Metabolic phenotypes in *Clostridioides difficile* and the Gut Microbiome

Deborah Adair Powers
Charlottesville, VA

M.S., University of Virginia, 2021
B.S., University of Virginia, 2019

A Dissertation presented to the Graduate Faculty of the
University of Virginia in Candidacy for the
Degree of Doctor of Philosophy

Department of Biochemistry and Molecular Genetics

University of Virginia
December 2024

This dissertation is dedicated to Jeff:
I wish you were still here;
I am grateful you were.

“The beautiful changes as a forest is changed
By a chameleon’s tuning his skin to it;
As a mantis, arranged
On a green leaf, grows
Into it, makes the leaf leafier, and proves
Any greenness is deeper than anyone knows.”
– Richard Wilbur, *The Beautiful Changes*

Table of Contents

Abstract

Clostridioides difficile infection (CDI) is a major healthcare-acquired infection caused by dysbiosis of the gut microbiome, typically due to broad-spectrum antibiotics. *C. difficile* pathogenesis is mediated by its two toxins, TcdA and TcdB which damage the gut epithelial barrier. Toxin production in *C. difficile* is metabolically sensitive, responding to a variety of external metabolites. However, the intracellular metabolic phenotypes of *C. difficile* connecting these external variables to a specific toxin state are unclear. In this dissertation, I focus on elucidating the metabolic phenotypes associated with low and high toxin states in *C. difficile*. By using publicly available RNAseq data from *C. difficile* grown in 16 unique conditions in conjunction with genome-scale metabolic models, I discovered metabolic differences between toxin states and strains of *C. difficile*. Furthermore, I found reactions that transformed *C. difficile* from a metabolic state associated with a high toxin state to a low toxin state in the model.

To further explore metabolic phenotypes in *C. difficile*, I next focused on metabolic interactions between *C. difficile* and the environment, using *C. difficile* strain type, diet, and gut commensals as the variables. By creating 105 strain-specific metabolic models, I was able to identify metabolic phenotypes that varied by both diet and strain. Notably, I found a subset of *C. difficile* strains that have increased flux through energy-generating redox pathways. I also found that simulations of *C. difficile*-commensal interactions were not affected by the *C. difficile* strain type, however they were affected by diet. This suggests a hierarchical relationship between diet, commensals, and *C. difficile* genetics.

Together, this dissertation provides a wholistic understanding of metabolic phenotypes in *C. difficile* related to toxin production, diet, genetics, and commensal interactions. I used disparate datasets and variables to investigate how they affect metabolism both in isolation and all together. These unique analyses have improved our understanding of the metabolic and ecological processes that are important to CDI pathogenesis, and they are a critical stepping stone towards improving treatment and management of disease.

Chapter 1: Introduction

The gut microbiome: composition, function, and ecology

The gut microbiome (GM) is a dynamic ecosystem, composed of approximately 500 unique species and over three million genes[1,2]. The taxonomic composition of the GM is ever shifting in response to numerous variables such as geographical location, age, diet, exercise, or genetics[2]. As a result, the specific microbiota composition of the gut varies significantly between individuals, such that half of the bacterial species identified are unique to an individual[3]. Despite variability in species-level composition, trends in composition emerge at the phylum-level: phyla such as Bacteroidetes, Firmicutes, and Actinobacteria predominate the average GM, accounting for over 90% of the GM[2]. The taxonomic diversity of the GM can be measured with two metrics: alpha and beta diversity. Alpha diversity measures the number and relative abundance of species within a single GM sample[4]. Beta diversity measures how similar or dissimilar GM communities are across different GM samples. Taxonomic diversity results in metabolic redundancy. Consequently, in contrast to the compositional variability of the GM, the metabolic functions of the intestine are much more conserved, with 57% of metabolic functions present in 75% or more of individuals[3].

The GM performs metabolic biotransformations crucial for human health. For example, bile acids (BAs), which digest fats and aid in the absorption of fat-soluble vitamins, depend on the GM for recycling; in BA metabolism, BAs are deconjugated via bile salt hydrolases (BSH) which are prevalent in Bacteroidetes, Firmicutes, and Actinobacteria phyla[5–7]. Deconjugated BAs can then be reabsorbed by the liver; approximately 95% of BAs are recycled in this efficient metabolic loop[5,7]. In cases of GM imbalance, decreased taxa with BSH functionality increases the primary BA pool leading to increased toxicity[5]. In addition to BA metabolism, the GM also ferments carbohydrates to short-chain fatty acids (SCFAs)[8,9]. SCFAs are essential to maintaining the intestinal epithelial barrier and are a primary energy source for enterocytes[10]. SCFAs are anti-inflammatory molecules and regulate intestinal macrophages and helper T cells[11]. Furthermore, the GM metabolizes neurotransmitters such as serotonin and dopamine; all three molecules can influence the production of stress hormones such as cortisol which impacts mood, appetite, and metabolism[12,13]. These microbiota-produced metabolites have been implicated in neurological disorders, such as dementia and depression[13–17]. Together, these metabolic processes comprise just a few of the numerous ways in which the GM buffers human health.

Common ecological principles govern GM behavior, determining the growth and carrying capacity of the system and the interspecific interactions of competition and symbiosis[18,19]. Bacterial population growth is regulated through quorum sensing molecules; these molecules either limit or stimulate growth depending on environmental cues such as resource depletion, biofilm formation, or virulence factor production[20]. Nutrient niches shape the interspecific interactions of the GM. Dietary and mucosal polysaccharides and amino acids plus their innumerable byproducts are consumed, competed for, and cross-fed by members of the GM[18,21]. Once a bacterial species has claimed a nutritional niche it holds a competitive advantage over new species. In a balanced ecosystem, the diversity, abundance, and nutritional resources of the GM contribute to colonization resistance (CR) against the outgrowth of a singular taxa or pathogen, such as the nosocomial pathogen *Clostridioides difficile*[22–24].

Clostridioides difficile biology and pathogenesis

Clostridioides difficile is a gram-positive, anaerobic bacteria in the Firmicutes phyla. Its genome is approximately 4.1-4.3 Mb and contains a high proportion of mobile genetic elements[25–27]. *C. difficile* has an ultra-low level of genome conservation; the core genome may be as low as 16% conserved[27]. Furthermore, there are large phylogenetic distances between strains, making species classification difficult[25]. Metabolically, Stickland fermentation is highly conserved between strains[27–29]. Stickland fermentation is a system of coupled oxidative and reductive reactions that transfer electrons between amino acids, resulting in the generation of ATP and other energy-rich molecules[29]. Stickland fermentation provides a significant competitive advantage to *C. difficile* as amino acids are an abundant resource in host mucins of the large intestine[29,30].

