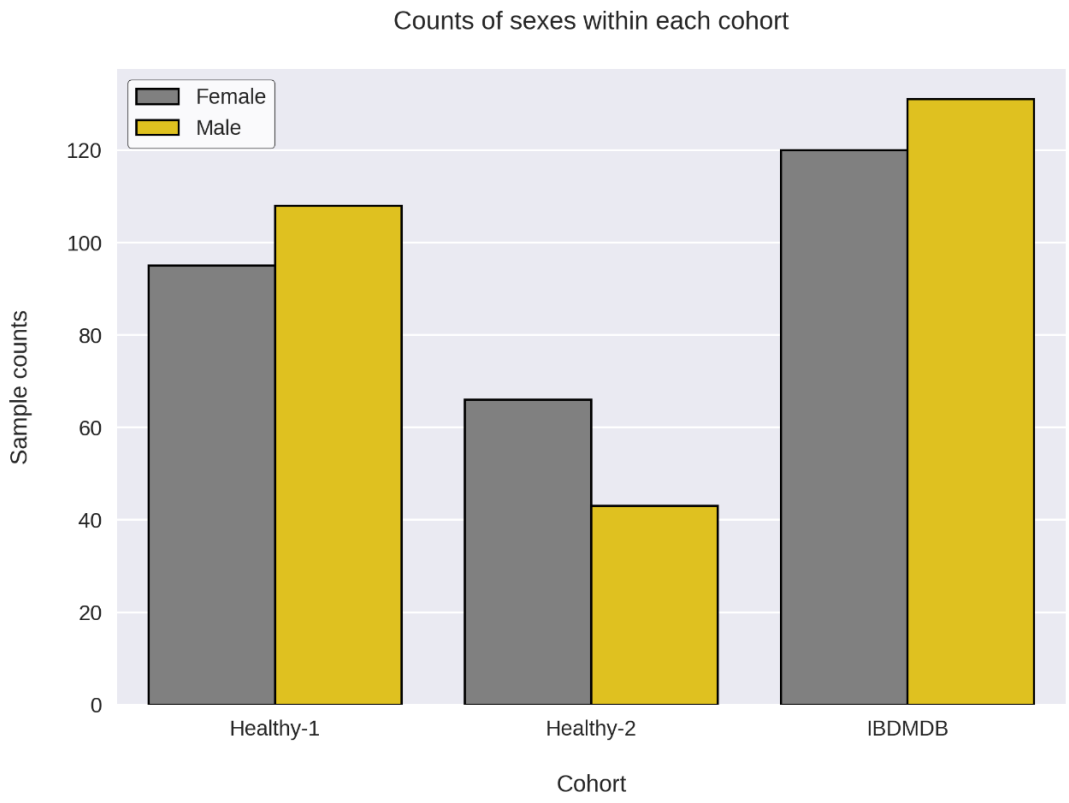


Figure 37: Sex counts by cohort.

The counts of samples from a subject of a given sex were plotted for each cohort.

There appears to be a greater proportion of females in the Healthy-2 cohort relative, but the IBDMDB and Healthy-1 cohorts appear to have similar proportions of each sex.

Tables

Table 2: Top-10 eigenvector centralities (EVC) per sample group.

Eigenvector centrality (EVC) of each node (bacterial species) in the IBD network was calculated. The ten nodes with the highest EVC in IBD were also found in the top-ten EVC nodes of the non-IBD, Healthy-1, or Healthy-2 networks except for Fusicatenibacter saccharivoran and Blautia hansenii.

Species	EVC	Diagnosis Group
Blautia_sp_N6H115	0.342211	IBD
Hungatella_hathewayi	0.313393	IBD
Blautia_obeum	0.304794	IBD
Clostridium_citroniae	0.289837	IBD
Clostridium_bolteae	0.283886	IBD
Agathobaculum_desmolans	0.274133	IBD
Fusicatenibacter_saccharivorans	0.266944	IBD
Blautia_hydrogenotrophica	0.258565	IBD
Clostridium_symbiosum	0.255309	IBD
Blautia_hansenii	0.202126	IBD
Blautia_sp_N6H115	0.357233	non-IBD
Blautia_obeum	0.297542	non-IBD
Clostridium_symbiosum	0.294405	non-IBD
Hungatella_hathewayi	0.281785	non-IBD
Blautia_hydrogenotrophica	0.265395	non-IBD
Clostridium_bolteae	0.252532	non-IBD
Ruminococcus_gnavus	0.25097	non-IBD
Clostridium_citroniae	0.247998	non-IBD
Anaerostipes_hadrus	0.20174	non-IBD
Agathobaculum_desmolans	0.201307	non-IBD
Bacteroides_fluxus	0.368378	Healthy-1
Bacteroides_caecimuris	0.33475	Healthy-1
Bacteroides_plebeius	0.322128	Healthy-1
Bacteroides_coprocola	0.255504	Healthy-1
Bacteroides_coprophilus	0.251902	Healthy-1
Bacteroides_ovatus	0.250976	Healthy-1
Butyricimonas_sp_H184	0.244785	Healthy-1
Bacteroides_fragilis	0.23319	Healthy-1
Bacteroides_thetaiotaomicron	0.218279	Healthy-1

Species	EVC	Diagnosis Group
Bacteroides_vulgatus	0.209061	Healthy-1
Fournierella_massiliensis	0.382647	Healthy-2
Provencibacterium_massiliense	0.331834	Healthy-2
Subdoligranulum_variabile	0.327014	Healthy-2
Ruthenibacterium_lactatiformans	0.281926	Healthy-2
Faecalicatena_contorta	0.247965	Healthy-2
Clostridium_bolteae	0.245986	Healthy-2
Merdibacter_massiliensis	0.234116	Healthy-2
Clostridium_asparagiforme	0.211324	Healthy-2
Butyricicoccus_pullicaecorum	0.205136	Healthy-2
Blautia_hydrogenotrophica	0.166324	Healthy-2