C. difficile colonizes the large intestine and induces host epithelial cell death and intestinal inflammation via its pathogenic exotoxins TcdA and TcdB and, in some clinical strains, binary toxin[31]. Once released, these toxins disrupt the colonic epithelial barrier[32]. TcdA and TcdB induce a portion of this damage by glucosylating, and

thereby inactivating, Rho GTPases in the host epithelial cells[33]. Rho GTPase inactivation disrupts the actin cytoskeleton and tight junctions of epithelial cells, resulting in a cytopathic phenotype of altered cell morphology and in an impaired epithelial barrier function[34]. A compromised epithelial barrier can result in progression of CDI to pseudomembranous colitis, toxic megacolon, and, in 9.3% of cases, death[35,36].

C. difficile is the leading contributor to healthcare acquired gastrointestinal (GI) infections, accounting for 73% of GI infections and costing an estimated $6.3 billion annually in the United States[35,37]. The primary risk factor for developing *C. difficile* infection (CDI) is broad-spectrum antibiotic use which clears the GM leaving an open niche for *C. difficile* to colonize (Fig 1.1)[22]. The standard treatment for CDI is broad-spectrum antibiotics, such as vancomycin or metronidazole[38]. These antibiotics are effective in resolving CDI initially; however, the antibiotics used to treat *C. difficile* act against the vegetative state but not the endospores and cause further GM disruption, priming the gut for recurrent CDI[24]. In approximately 30% of patients the loss of a healthy microbiome leads to recurrent CDI, which is defined as an episode of CDI occurring within 8 weeks from the initial treatment[39]. Patients with recurrent CDI are treated with a fecal microbiota transplant (FMT) which is highly effective in resolving disease in 90% of cases[40,41]. FMT is an effective treatment because it restores a balanced state of the GM and leverages the innate immune defense of CR[19,22,40]. Thus, CDI onset and resolution can both be explained through ecological principles (Fig 1.1).

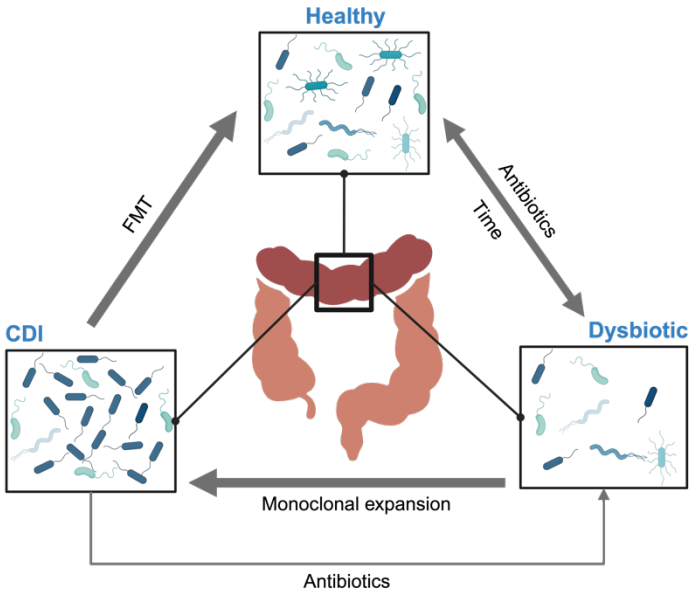


Figure 1.1 *C. difficile* pathogenesis. Antibiotics disrupt the normal gut flora, leading to a dysbiotic state with decreased abundance and diversity. This creates an open niche for monoclonal expansion of opportunistic pathogens, such as *C. difficile.* CDI can be treated with antibiotics or FMT. Treatment with antibiotics results in either a recurring cycle of infection or eventual recovery of a healthy microbiome. Figure created with BioRender.

Dementia and the gut microbiome

Dementia affects approximately 6.5 million Americans over the age of 65, costing an estimated 321 billion dollars annually[42,43]. It is defined as a syndrome that presents with deterioration of memory, language, thinking, and behavior[43,44]. Dementias are further categorized by the physiological cause of disease; Alzheimer's disease (AD) is the most common cause, comprising 60-80% of dementias[43]. In AD, pathogenesis occurs through build-up of beta-amyloid plaques around neurons and neurofibrillary tangles of protein tau inside neurons[43,45]. This neurodegeneration is progressive, manifesting a continuum of symptoms from normal cognition to mild cognitive impairment to dementia, typically over a time span of 10-20 years[43,45].

The GM plays a complicated role in dementia. The GM can metabolize neurotransmitters and modulators such as short-chain fatty acids (SCFA), gamma-amino butyric acid (GABA), acetylcholine, dopamine, glutamate, and serotonin[44,45]. These neurotransmitters act through the gut-brain axis, a communication pathway between the enteric and central nervous systems (ENS, CNS)[13,44–46]. Furthermore, intestinal dysbiosis and inflammation are consistent symptoms across neurodegenerative disorders which may be due to general diffusion of neuroinflammation of the CNS via the gut-brain axis[13,45]. Mapping the associations between the GM composition and function and neurocognition has been a recent focus in dementia research; several bacterial families are differentially abundant in AD versus controls, though findings currently lack consistency due to a variety of factors[47–59].

Systems biology approaches

Systems biology is a framework for understanding how biological systems function as a whole; it examines how the individual components of a system interact and contribute to a behavior of interest in a system. A system can be defined at many scales, such as a cell, an organ, a disease, or an ecosystem. Systematic approaches frequently leverage high-throughput datasets and computational models to generate hypotheses and isolate the

variables influencing the system. Applying this framework to complex systems like the GM, CDI, or dementia allows us to study causal relationships between variables and outcomes.

The primary systems biology tool used in this dissertation is genome-scale metabolic modeling. Metabolic models are mathematical descriptions of the metabolic network of an organism based on its genome sequence. A genome-scale metabolic network reconstruction (GENRE) consists of two main components: 1) a stoichiometric matrix of all the reactions and metabolites within a network and 2) Boolean gene-protein-reaction rules which define a gene's relationship to a protein and the reaction catalyzed by that protein (Fig 1.2A-B)[60,61]. Flux balance analysis (FBA) simulates metabolic flux through a GENRE under the constraints of stoichiometric mass-balance and optimization of an objective function, for example maximizing biomass production[60,62]. FBA results predict the metabolic phenotype of an organism under the given constraints. GENREs can be manipulated at three biologically distinct levels (gene, reaction, and metabolite) and numerous biologically relevant objective functions can be set[61,63]. Together, these variables enable myriad permutations and explorations of possible metabolic phenotypes

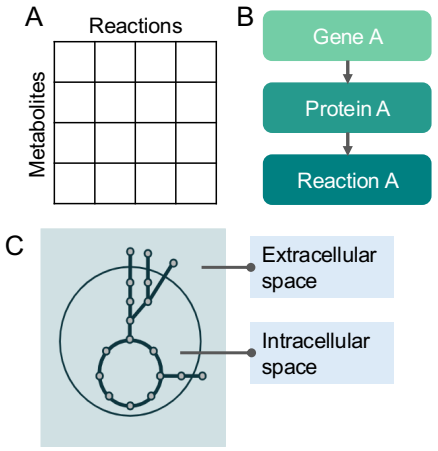


Figure 1.2 Components of a GENRE. (A) An S-matrix consisting of all the metabolic reactions and metabolites in a network. **(B)** An example of GPR rules. **(C)** A visual representation of the metabolic network that consists of intra- and extracellular spaces. Metabolic flux can be simulated in using network.