References

1. Agnello, Luisa, Chiara Bellia, Lucio Lo Coco, Silvana Vitale, Felicia Coraci, Filippa Bonura, Rossella Gnoffo, et al. 2014. “Vitamin K Deficiency Bleeding Leading to the Diagnosis of Crohn’s Disease.” *Annals of Clinical and Laboratory Science* 44 (3): 337–40.
2. Aitchison, J. 1982. “The Statistical Analysis of Compositional Data.” *Journal of the Royal Statistical Society. Series B (Methodological)* 44 (2): 40. http://www.jstor.org/stable/2345821.
3. Benjamini, Yoav, and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1): 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
4. Bernalier-Donadille, A. 2010. “Fermentative Metabolism by the Human Gut Microbiota.” *Gastroentérologie Clinique et Biologique*, The intestinal microbiota: Equilibrium and disorders, 34 (October): S16–22. https://doi.org/10.1016/S0399-8320(10)70016-6.
5. Bhat, M., E. Pasini, J. Copeland, M. Angeli, S. Husain, D. Kumar, E. Renner, et al. 2017. “Impact of Immunosuppression on the Metagenomic Composition of the Intestinal Microbiome: A Systems Biology Approach to Post-Transplant Diabetes.” *Scientific Reports* 7 (1): 1–12. https://doi.org/10.1038/s41598-017-10471-2.
6. Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” *Bioinformatics* 30 (15): 2114–20. https://doi.org/10.1093/bioinformatics/btu170.

7. Bonacich, Phillip. 1972. “Factoring and Weighting Approaches to Status Scores and Clique Identification.” *The Journal of Mathematical Sociology*. https://doi.org/10.1080/0022250X.1972.9989806.

8. Breiman, Leo. 2001. “Random Forests.” *Machine Learning*. https://doi.org/10.1023/A:1010933404324.

9. Brown, Christopher T, Austin G Davis-Richardson, Adriana Giongo, Kelsey A Gano, David B Crabb, Nabanita Mukherjee, George Casella, et al. 2011. “Gut Microbiome Metagenomics Analysis Suggests a Functional Model for the Development of Autoimmunity for Type 1 Diabetes.” *PLOS ONE* 6 (10): e25792. https://doi.org/10.1371/journal.pone.0025792.

10. Chiara, Mentella Maria, Scaldaferri Franco, Pizzoferrato Marco, Gasbarrini Antonio, and Miggiano Giacinto Abele Donato. 2020. “Nutrition, Ibd and Gut Microbiota: A Review.” *Nutrients* 12 (4): 1–20. https://doi.org/10.3390/nu12040944.

11. Clooney, Adam G, Julia Eckenberger, Emilio Laserna-Mendieta, Kathryn A Sexton, Matthew T Bernstein, Kathy Vagianos, Michael Sargent, et al. 2020. “Ranking Microbiome Variance in Inflammatory Bowel Disease: A Large Longitudinal Intercontinental Study.” *Gut*, gutjnl-2020-321106. https://doi.org/10.1136/gutjnl-2020-321106.

12. Coyte, Katharine Z., Jonas Schluter, and Kevin R. Foster. 2015. “The Ecology of the Microbiome: Networks, Competition, and Stability.” *Science* 350 (6261): 663–66. https://doi.org/10.1126/science.aad2602.

13. Crost, Emmanuelle H., Louise E. Tailford, Marie Monestier, David Swarbreck, Bernard Henrissat, Lisa C. Crossman, Nathalie Juge, et al. 2016. “The Mucin-Degradation Strategy of Ruminococcus Gnavus: The Importance of Intramolecular Trans-Sialidases.” *Gut Microbes* 7 (4): 302–12. https://doi.org/10.1080/19490976.2016.1186334.

14. Desai, Mahesh S., Anna M. Seekatz, Nicole M. Koropatkin, Nobuhiko Kamada, Christina A. Hickey, Mathis Wolter, Nicholas A. Pudlo, et al. 2016. “A Dietary Fiber-Deprived Gut Microbiota Degrades the Colonic Mucus Barrier and Enhances Pathogen Susceptibility.” *Cell* 167 (5): 1339-1353.e21. https://doi.org/10.1016/j.cell.2016.10.043.

15. Dethlefsen, Les, Sue Huse, Mitchell L. Sogin, and David A. Relman. 2008. “The Pervasive Effects of an Antibiotic on the Human Gut Microbiota, as Revealed by Deep 16s RRNA Sequencing.” *PLoS Biology*. https://doi.org/10.1371/journal.pbio.0060280.

16. Díaz-Uriarte, Ramón, and Sara Alvarez de Andrés. 2006. “Gene Selection and Classification of Microarray Data Using Random Forest.” *BMC Bioinformatics* 7: 1–13. https://doi.org/10.1186/1471-2105-7-3.

17. Duranti, Sabrina, Federica Gaiani, Leonardo Mancabelli, Christian Milani, Andrea Grandi, Angelo Bolchi, Andrea Santoni, et al. 2016. “Elucidating the Gut Microbiome of

Ulcerative Colitis: Bifidobacteria as Novel Microbial Biomarkers." *FEMS Microbiology Ecology* 92 (12): fiw191. https://doi.org/10.1093/femsec/fiw191.

18. "FDA Briefing Document Gastrointestinal Drug Advisory Committee Meeting." 2018.
19. Fenn, Kathrin, Philip Strandwitz, Eric J. Stewart, Eric Dimise, Sarah Rubin, Shreya Gurubacharya, Jon Clardy, and Kim Lewis. 2017. "Quinones Are Growth Factors for the Human Gut Microbiota." *Microbiome* 5 (1): 161. https://doi.org/10.1186/s40168-017-0380-5.
20. Flores, Avegail, Ezra Burstein, Daisha J. Cipher, and Linda A. Feagins. 2015. "Obesity in Inflammatory Bowel Disease: A Marker of Less Severe Disease." *Digestive Diseases and Sciences* 60 (8): 2436–45. https://doi.org/10.1007/s10620-015-3629-5.
21. Fox, G. E., J. D. Wisotzkey, and P. Jurtshuk. 1992. "How Close Is Close: 16S RRNA Sequence Identity May Not Be Sufficient to Guarantee Species Identity." *International Journal of Systematic Bacteriology* 42 (1): 166–70. https://doi.org/10.1099/00207713-42-1-166.
22. Fox, George E, Linda J Magrum, William E Balcht, Ralph S Wolfef, and Carl R Woese. 1977. "Classification of Methanogenic Bacteria by 16S Ribosomal RNA Characterization (Comparative Oligonucleotide Cataloging/Phylogeny/Molecular Evolution)." *Evolution* 74 (10): 4537–41. https://doi.org/10.1073/pnas.74.10.4537.
23. Frank, Daniel N., Charles E. Robertson, Christina M. Hamm, Zegbeh Kpadeh, Tianyi Zhang, Hongyan Chen, Wei Zhu, et al. 2011. "Disease Phenotype and Genotype Are Associated with Shifts in Intestinal-Associated Microbiota in Inflammatory Bowel Diseases." *Inflammatory Bowel Diseases* 17 (1): 179–84. https://doi.org/10.1002/ibd.21339.
24. Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2008. "Sparse Inverse Covariance Estimation with the Graphical Lasso." *Biostatistics* 9 (3): 432–41. https://doi.org/10.1093/biostatistics/kxm045.
25. Friedman, Jonathan, and Eric J. Alm. 2012. "Inferring Correlation Networks from Genomic Survey Data." Edited by Christian von Mering. *PLoS Computational Biology* 8 (9): e1002687. https://doi.org/10.1371/journal.pcbi.1002687.
26. Gevers, Dirk, Subra Kugathasan, Lee A. Denson, Yoshiki Vázquez-Baeza, Will Van Treuren, Boyu Ren, Emma Schwager, et al. 2014. "The Treatment-Naive Microbiome in New-Onset Crohn's Disease." *Cell Host & Microbe* 15 (3): 382–92. https://doi.org/10.1016/j.chom.2014.02.005.
27. Gevers, Dirk, Subra Kugathasan, Lee A. Denson, Yoshiki Vázquez-Baeza, Will Van Treuren, Boyu Ren, Emma Schwager, et al. 2014. "The Treatment-Naive Microbiome in New-Onset Crohn's Disease." *Cell Host and Microbe* 15 (3): 382–92. https://doi.org/10.1016/j.chom.2014.02.005.

28. Gloor, Gregory B., Jean M. Macklaim, Vera Pawlowsky-Glahn, and Juan J. Egozcue. 2017. “Microbiome Datasets Are Compositional: And This Is Not Optional.” *Frontiers in Microbiology* 8 (NOV): 1–6. https://doi.org/10.3389/fmicb.2017.02224.

29. Gupta, Ankit, Darshan B. Dhakan, Abhijit Maji, Rituja Saxena, Vishnu Prasoodanan P.K., Shruti Mahajan, Joby Pulikkan, et al. 2019. “ Association of Flavonifractor Plautii , a Flavonoid-Degrading Bacterium, with the Gut Microbiome of Colorectal Cancer Patients in India .” *MSystems* 4 (6): 1–20. https://doi.org/10.1128/msystems.00438-19.

30. Haft, D. H. 2001. “TIGRFAMs: A Protein Family Resource for the Functional Identification of Proteins.” *Nucleic Acids Research* 29 (1): 41–43. https://doi.org/10.1093/nar/29.1.41.

31. Hagberg, Aric A, Daniel A Schult, and Pieter J Swart. 2008. “Exploring Network Structure, Dynamics, and Function Using NetworkX,” 6.

32. Halfvarson, Jonas, Colin J. Brislawn, Regina Lamendella, Yoshiki Vázquez-Baeza, William A. Walters, Lisa M. Bramer, Mauro D’Amato, et al. 2017. “Dynamics of the Human Gut Microbiome in Inflammatory Bowel Disease.” *Nature Microbiology* 2 (February): 1–7. https://doi.org/10.1038/nmicrobiol.2017.4.

33. Hall, Andrew Brantley, Moran Yassour, Jenny Sauk, Ashley Garner, Xiaofang Jiang, Timothy Arthur, Georgia K. Lagoudas, et al. 2017. “A Novel Ruminococcus Gnavus Clade Enriched in Inflammatory Bowel Disease Patients.” *Genome Medicine* 9 (1): 1–12. https://doi.org/10.1186/s13073-017-0490-5.

34. Heintz-Buschart, Anna, and Paul Wilmes. 2018. “Human Gut Microbiome: Function Matters.” *Trends in Microbiology* 26 (7): 563–74. https://doi.org/10.1016/j.tim.2017.11.002.

35. Hookman, Perry, and Jamie S. Barkin. 2009. “Clostridium Difficile Associated Infection, Diarrhea and Colitis.” *World Journal of Gastroenterology* 15 (13): 1554–80. https://doi.org/10.3748/wjg.15.1554.

36. Hunter, Sarah, Rolf Apweiler, Teresa K. Attwood, Amos Bairoch, Alex Bateman, David Binns, Peer Bork, et al. 2009. “InterPro: The Integrative Protein Signature Database.” *Nucleic Acids Research* 37 (Database): D211–15. https://doi.org/10.1093/nar/gkn785.