Specific Aims

Aim 1: Define the metabolic phenotypes in *C. difficile* related to toxin production.

1.1 Integrate publicly available RNA-seq data from *C. difficile* grown in different environments with *C. difficile* GENREs to create context-specific models.

1.2 Determine the metabolic phenotypes associated with low and high toxin states using the context specific models.

1.3 Identify metabolic switches that can shift a metabolic model from a high toxin state to a low toxin state.

Aim 2: Categorize the effects of strain, diet, and microbiome interactions on *C. difficile* metabolic phenotypes.

2.1 Reconstruct a panel of metabolic models for 100 *C. difficile* strains using whole-genome sequences.

2.2 Categorize the metabolic phenotypes of *C. difficile* as a function of strain, diet, and microbiome interactions using the 100 strain models.

2.3 Validate microbiome interaction predictions using *in vitro* co-culture methods.

References

1. Hou, K. *et al.* Microbiota in health and diseases. *Sig Transduct Target Ther* **7**, 135 (2022).
2. Rinninella, E. *et al.* What is the Healthy Gut Microbiota Composition? A Changing Ecosystem across Age, Environment, Diet, and Diseases. *Microorganisms* **7**, 14 (2019).
3. Visconti, A. *et al.* Interplay between the human gut microbiome and host metabolism. *Nat Commun* **10**, 4505 (2019).
4. Wensel, C. R., Pluznick, J. L., Salzberg, S. L. & Sears, C. L. Next-generation sequencing: insights to advance clinical investigations of the microbiome. *J Clin Invest* **132**, e154944 (2022).
5. Collins, S. L., Stine, J. G., Bisanz, J. E., Okafor, C. D. & Patterson, A. D. Bile acids and the gut microbiota: metabolic interactions and impacts on disease. *Nat Rev Microbiol* **21**, 236–247 (2023).
6. Winston, J. A. & Theriot, C. M. Diversification of host bile acids by members of the gut microbiota. *Gut Microbes* **11**, 158–171 (2020).
7. Ticho, A. L., Malhotra, P., Dudeja, P. K., Gill, R. K. & Alrefai, W. A. Intestinal Absorption of Bile Acids in Health and Disease. *Compr Physiol* **10**, 21–56 (2019).
8. Deehan, E. C. *et al.* Modulation of the Gastrointestinal Microbiome with Nondigestible Fermentable Carbohydrates To Improve Human Health. *Microbiol Spectr* **5**, 5.5.04 (2017).
9. Makki, K., Deehan, E. C., Walter, J. & Bäckhed, F. The Impact of Dietary Fiber on Gut Microbiota in Host Health and Disease. *Cell Host & Microbe* **23**, 705–715 (2018).

10. Morrison, D. J. & Preston, T. Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism. *Gut Microbes* **7**, 189–200 (2016).
11. Zheng, D., Liwinski, T. & Elinav, E. Interaction between microbiota and immunity in health and disease. *Cell Res* **30**, 492–506 (2020).
12. Gershon, M. D. & Margolis, K. G. The gut, its microbiome, and the brain: connections and communications. *Journal of Clinical Investigation* **131**, e143768 (2021).
13. Cryan, J. F., O'Riordan, K. J., Sandhu, K., Peterson, V. & Dinan, T. G. The gut microbiome in neurological disorders. *Lancet Neurol* **19**, 179–194 (2020).
14. Valles-Colomer, M. *et al.* The neuroactive potential of the human gut microbiota in quality of life and depression. *Nat Microbiol* **4**, 623–632 (2019).
15. Lin, L., Zheng, L. J. & Zhang, L. J. Neuroinflammation, Gut Microbiome, and Alzheimer's Disease. *Mol Neurobiol* **55**, 8243–8250 (2018).
16. Liu, L. *et al.* Gut microbiota and its metabolites in depression: from pathogenesis to treatment. *eBioMedicine* **90**, 104527 (2023).
17. Hung, C.-C., Chang, C.-C., Huang, C.-W., Nouchi, R. & Cheng, C.-H. Gut microbiota in patients with Alzheimer's disease spectrum: a systematic review and meta-analysis. *Aging (Albany NY)* **14**, 477–496 (2022).
18. Culp, E. J. & Goodman, A. L. Cross-feeding in the gut microbiome: Ecology and mechanisms. *Cell Host Microbe* **31**, 485–499 (2023).
19. Hitch, T. C. A. *et al.* Microbiome-based interventions to modulate gut ecology and the immune system. *Mucosal Immunology* **15**, 1095–1113 (2022).
20. Tudela, H., Claus, S. P. & Saleh, M. Next Generation Microbiome Research: Identification of Keystone Species in the Metabolic Regulation of Host-Gut Microbiota Interplay. *Front Cell Dev Biol* **9**, 719072 (2021).
21. Sanchez, A. *et al.* The community-function landscape of microbial consortia. *Cell Systems* **14**, 122–134 (2023).
22. Kamiya, S. Microbial ecology between Clostridioides difficile and gut microbiota. *Biosci Microbiota Food Health* **42**, 229–235 (2023).
23. Furuichi, M. *et al.* Commensal consortia decolonize Enterobacteriaceae via ecological control. *Nature* (2024) doi:10.1038/s41586-024-07960-6.
24. Theriot, C. M. & Young, V. B. Interactions Between the Gastrointestinal Microbiome and Clostridium difficile. *Annu Rev Microbiol* **69**, 445–461 (2015).
25. Knight, D. R. *et al.* Major genetic discontinuity and novel toxigenic species in Clostridioides difficile taxonomy. *Elife* **10**, e64325 (2021).
26. Knight, D. R., Elliott, B., Chang, B. J., Perkins, T. T. & Riley, T. V. Diversity and Evolution in the Genome of Clostridium difficile. *Clin Microbiol Rev* **28**, 721–741 (2015).
27. Scaria, J. *et al.* Analysis of Ultra Low Genome Conservation in Clostridium difficile. *PLoS ONE* **5**, e15147 (2010).
28. Kulecka, M. *et al.* Higher genome variability within metabolism genes associates with recurrent Clostridium difficile infection. *BMC Microbiol* **21**, 36 (2021).
29. Pavao, A. *et al.* Reconsidering the in vivo functions of Clostridial Stickland amino acid fermentations. *Anaerobe* **76**, 102600 (2022).
30. Fletcher, J. R. *et al.* Clostridioides difficile exploits toxin-mediated inflammation to alter the host nutritional landscape and exclude competitors from the gut microbiota. *Nat Commun* **12**, 462 (2021).
31. Chandrasekaran, R. & Lacy, D. B. The role of toxins in Clostridium difficile infection. *FEMS Microbiol Rev* **41**, 723–750 (2017).
32. Voth, D. E. & Ballard, J. D. Clostridium difficile toxins: mechanism of action and role in disease. *Clin Microbiol Rev* **18**, 247–263 (2005).
33. Vedantam, G. *et al.* Clostridium difficile infection: toxins and non-toxin virulence factors, and their contributions to disease establishment and host response. *Gut Microbes* **3**, 121–134 (2012).
34. Aktories, K., Schwan, C. & Jank, T. Clostridium difficile Toxin Biology. *Annu Rev Microbiol* **71**, 281–307 (2017).
35. Lessa, F. C., Winston, L. G., McDonald, L. C., & Emerging Infections Program C. difficile Surveillance Team. Burden of Clostridium difficile infection in the United States. *N Engl J Med* **372**, 2369–2370 (2015).
36. Olsen, M. A., Yan, Y., Reske, K. A., Zilberberg, M. D. & Dubberke, E. R. Recurrent Clostridium difficile infection is associated with increased mortality. *Clin Microbiol Infect* **21**, 164–170 (2015).