37. Huttenhower, Curtis, Dirk Gevers, Rob Knight, Sahar Abubucker, Jonathan H. Badger, Asif T. Chinwalla, Heather H. Creasy, et al. 2012. “Structure, Function and Diversity of the Healthy Human Microbiome.” *Nature* 486 (7402): 207–14. https://doi.org/10.1038/nature11234.

38. Hyatt, Doug, Gwo-Liang Chen, Philip F LoCascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. 2010. “Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification.” *BMC Bioinformatics* 11 (1): 119. https://doi.org/10.1186/1471-2105-11-119.

39. Ibal, Jerald Conrad, Huy Quang Pham, Chang Eon Park, and Jae Ho Shin. 2019. “Information about Variations in Multiple Copies of Bacterial 16S RRNA Genes May Aid in Species Identification.” *PLoS ONE* 14 (2): 1–15. https://doi.org/10.1371/journal.pone.0212090.

40. Johnson, Abigail J., Pajau Vangay, Gabriel A. Al-Ghalith, Benjamin M. Hillmann, Tonya L. Ward, Robin R. Shields-Cutler, Austin D. Kim, et al. 2019. “Daily Sampling Reveals Personalized Diet-Microbiome Associations in Humans.” *Cell Host & Microbe* 25 (6): 789-802.e5. https://doi.org/10.1016/j.chom.2019.05.005.

41. Kho, Zhi Y., and Sunil K. Lal. 2018. “The Human Gut Microbiome - A Potential Controller of Wellness and Disease.” *Frontiers in Microbiology* 9 (AUG): 1835. https://doi.org/10.3389/fmicb.2018.01835.

42. Kish, Lisa, Naomi Hotte, Gilaad G. Kaplan, Renaud Vincent, Robert Tso, Michael Gänzle, Kevin P. Rioux, et al. 2013. “Environmental Particulate Matter Induces Murine Intestinal Inflammatory Responses and Alters the Gut Microbiome.” *PLoS ONE* 8 (4). https://doi.org/10.1371/journal.pone.0062220.

43. Krasinski, S. D., R. M. Russell, B. C. Furie, S. F. Kruger, and P. F. Jacques. 1985. “The Prevalence of Vitamin K Deficiency in Chronic Gastrointestinal Disorders.” *American Journal of Clinical Nutrition*. https://doi.org/10.1093/ajcn/41.3.639.

44. Kukuruzovic, R., David R. Brewster, E. Gray, and N. M. Anstey. 2003. “Increased Nitric Oxide Production in Acute Diarrhoea Is Associated with Abnormal Gut Permeability, Hypokalaemia and Malnutrition in Tropical Australian Aboriginal Children.” *Transactions of the Royal Society of Tropical Medicine and Hygiene*. https://doi.org/10.1016/S0035-9203(03)90044-7.

45. Kurtz, Zachary D., Christian L. Müller, Emily R. Miraldi, Dan R. Littman, Martin J. Blaser, and Richard A. Bonneau. 2015. “Sparse and Compositionally Robust Inference of Microbial Ecological Networks.” Edited by Christian von Mering. *PLoS Computational Biology* 11 (5): 1–25. https://doi.org/10.1371/journal.pcbi.1004226.

46. Langmead, Ben, and Steven L Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59. https://doi.org/10.1038/nmeth.1923.

47. Laudadio, Ilaria, Valerio Fulci, Francesca Palone, Laura Stronati, Salvatore Cucchiara, and Claudia Carissimi. 2018. “Quantitative Assessment of Shotgun Metagenomics and 16S RDNA Amplicon Sequencing in the Study of Human Gut Microbiome.” *OMICS: A Journal of Integrative Biology* 22: 248–54. https://doi.org/10.1089/omi.2018.0013.

48. Lidder, Satnam, and Andrew J. Webb. 2013. “Vascular Effects of Dietary Nitrate (as Found in Green Leafy Vegetables and Beetroot) via the Nitrate-Nitrite-Nitric Oxide Pathway.” *British Journal of Clinical Pharmacology*. https://doi.org/10.1111/j.1365-2125.2012.04420.x.

49. Lloyd-Price, Jason, Cesar Arze, Ashwin N. Ananthakrishnan, Melanie Schirmer, Julian Avila-Pacheco, Tiffany W. Poon, Elizabeth Andrews, et al. 2019. “Multi-Omics of the Gut Microbial Ecosystem in Inflammatory Bowel Diseases.” *Nature* 569 (7758): 655–62. https://doi.org/10.1038/s41586-019-1237-9.

50. Loftus, Mark, Sayf Al-Deen Hassouneh, and Shibu Yooseph. 2021. “Bacterial Associations in the Healthy Human Gut Microbiome across Populations.” *Scientific Reports* 11 (1): 1–14. https://doi.org/10.1038/s41598-021-82449-0.

51. Loomba, Rohit, Victor Seguritan, Weizhong Li, Tao Long, Niels Klitgord, Archana Bhatt, Parambir Singh Dulai, et al. 2017. “Gut Microbiome-Based Metagenomic Signature for Non-Invasive Detection of Advanced Fibrosis in Human Nonalcoholic Fatty Liver Disease.” *Cell Metabolism* 25 (5): 1054-1062.e5. https://doi.org/10.1016/j.cmet.2017.04.001.

52. Mann, H. B., and D. R. Whitney. 1947. “On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other.” *The Annals of Mathematical Statistics* 18 (1): 50–60. https://doi.org/10.1214/aoms/1177730491.

53. Methé, Barbara A., Karen E. Nelson, Mihai Pop, Heather H. Creasy, Michelle G. Giglio, Curtis Huttenhower, Dirk Gevers, et al. 2012. “A Framework for Human Microbiome Research.” *Nature* 486 (7402): 215–21. https://doi.org/10.1038/nature11209.