37. Zhang, S. *et al.* Cost of hospital management of Clostridium difficile infection in United States—a meta-analysis and modelling study. *BMC Infect Dis* **16**, 447 (2016).
38. Al-Jashaami, L. S. & DuPont, H. L. Management of Clostridium difficile Infection. *Gastroenterol Hepatol (N Y)* **12**, 609–616 (2016).
39. Song, J. H. & Kim, Y. S. Recurrent Clostridium difficile Infection: Risk Factors, Treatment, and Prevention. *Gut Liver* **13**, 16–24 (2019).
40. Baktash, A. *et al.* Mechanistic Insights in the Success of Fecal Microbiota Transplants for the Treatment of Clostridium difficile Infections. *Front Microbiol* **9**, 1242 (2018).
41. Baunwall, S. M. D. *et al.* Faecal microbiota transplantation for first or second Clostridioides difficile infection (EarlyFMT): a randomised, double-blind, placebo-controlled trial. *The Lancet Gastroenterology & Hepatology* **7**, 1083–1091 (2022).
42. Rajan, K. B. *et al.* Population estimate of people with clinical Alzheimer's disease and mild cognitive impairment in the United States (2020-2060). *Alzheimers Dement* **17**, 1966–1975 (2021).
43. 2022 Alzheimer's disease facts and figures. *Alzheimers Dement* **18**, 700–789 (2022).
44. Shabbir, U., Arshad, M. S., Sameen, A. & Oh, D.-H. Crosstalk between Gut and Brain in Alzheimer's Disease: The Role of Gut Microbiota Modulation Strategies. *Nutrients* **13**, 690 (2021).
45. Sochocka, M. *et al.* The Gut Microbiome Alterations and Inflammation-Driven Pathogenesis of Alzheimer's Disease-a Critical Review. *Mol Neurobiol* **56**, 1841–1851 (2019).
46. Liu, S., Gao, J., Zhu, M., Liu, K. & Zhang, H.-L. Gut Microbiota and Dysbiosis in Alzheimer's Disease: Implications for Pathogenesis and Treatment. *Mol Neurobiol* **57**, 5026–5043 (2020).
47. Haran, J. P. *et al.* Alzheimer's Disease Microbiome Is Associated with Dysregulation of the Anti-Inflammatory P-Glycoprotein Pathway. *mBio* **10**, e00632-19 (2019).
48. Hou, M., Xu, G., Ran, M., Luo, W. & Wang, H. APOE-ε4 Carrier Status and Gut Microbiota Dysbiosis in Patients With Alzheimer Disease. *Front Neurosci* **15**, 619051 (2021).
49. Laske, C. *et al.* Signature of Alzheimer's Disease in Intestinal Microbiome: Results From the AlzBiom Study. *Front. Neurosci.* **16**, 792996 (2022).
50. Li, B. *et al.* Mild cognitive impairment has similar alterations as Alzheimer's disease in gut microbiota. *Alzheimers Dement* **15**, 1357–1366 (2019).
51. Ling, Z. *et al.* Structural and Functional Dysbiosis of Fecal Microbiota in Chinese Patients With Alzheimer's Disease. *Front Cell Dev Biol* **8**, 634069 (2020).
52. Liu, P. *et al.* Altered microbiomes distinguish Alzheimer's disease from amnestic mild cognitive impairment and health in a Chinese cohort. *Brain Behav Immun* **80**, 633–643 (2019).
53. Nagpal, R., Neth, B. J., Wang, S., Craft, S. & Yadav, H. Modified Mediterranean-ketogenic diet modulates gut microbiome and short-chain fatty acids in association with Alzheimer's disease markers in subjects with mild cognitive impairment. *EBioMedicine* **47**, 529–542 (2019).
54. Saji, N. *et al.* Relationship between dementia and gut microbiome-associated metabolites: a cross-sectional study in Japan. *Sci Rep* **10**, 8088 (2020).
55. Sheng, C. *et al.* Altered Gut Microbiota in Adults with Subjective Cognitive Decline: The SILCODE Study. *J Alzheimers Dis* **82**, 513–526 (2021).
56. Vogt, N. M. *et al.* Gut microbiome alterations in Alzheimer's disease. *Sci Rep* **7**, 13537 (2017).
57. Xi, J. *et al.* Disturbed microbial ecology in Alzheimer's disease: evidence from the gut microbiota and fecal metabolome. *BMC Microbiol* **21**, 226 (2021).
58. Yıldırım, S. *et al.* Stratification of the Gut Microbiota Composition Landscape across the Alzheimer's Disease Continuum in a Turkish Cohort. *mSystems* **7**, e0000422 (2022).
59. Zhuang, Z.-Q. *et al.* Gut Microbiota is Altered in Patients with Alzheimer's Disease. *J Alzheimers Dis* **63**, 1337–1346 (2018).
60. Rawls, K. D. *et al.* A simplified metabolic network reconstruction to promote understanding and development of flux balance analysis tools. *Computers in Biology and Medicine* **105**, 64–71 (2019).
61. Haggart, C. R., Bartell, J. A., Saucerman, J. J. & Papin, J. A. Whole-Genome Metabolic Network Reconstruction and Constraint-Based Modeling⋆. in *Methods in Enzymology* vol. 500 411–433 (Elsevier, 2011).
62. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nat Biotechnol* **28**, 245–248 (2010).
63. Schuetz, R., Kuepfer, L. & Sauer, U. Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Molecular Systems Biology* **3**, 119 (2007).

Chapter 2: Network analysis of toxin production in *Clostridioides difficile* identifies key metabolic dependencies*

Authors
Deborah A. Powers[1], Matthew L. Jenior[2], Glynis L. Kolling[2], Jason A. Papin[1,2]

1. Biochemistry and Molecular Genetics, School of Medicine, University of Virginia, Charlottesville, Virginia, United States of America
2. Biomedical Engineering, School of Engineering, University of Virginia, Charlottesville, Virginia, United States of America

Abstract
Clostridioides difficile pathogenesis is mediated through its two toxin proteins, TcdA and TcdB, which induce intestinal epithelial cell death and inflammation. It is possible to alter *C. difficile* toxin production by changing various metabolite concentrations within the extracellular environment. However, it is unknown which intracellular metabolic pathways are involved and how they regulate toxin production. To investigate the response of intracellular metabolic pathways to diverse nutritional environments and toxin production states, we use previously published genome-scale metabolic models of *C. difficile* strains CD630 and CDR20291 (iCdG709 and iCdR703). We integrated publicly available transcriptomic data with the models using the RIPTiDe algorithm to create 16 unique contextualized *C. difficile* models representing a range of nutritional environments and toxin states. We used Random Forest with flux sampling and shadow pricing analyses to identify metabolic patterns correlated with toxin states and environment. Specifically, we found that arginine and ornithine uptake is particularly active in low toxin states. Additionally, uptake of arginine and ornithine is highly dependent on intracellular fatty acid and large polymer metabolite pools. We also applied the metabolic transformation algorithm (MTA) to identify model perturbations that shift metabolism from a high toxin state to a low toxin state. This analysis expands our understanding of toxin production in *C. difficile* and identifies metabolic dependencies that could be leveraged to mitigate disease severity.

* The material presented in this chapter has been previously published here:
Powers DA, Jenior ML, Kolling GL, Papin JA (2023) Network analysis of toxin production in Clostridioides difficile identifies key metabolic dependencies. PLoS Comput Biol 19(4): e1011076. https://doi.org/10.1371/journal.pcbi.1011076

Introduction

Clostridioides difficile is the leading contributor to healthcare-acquired gastrointestinal (GI) infections, accounting for 73% of GI infections and costing an estimated $6.3 billion annually in the United States[1,2]. The primary risk factor for developing *C. difficile* infection (CDI) is broad-spectrum antibiotic usage which alters the structure and composition of the gut microbiota, allowing *C. difficile* to outgrow nonpathogenic competitors[3]. CDI is recurrent in 30% of cases and has a mortality rate of 9.3%, causing 29,000 deaths in the US annually[1,4].

CDI pathogenesis is primarily mediated by two large *Clostridia* toxins, TcdA and TcdB[5,6]. These toxins induce cytopathic and cytotoxic effects, including changes in epithelial cell morphology, cell cycle modulation, disruption of the colonic epithelial barrier, induction of apoptosis, and activation of an acute innate inflammatory response[6–10]. TcdA and TcdB induce a portion of this damage by glucosylating, and thereby inactivating, Rho GTPases in the host intestinal epithelial cells[7]. Rho GTPase inactivation disrupts the actin cytoskeleton and tight junctions of epithelial cells, resulting in a cytopathic phenotype of altered cell morphology and impaired epithelial barrier function[11]. A compromised colonic epithelial barrier can lead to increased intestinal permeability and fluid secretion which furthers intestinal inflammation and damage[5]. Together, these toxins are crucial to the establishment of CDI.

Regulation of *C. difficile* toxin synthesis is complex. TcdA and TcdB are encoded by genes on the pathogenicity locus (PaLoc) along with a transcriptional regulator, TcdR[12]. TcdR, in turn, is negatively controlled by metabolically sensitive global regulators such as CodY and CcpA which inhibit TcdR in the presence of intracellular branched chain amino acids (BCAAs) or fructose-biphosphate (FBP), respectively (Fig 2.1)[12,13]. In addition to these intracellular metabolites, the extracellular environment can also influence toxin production in *C. difficile*. Multiple defined media experiments with *C. difficile* demonstrate that certain carbohydrates, amino acids, and short chain fatty acids (SCFA) promote the increase or decrease of toxin production (Fig 2.1)[14–19]. Furthermore, *C. difficile* toxin production responses are dependent on the surrounding microbial community, as the microbial community shapes the nutritional environment of *C. difficile*[20,21]. Toxin response to environment may also be strain-dependent, further compounding the complexity of toxin production regulation[20]. While the exact relationship between the regulation of *C. difficile* toxin synthesis and extracellular environment is unclear, there is evidence that this process is linked to extra- and intracellular metabolism.

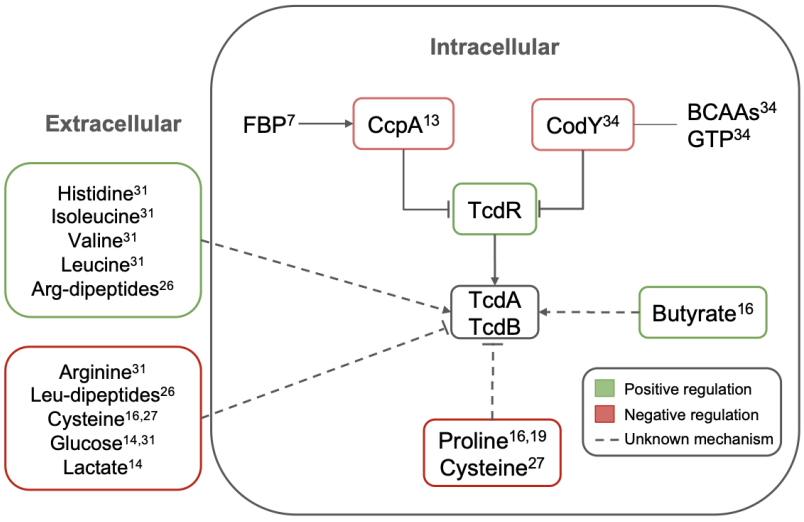


Figure 2.1. *C. difficile* toxin production is regulated by multiple metabolic signals. The transcription of *tcdA* and *tcdB* to synthesize toxins TcdA and TcdB is positively regulated by TcdR which is in turn negatively regulated by CcpA and CodY. Each of these components are regulated by multiple metabolic signals. Fructose bis-phosphate (FBP); branched-chain amino acids (BCAA); Guanosine triphosphate (GTP). Numbers correspond to references with evidence for the indicated regulatory mechanism.

To investigate the metabolic states contributing to shifts in toxin production related to changes in the environment, we use previously established and curated genome-scale metabolic network reconstructions (GENREs) of *C. difficile*[22]. Metabolic modeling provides a unique systems approach to studying metabolism. Briefly, a GENRE describes the gene-protein-reaction associations for all of the metabolic genes of an organism

which can then be simulated to predict the flux through the metabolic network[23]. We can represent the metabolic state of *C. difficile* in a specific context by integrating transcriptomic data from multiple studies (Table S2.1) with the metabolic model via the RIPTiDe algorithm[24]. Using this approach, we created context-specific metabolic models of two *C. difficile* strains (630 and R20291) for 16 different environmental conditions with a range of toxin production states. We analyzed these models using a combination of machine learning and flux balance analysis (FBA) methods and found that arginine and ornithine metabolism is more active in models with low toxin production. Moreover, we found that arginine and ornithine metabolism may be influenced by intracellular fatty acid and large polymer pools. Finally, we applied a modified metabolic transformation algorithm (mMTA) to identify pivotal reactions for transitioning from a metabolic state associated with high toxin production to one associated with low toxin production[25]. Integrating gene expression data with a metabolic network simultaneously gives the gene expression data a metabolic context and the metabolic network biological relevance, optimizing our predictive power. Metabolic network analysis of microbial pathogens can catalyze biological discoveries which can then translate to new therapeutic leads.