54. Moreno-Vivián, Conrado, Purificación Cabello, Manuel Martínez-Luque, Rafael Blasco, and Francisco Castillo. 1999. “Prokaryotic Nitrate Reduction: Molecular Properties and Functional Distinction among Bacterial Nitrate Reductases.” *Journal of Bacteriology* 181 (21): 6573–84. https://doi.org/10.1128/jb.181.21.6573-6584.1999.

55. Moustafa, Ahmed, Weizhong Li, Ericka L. Anderson, Emily H.M. Wong, Parambir S. Dulai, William J. Sandborn, William Biggs, et al. 2018. “Genetic Risk, Dysbiosis, and Treatment Stratification Using Host Genome and Gut Microbiome in Inflammatory Bowel Disease.” *Clinical and Translational Gastroenterology* 9 (1): e132-8. https://doi.org/10.1038/ctg.2017.58.

56. Musumeci, Laura, Alessandro Maugeri, Santa Cirmi, Giovanni Enrico, Caterina Russo, Sebastiano Gangemi, Gioacchino Calapai, et al. 2020. “Citrus Fruits and Their Flavonoids in Inflammatory Bowel Disease: An Overview.” *Natural Product Research* 34 (1): 122–36. https://doi.org/10.1080/14786419.2019.1601196.

57. Nagata, Naoyoshi, Mari Tohya, Shinji Fukuda, Wataru Suda, Suguru Nishijima, Fumihiko Takeuchi, Mitsuru Ohsugi, et al. 2019. “Effects of Bowel Preparation on the Human Gut Microbiome and Metabolome.” *Scientific Reports* 9 (1): 1–8. https://doi.org/10.1038/s41598-019-40182-9.

58. Newman, M E J. 2006. “Modularity and Community Structure in Networks.” *Proceedings of the National Academy of Sciences* 103 (23): 8577–82. https://doi.org/10.1073/pnas.0601602103.

59. O’Leary, Nuala A, Mathew W Wright, J Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, et al. 2016. “Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation.” *Nucleic Acids Research* 44 (D1): D733–45. https://doi.org/10.1093/nar/gkv1189.

60. Ohland, Christina L., and Christian Jobin. 2015. “Microbial Activities and Intestinal Homeostasis: A Delicate Balance Between Health and Disease.” *CMGH*. https://doi.org/10.1016/j.jcmgh.2014.11.004.

61. Park, Ji Won, Barbora Piknova, Paul L. Huang, Constance T. Noguchi, and Alan N. Schechter. 2013. “Effect of Blood Nitrite and Nitrate Levels on Murine Platelet Function.” *PLoS ONE*. https://doi.org/10.1371/journal.pone.0055699.

62. Pasolli, Edoardo, Duy Tin Truong, Faizan Malik, Levi Waldron, and Nicola Segata. 2016. “Machine Learning Meta-Analysis of Large Metagenomic Datasets: Tools and Biological Insights.” *PLoS Computational Biology* 12 (7): 1–26. https://doi.org/10.1371/journal.pcbi.1004977.

63. Pearson, K. 1896. “Mathematical Contributions to the Theory of Evolution.” *Proeedings of the Royal Society* 60 (1834): 489–98. http://books.google.com/books?hl=en&lr=&id=aIU_AQAAIAAJ&oi=fnd&pg=PA1&dq=Mathematical+Contributions+to+the+Theory+of+Evolution&ots=6q0ynawAzT&sig=FdqqMWpdG0a5gRGfvPbW2BRUw8I.

64. Pérez-Gutiérrez, Rocío Anaís, Varinia López-Ramírez, África Islas, Luis David Alcaraz, Ismael Hernández-González, Beatriz Carely Luna Olivera, Moisés Santillán, et al. 2013. “Antagonism Influences Assembly of a Bacillus Guild in a Local Community and Is Depicted as a Food-Chain Network.” *ISME Journal* 7 (3): 487–97. https://doi.org/10.1038/ismej.2012.119.

65. Petrov, V. A., I. V. Saltykova, I. A. Zhukova, V. M. Alifirova, N. G. Zhukova, Yu B. Dorofeeva, A. V. Tyakht, et al. 2017. “Analysis of Gut Microbiota in Patients with Parkinson’s Disease.” *Bulletin of Experimental Biology and Medicine* 162 (6): 734–37. https://doi.org/10.1007/s10517-017-3700-7.

66. Ranjan, Ravi, Asha Rani, Ahmed Metwally, Halvor S. McGee, and David L. Perkins. 2016. “Analysis of the Microbiome: Advantages of Whole Genome Shotgun versus 16S Amplicon Sequencing.” *Biochemical and Biophysical Research Communications* 469 (4): 967–77. https://doi.org/10.1016/j.bbrc.2015.12.083.

67. Rastogi, Rajat, Martin Wu, Indrani Dasgupta, and George E. Fox. 2009. “Visualization of Ribosomal RNA Operon Copy Number Distribution.” *BMC Microbiology* 9: 208. https://doi.org/10.1186/1471-2180-9-208.

68. Roguet, Adélaïde, A. Murat Eren, Ryan J. Newton, and Sandra L. McLellan. 2018. “Fecal Source Identification Using Random Forest.” *Microbiome* 6 (1): 1–15. https://doi.org/10.1186/s40168-018-0568-3.

69. Rossum, Guido and Drake, Fred L. Van. 2009. *Python3 Reference Manual*. CreateSpace.

70. Ruhnau, Britta. 2000. “Eigenvector-Centrality - a Node-Centrality.” *Social Networks* 22 (4): 357–65. https://doi.org/10.1016/S0378-8733(00)00031-9.