Results

Contextualized metabolic models of *C. difficile*

Multiple studies have shown that toxin production in *C. difficile* can be altered by changing its nutritional environment; therefore, we wanted to investigate the intracellular metabolism of *C. difficile* grown in a variety of media conditions, inducing different toxin production states[13–16,18,20,26,27]. To do this, we used published GENREs of *C. difficile* strains 630 and R20291 (iCdG709 and iCdR703, respectively)[22]. We compiled a set of publicly available RNA-sequencing data for each of these strains that considers a range of environmental conditions (Table S2.1). For each condition, we classified the toxin states as low or high based on the median *tcdA* transcript reads per million (RPM) across all conditions (Fig S2.1). Using the RIPTiDe algorithm[24], we integrated the transcript data with the appropriate strain model to generate a total of 16 contextualized models of *C. difficile* (Fig 2.2A).

On average, the RIPTiDe models for low toxin conditions include more genes than the high toxin RIPTiDe models but retain similar metabolite numbers (Fig 2.2A). The RIPTiDe models do share a core set of 483 reactions, which accounts for approximately 80% of the reactions in each model. There are strain differences, with 630 models accounting for more genes, reactions, and metabolites compared to R20291 models (Fig 2.2A). Principle component analysis (PCA) of flux samples of these models demonstrates broad clustering by strain as well as by context (Fig S2.2B-F); complete flux sampling details can be found in the Materials and Methods. Overall, these models reflect metabolic differences by strain and toxin level with more metabolic genes and reactions included in both 630 models and in low toxin models.

Metabolic differences between strains and toxin states

To identify reactions important in distinguishing between metabolic states of low and high toxin conditions, we applied a Random Forest classifier to the flux sampling data (Fig S2.3). Briefly, after randomly down-sampling the flux sampling data to 100 flux samples per RIPTiDe model, we used random stratified groups to split the flux sampling data, with a 75-25 train-test ratio (Fig S2.3). The classifier had a mean accuracy of 97%; features were ranked by their Gini score to identify reactions important in distinguishing between toxin states (Methods). From the Random Forest analysis, we determined that arginine and ornithine reactions are more active in low toxin conditions (Fig 2.2B). Specifically, growth media supplemented with bile acid (deoxycholate and cholate) and calprotectin result in increased metabolic flux through arginine and ornithine transport reactions which move towards Stickland fermentation and NAD+ production via D-proline catabolism (Fig 2.2B-C). Other metabolic processes that are more active in low toxin conditions include reactions involved in carbohydrate metabolism and nucleotide metabolism (Fig 2.2B-C). However, for most important reactions from Random Forest, the flux difference across all conditions was neither large nor significant. These results indicate that while metabolism can be predictive of toxin outcomes, it may be due to the cumulative effect of many tightly controlled reactions rather than large changes in flux through a few key reactions.