71. Saulnier, Delphine M., Kevin Riehle, Toni Ann Mistretta, Maria Alejandra Diaz, Debasmita Mandal, Sabeen Raza, Erica M. Weidler, et al. 2011. “Gastrointestinal Microbiome Signatures of Pediatric Patients with Irritable Bowel Syndrome.” *Gastroenterology* 141 (5): 1782–91. https://doi.org/10.1053/j.gastro.2011.06.072.

72. Saunders, Aaron M, Mads Albertsen, Jes Vollertsen, and Per H Nielsen. 2016. “The Activated Sludge Ecosystem Contains a Core Community of Abundant Organisms.” *The ISME Journal* 10 (1): 11–20. https://doi.org/10.1038/ismej.2015.117.

73. Schirmer, Melanie, Lee Denson, Hera Vlamakis, Eric A. Franzosa, Sonia Thomas, Nathan M. Gotman, Paul Rufo, et al. 2018. “Compositional and Temporal Changes in the Gut Microbiome of Pediatric Ulcerative Colitis Patients Are Linked to Disease Course.” *Cell Host and Microbe* 24 (4): 600-610.e4. https://doi.org/10.1016/j.chom.2018.09.009.

74. Schoon, E. J., M. C.A. Müller, C. Vermeer, L. J. Schurgers, R. J.M. Brummer, and R. W. Stockbrügger. 2001. “Low Serum and Bone Vitamin K Status in Patients with Longstanding Crohn’s Disease: Another Pathogenetic Factor of Osteoporosis in Crohn’s Disease?” *Gut* 48 (4): 473–77. https://doi.org/10.1136/gut.48.4.473.

75. Schubert, Alyxandria M., Mary A.M. Rogers, Cathrin Ring, Jill Mogle, Joseph P. Petrosino, Vincent B. Young, David M. Aronoff, and Patrick D. Schloss. 2014. “Microbiome Data Distinguish Patients with Clostridium Difficile Infection and Non-c. Difficile-Associated Diarrhea from Healthy Controls.” *MBio* 5 (3): 1–9. https://doi.org/10.1128/mBio.01021-14.

76. Shannon, C. E. 1948. “A Mathematical Theory of Communication.” *Bell System Technical Journal*. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

77. Sheehan, Donal, Carthage Moran, and Fergus Shanahan. 2015. “The Microbiota in Inflammatory Bowel Disease.” *Journal of Gastroenterology* 50 (5): 495–507. https://doi.org/10.1007/s00535-015-1064-1.

78. Shen, Aimee, Adrianne N. Edwards, Mahfuzur R. Sarker, and Daniel Paredes-Sabja. 2019. “Sporulation and Germination in Clostridial Pathogens.” *Gram-Positive Pathogens* 7 (6): 903–26. https://doi.org/10.1128/9781683670131.ch56.

79. Shi, Tao, David Seligson, Arie S. Belldegrun, Aarno Palotie, and Steve Horvath. 2005. “Tumor Classification by Tissue Microarray Profiling: Random Forest Clustering Applied to Renal Cell Carcinoma.” *Modern Pathology* 18 (4): 547–57. https://doi.org/10.1038/modpathol.3800322.

80. Thomas, Andrew Maltez, Paolo Manghi, Francesco Asnicar, Edoardo Pasolli, Federica Armanini, Moreno Zolfo, Francesco Beghini, et al. 2019. “Metagenomic Analysis of Colorectal Cancer Datasets Identifies Cross-Cohort Microbial Diagnostic Signatures and a Link with Choline Degradation.” *Nature Medicine* 25 (4): 667–78. https://doi.org/10.1038/s41591-019-0405-7.

81. Tiso, Mauro, and Alan N. Schechter. 2015. “Nitrate Reduction to Nitrite, Nitric Oxide and Ammonia by Gut Bacteria under Physiological Conditions.” *PLoS ONE* 10 (3): 1–18. https://doi.org/10.1371/journal.pone.0119712.

82. Tsilimigras, Matthew C.B., and Anthony A. Fodor. 2016. “Compositional Data Analysis of the Microbiome: Fundamentals, Tools, and Challenges.” *Annals of Epidemiology* 26 (5): 330–35. https://doi.org/10.1016/j.annepidem.2016.03.002.

83. Ursell, Luke K., Jessica L. Metcalf, Laura Wegener Parfrey, and Rob Knight. 2012. “Defining the Human Microbiome.” *Nutrition Reviews* 70 (SUPPL. 1). https://doi.org/10.1111/j.1753-4887.2012.00493.x.

84. Varoquaux, G., L. Buitinck, G. Louppe, O. Grisel, F. Pedregosa, and A. Mueller. 2015. “Scikit-Learn.” *GetMobile: Mobile Computing and Communications* 19 (1): 29–33. https://doi.org/10.1145/2786984.2786995.

85. Veauthier, Brian, and Jaime R. Hornecker. 2018. “Crohn’s Disease: Diagnosis and Management.” *American Family Physician* 98 (11): 661–69.

86. Vich Vila, Arnau, Floris Imhann, Valerie Collij, Soesma A. Jankipersadsing, Thomas Gurry, Zlatan Mujagic, Alexander Kurilshikov, et al. 2018. “Gut Microbiota Composition and Functional Changes in Inflammatory Bowel Disease and Irritable Bowel Syndrome.” *Science Translational Medicine* 10 (472): eaap8914. https://doi.org/10.1126/scitranslmed.aap8914.

87. Walther, Barbara, J. Philip Karl, Sarah L. Booth, and Patrick Boyaval. 2013. “Menaquinones, Bacteria, and the Food Supply: The Relevance of Dairy and Fermented Food Products to Vitamin K Requirements.” *Advances in Nutrition* 4 (4): 463–73. https://doi.org/10.3945/an.113.003855.

88. Wermuth, Nanny, and Steffen Lilholt Lauritzen. 1990. “On Substantive Research Hypotheses, Conditional Independence Graphs and Graphical Chain Models.” *Journal of*

the Royal Statistical Society: Series B (Methodological) 52 (1): 21–50. https://doi.org/10.1111/j.2517-6161.1990.tb01771.x.