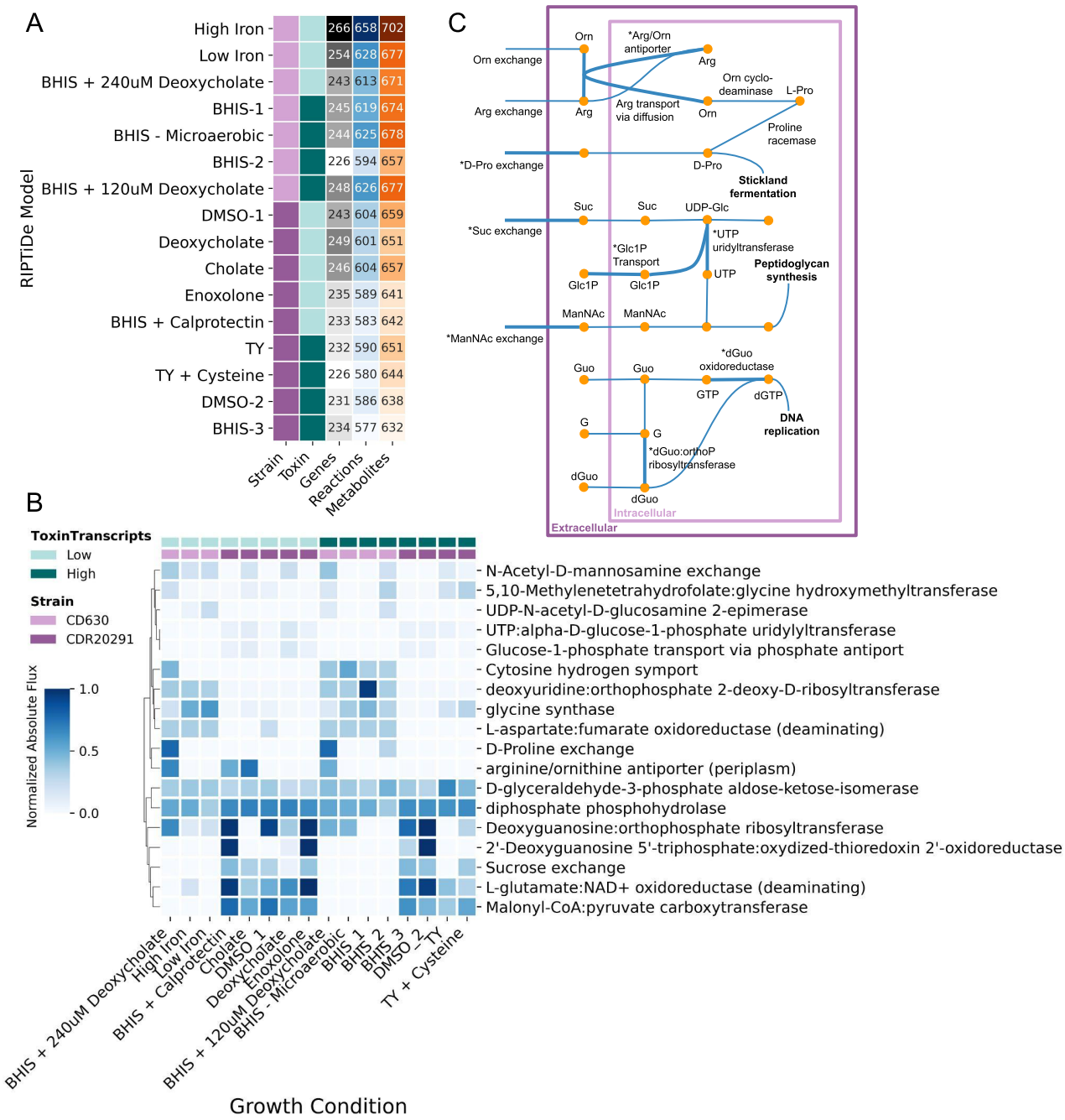


Figure 2.2. Metabolic differences between toxin states are driven by strain. (A) Summary table of the RIPTiDe contextualized models including the strain, toxin production level, and number of genes, reactions, and metabolites. (B) Normalized, absolute flux values for reactions indicated by Random Forest classifier as important for distinguishing between toxin levels. *C. difficile* strains 630 and R20291 are shown by light and dark purple respectively. Toxin transcript levels are shown by light (low) and dark (high) teal. (C) Map of reactions in the metabolic model identified by Random Forest analysis as listed in panel (A). Arg: Arginine, Orn: Ornithine, Pro: Proline, Suc: Sucrose, UDP-Glc: UDP-Glucose, Glc1P: Glucose-1-phospate, ManNAc: N-acetyl-D-mannosamine, Guo: Guanosine, dGuo: Deoxyguanosine, G: Guanine.

Because of the flux sampling results, we decided to investigate how sensitive growth and toxin production would be to disruptions in flux balance and intracellular metabolite concentrations. To approach this question, we conducted a shadow pricing analysis. Briefly, shadow pricing is the dual problem to flux balance analysis (FBA) in which shadow prices capture the sensitivity of an objective function (e.g., biomass) to changes in metabolites levels[28,29]. Thus, increases in levels of metabolites with negative shadow prices reduce flux through the objective function while increases in levels of metabolites with positive shadow prices increase flux through the objective function. For this analysis we iteratively set the top 20 reactions from the Random Forest analysis as the objective function and solved the dual problem. Reactions whose flux is predominantly increased by increasing metabolite concentrations may indicate tightly regulated reactions where a consistent flux is important or perhaps that the

reactions are not metabolically regulated. Conversely, reactions whose flux is decreased by many metabolites may be more responsive to the environment, allowing *C. difficile* to optimize its cellular function. Overall, we find that arginine and ornithine transport and L-aspartate:fumarate oxidoreductase are particularly sensitive to disruptions in metabolite concentrations with over 40 metabolites with a shadow price less than -5 (Fig 2.3). Both reactions are primarily sensitive to fatty acid and large polymer metabolite pools which have relatively high negative shadow prices (Fig 2.3B), indicating a possible source of intracellular metabolic regulation.

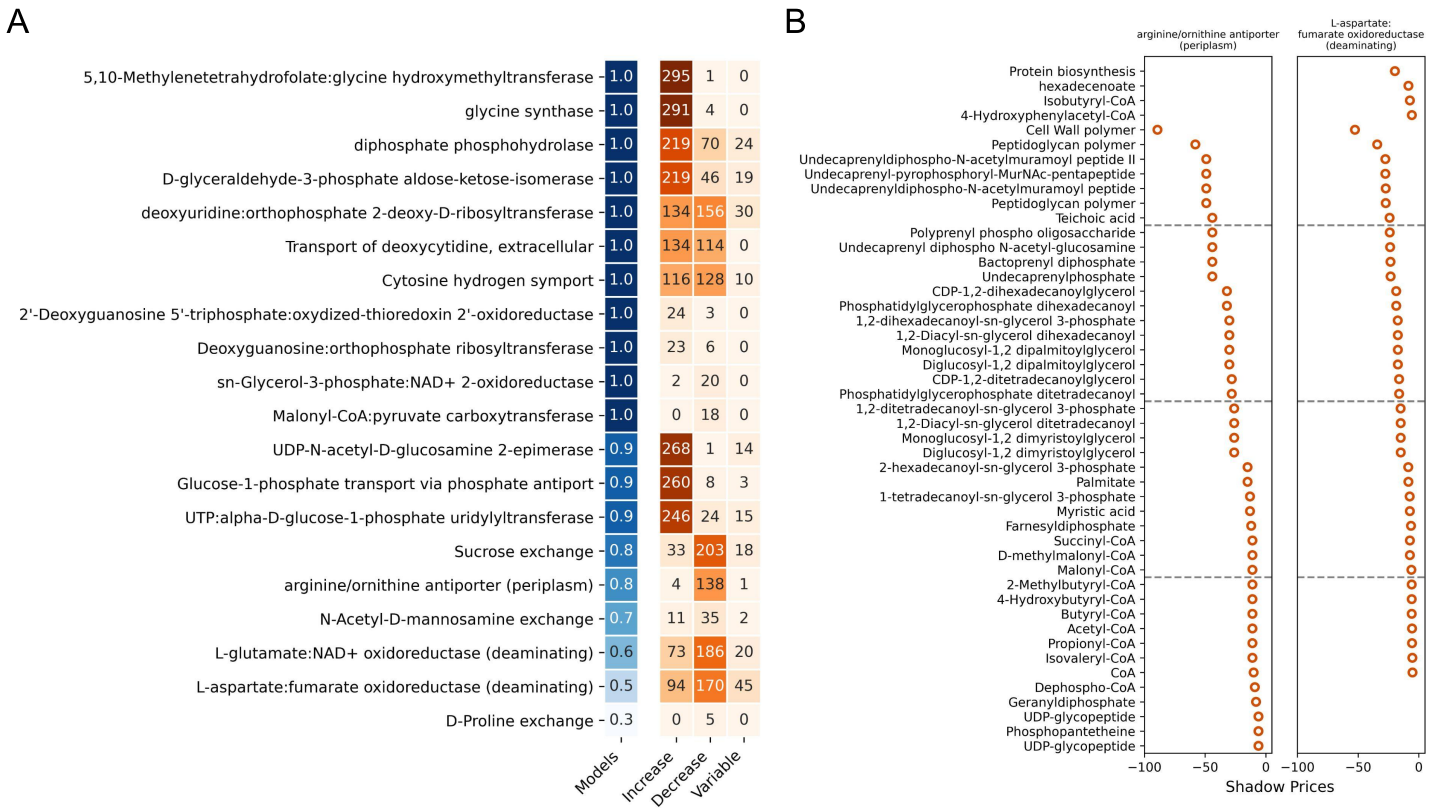


Figure 2.3. Flux through arginine and ornithine reactions is sensitive to intracellular metabolite concentrations. (A) Summary of the shadow pricing analysis with the top 20 reactions from the Random Forest classifier set as the objective function. The number in the "models" column (blue) corresponds to the fraction of contextualized models that were able to carry flux with the indicated reaction set as the objective function (OF). The values in the orange columns indicate the following: Increase: the number of metabolites for which an increased level results in increased flux through the OF (median shadow price > 0, range < 2); Decrease: the number of metabolites for which an increased level results in decreased flux through the OF (median shadow price < -0.1, range < 2); and Variable: the number of metabolites whose shadow price varied across RIPTiDe models (range > 2). For example, in the first row of panel (A), the OF was able to carry flux in all of the models, 294 metabolites had no impact on flux through the OF in all of the models, 1 metabolite limited flux through the OF in all of the models, and 5 metabolites had different effects on flux through the OF across all of the models. (B) Shadow prices for limiting metabolites in arginine/ornithine and aspartate metabolism reactions. The metabolites categorized as sensitive in panel (A) for these OFs and with a shadow price < -5 are shown. Increasing negative values indicates increasing reaction flux sensitivity to the metabolite.