89. Wu, Shaoping, Jianxun Yi, Yong Guo Zhang, Jingsong Zhou, and Jun Sun. 2015. "Leaky Intestine and Impaired Microbiome in an Amyotrophic Lateral Sclerosis Mouse Model." *Physiological Reports* 3 (4): 1–10. https://doi.org/10.14814/phy2.12356.

90. Xia, Li C., Jacob A. Cram, Ting Chen, Jed A. Fuhrman, and Fengzhu Sun. 2011. "Accurate Genome Relative Abundance Estimation Based on Shotgun Metagenomic Reads." Edited by Emmanuel Dias-Neto. *PLoS ONE* 6 (12): e27992. https://doi.org/10.1371/journal.pone.0027992.

91. Yatsunenko, Tanya, Federico E Rey, Mark J Manary, Indi Trehan, Maria Gloria Dominguez-Bello, Monica Contreras, Magda Magris, et al. 2012. "Human Gut Microbiome Viewed across Age and Geography." *Nature* 486 (7402): 222–27. https://doi.org/10.1038/nature11053.

Availability of data and materials

- Healthy-1 Cohort
    - Human microbiome project: Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, et al. The NIH Human Microbiome Project. Genome Res. 2009;19(12):2317–23.
    - SRA: PRJNA48479 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA48479/)
- Healthy-2 Cohort
    - Johnson AJ, Vangay P, Al-Ghalith GA, Hillmann BM, Ward TL, Shields-Cutler RR, et al. Daily Sampling Reveals Personalized Diet-Microbiome Associations in Humans. Cell Host Microbe [Internet]. 2019 Sep 9;25(6):789-802.e5. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1931312819302501
    - ENA: PRJEB29065 (https://www.ebi.ac.uk/ena/browser/view/PRJEB29065)
- IBD Multi-omics Database cohort
    - Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, Andrews E, Ajami NJ, Bonham KS, Brislawn CJ, Casero D. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature. 2019 May;569(7758):655-62.

- SRA: PRJNA398089 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA398089/)

- All analysis and pertinent data files can be found at https://github.com/syooseph/YoosephLab/blob/master/MicrobiomeNetworks/IBD/

CHAPTER 4: CONCLUSION

To understand the effects caused by changes in the gut microbiome, analysis cannot be restricted to an examination of the presence, absence, or relative abundances of bacteria within the community. It is also important to examine the bacterial associations that compose the microbiome as well as the functional implications of the changes in the gut microbiome. To this end, we used WGS sequencing data to enable us to accurately identify changes in species-level taxonomic profiles and functional capacity of the microbiomes. We also used log-ratio transformations to account for the compositional nature of the sequencing data and developed a Graphical Gaussian Model to accurately infer bacterial associations. Finally, we included multiple independent cohorts to ensure that our findings could be replicated in independent cohorts thereby corroborating our findings.

Firstly, we studied four independent healthy cohorts from different geographic regions to examine the variation in the healthy gut microbiome and identify patterns within the human gut microbiome that may be associated with health. While we found significant variation in the differential abundances of the bacteria in the gut microbiome, approximately 95% of bacterial species were present in all cohorts, indicating a significant overlap in the presence of bacterial species across healthy subjects. Bacterial association networks of the healthy gut microbiomes also exhibited many similarities in their properties and demonstrated conserved structures. Specifically, 20 bacterial species were found to have the same 14 associations amongst each other across all healthy cohorts. All bacterial association networks also exhibited a preference

for species with similar taxonomy or function to associate positively with one another. Finally, analysis of functional capacity of all healthy cohorts demonstrated little variation, indicating that healthy populations have similar functional capabilities regardless of the differences in bacterial differential abundances. We demonstrate that gut microbiomes across healthy human populations from different geographical regions have similar species present, similar association network organization and properties, and similar functional capacities. These findings demonstrate that while possible differences within cohorts exist due to diet, genetics, and other geographically distinct influences, there is still a large amount of similarity within healthy gut microbiomes.

Next, we used WGS sequencing of IBD patients to identify important differences between American IBD gut microbiomes and American healthy subjects. To corroborate our findings and increase their generalizability, we utilized two independent healthy cohorts as well as the internal control group. We identified 34 bacterial species that were significantly elevated in IBD. While these species were elevated in IBD, they still appeared to play important roles, as measured by the number of associations they were involved in, in the bacterial association networks of the healthy gut microbiome. Additionally, analysis of functional capacity of the IBD gut microbiome revealed lower capacity for menaquinone synthesis, an essential vitamin (vitamin K) not produced by humans, which is involved the regulation of osteoporosis and rectal bleeding. The functional capacity analysis also revealed an increased capacity for nitrate reduction which can contribute to intestinal bleeding, as well as increased intestinal motility leading to diarrhea. It was also revealed that IBD gut microbiomes displayed elevated capacity for polysaccharide metabolism, possibly due to the increased relative

abundance of known mucin-degrading bacteria such as *Rumincoccus gnavus, Rumincoccus. torques,* and *Clostridium symbiosum.* These findings illustrate a link between the gut microbiome and IBD-related symptoms as well as provide potential targets for symptom management in IBD patients.

By using WGS sequencing in conjunction with compositionally appropriate analysis and network inference, we identified important species-level patterns in the relative abundances, community interactions, and functional capacity of the human gut microbiome in healthy subjects and IBD patients. Furthermore, the identified patterns were compared in the context of health and disease and some patterns were found to be associated with IBD-related symptoms such as rectal bleeding, diarrhea, inflammation, and mucin degradation. Finally, we corroborated our findings by using multiple healthy cohorts to ensure that our results are robust and are not limited to cohort-specific signals. Our findings illustrated that the gut microbiome is linked to IBD-related symptoms and identified specific pathways and bacterial species as potential targets in the management of IBD symptoms.

APPENDIX A: DIFFERENTIALLY ABUNDANDANT BACTERIAL SPECIES

Species	Differential Abundance	Search Results	Function	Keywords	Citations
Flavonifractor plautii	Elevated in IBD	Associated with IBD	Degrades beneficial flavinoids	Inflammatory	(Gupta et al. 2019; Musumeci et al. 2020)
Faecalicatena contorta	Elevated in IBD	Novel	Novel	Novel	Novel
Blautia hydrogenotrophica	Elevated in IBD	Associated with IBD	Multi-drug resistance	Antibiotic resistance	Novel
Clostridium bolteae	Elevated in IBD	Associated with IBD			(Rodriguez et al. 2020)
Clostridium citroniae	Elevated in IBD	Novel	Novel	Novel	Novel
Lachnoclostridium sp_YL32	Elevated in IBD	Associated with IBD			(Liang et al. 2019)
Oscillibacter sp_PEA192	Elevated in IBD	Novel	Novel	Novel	Novel
Clostridium symbiosum	Elevated in IBD	Associated with CRC and rectal bleeding	Transfers vanB genes to commensals	Antibiotic resistance, Mucin-related	(Xie et al. 2017; Launay et al. 2006)
Blautia obeum	Elevated in IBD	Elevated in Healthy			(Theriot and Petri 2020)
Ruminococcus gnavus	Elevated in IBD	Associated with IBD	Foments TNF-a from dendritic cells	Mucin-related, Inflammatory	(Hall et al. 2017; Beaud, Tailliez, and Aba-Mondoloni 2005; Henke et al. 2019; Chua et al. 2018; Imam et al. 2018; Hansen, Skov, and Justesen 2013; Crost et al. 2016)
Blautia hansenii	Elevated in IBD	Novel	Novel	Novel	Novel
Blautia sp_N6H115	Elevated in IBD	Novel	Novel	Novel	Novel
Merdimonas faecis	Elevated in IBD	Associated with obesity			(Schoch et al. 2012)
Dorea formicigenerans	Elevated in IBD	Associated with IBD			(Nomura et al. 2005)
Clostridium asparagiforme	Elevated in IBD	Elevated in Healthy			(Lett, Costello, and Roberts-Thomson 2020)
Intestinibacter bartlettii	Elevated in IBD	Novel	Novel	Novel	Novel
Ruminococcus torques	Elevated in IBD	Associated with CD	Associated with upper GI involvement		(Kwak et al. 2020)
Hungatella hathewayi	Elevated in IBD	Associated with CD			(Rodriguez et al. 2020)
Butyricicoccus pullicaecorum	Elevated in IBD	Elevated in Healthy			(Eeckhaut et al. 2013)
Agathobaculum desmolans	Elevated in IBD	Novel	Novel	Novel	Novel
Dorea longicatena	Elevated in IBD	Novel	Novel	Novel	Novel

Species	Differential Abundance	Search Results	Function	Keywords	Citations
Fournierella massiliensis	Elevated in IBD	Novel	Novel	Novel	Novel
Eubacterium hallii	Elevated in IBD	Novel	Novel	Novel	Novel
Fusicatenibacter saccharivorans	Elevated in IBD	Novel	Novel	Novel	Novel
Erysipelotrichaceae bacterium	Elevated in IBD	Novel	Novel	Novel	Novel
Holdemania massiliensis	Elevated in IBD	Novel	Novel	Novel	Novel
Lachnoclostridium phocaeense	Elevated in IBD	Novel	Novel	Novel	Novel
Negativibacillus massiliensis	Elevated in IBD	Novel	Novel	Novel	Novel
Clostridium scindens	Elevated in IBD	Novel	Modulates steroid signaling	Steroid signaling	(Morris, Winter, and Cato 1985)
Anaerostipes hadrus	Elevated in IBD	Associated with IBD	Exacerbates inflammation in mice with colitis	Inflammatory	(Zhang et al. 2016)
Subdoligranulum variabile	Elevated in IBD	Associated with IBS	Cytokine release in PI-IBS	Inflammatory	(Sundin et al. 2015)
Faecalibacterium prausnitzii	Elevated in IBD	Associated with Healthy	Anti-inflammatory		(Zhao et al. 2020; Zhou et al. 2018; Breyner et al. 2017; Burkqvist et al. 2019)
Mordavella sp_MarseilleP3756	Elevated in IBD	Novel	Novel	Novel	Novel

APPENDIX B: CONSENTS FOR PUBLICATION

We, the authors, give our permission to include data and materials described in Loftus et. Al 2021 (below) in the dissertation contents of Mr. Sayf Al-Deen Hassouneh for Doctor of Philosophy in Biomedical Sciences at the University of Central Florida.
Article Title: Bacterial associations in the healthy human gut microbiome across populations
Authors: Mark Loftus*, Sayf Al-Deen Hassouneh*, Shibu Yooseph
Journal: Scientific Reports
DOI: https://doi.org/10.1038/s41598-021-82449-0
Published: 02 February 2021

We, the authors, give our permission to include data and materials described in Hassouneh et. Al 2021 (below) in the dissertation contents of Mr. Sayf Al-Deen Hassouneh for Doctor of Philosophy in Biomedical Sciences at the University of Central Florida.
Article Title: Linking Inflammatory Bowel Disease Symptoms to Changes in the Gut Microbiome Structure and Function
Authors: Sayf Al-Deen Hassouneh, Mark Loftus, Shibu Yooseph
Journal: Frontiers in Microbiology
DOI: 10.3389/fmicb.2021.673632
Published: Pending publication

* Indicates that authors contributed equally.