Metabolic transformation between toxin states

From the flux sampling and shadow pricing analyses above, we described the metabolic state of *C. difficile* under 16 specific conditions and identified metabolic differences between low and high toxin producing conditions. We next sought to investigate whether there were pivotal reaction knockouts that could transition the model from a metabolic state associated with a high toxin transcript level to a metabolic state associated with a low toxin transcript level. For this analysis we used a modified metabolic transformation algorithm (mMTA) which identifies key reactions to switch a cell from one metabolic state to another. To do this, MTA classifies reactions as changed or unchanged based on whether there was differential flux between a reference and target state; it then solves a mixed integer quadratic programming (MIQP) problem that maximizes change in flux in the direction of the target flux for the changed reaction set while minimizing flux changes for the unchanged reaction set[25]. We modified the original MTA for compatibility with COBRApy and to use updated modeling methods throughout its implementation. These changes and considerations are detailed in Materials and Methods. For our reference

and target states we selected conditions that induce high and low toxin states respectively and that have transcriptomic data generated from the same study to reduce inter-study experimental variation. Therefore, we chose the iCdG709 RIPTiDe models contextualized with data from GSE165116, setting BHIS (high toxin) as the reference state and BHIS + 240 uM Deoxycholate (low toxin) as the target state (Table S2.1, Fig S2.1). Using mMTA, we classified 101 reactions as significantly different between the reference and target state. Out of this reaction set, 87 reactions were successfully changed in the desired flux direction in at least one reaction knockout simulation (Fig S2.6C). 396 out of 594 reaction knockouts resulted in a feasible solution; for each feasible reaction knockout, a transformation score (TS) was calculated (Fig S2.6B). The reaction knockouts with the top 50 TSs induced flux transformation in reactions related to energy metabolism, such as carbohydrate metabolism and Stickland fermentation (Fig 2.4). Of these transformed reactions, the Stickland fermentation reactions and most of the amino acid metabolism reactions are involved in isoleucine fermentation (Fig 2.4, Supplemental Data File 2.3). Isoleucine is an important energy substrate that is metabolized via oxidative Stickland fermentation to form ATP [17]. This finding not only supports the importance of energy metabolism in toxin production, but also highlights the nutritional flexibility of *C. difficile* to acquire energy from available resources.

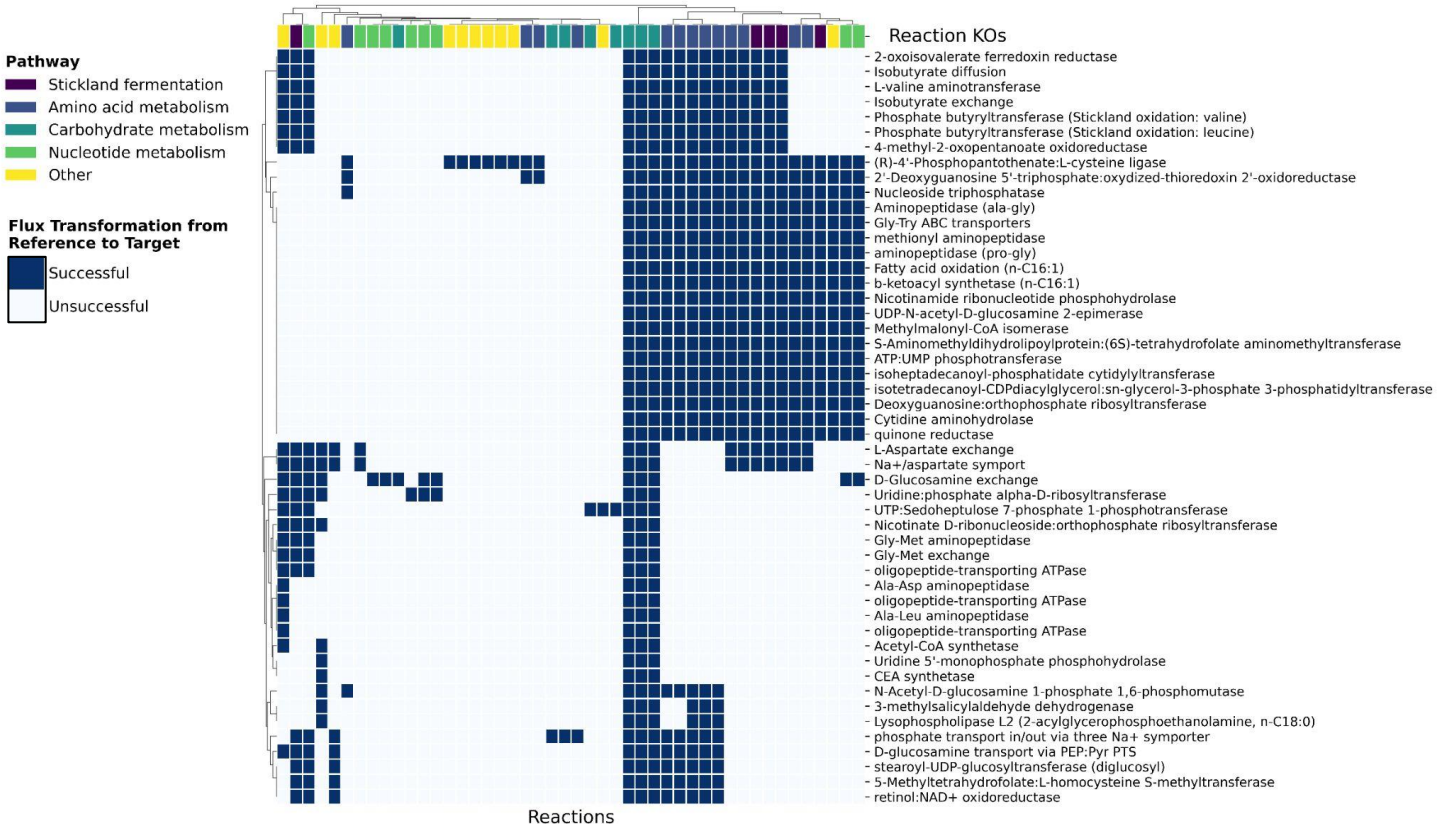


Figure 2.4. Modified MTA identifies key reaction knockouts and pathways for transformation from a high to low toxin state. The mMTA algorithm runs a reaction KO simulation to optimize changes in reaction flux that transform the model from the reference metabolic state (high toxin) to the target metabolic state (low toxin). The reaction knockouts with the highest transformation scores are shown on the y-axis. The reactions whose flux changed under these KO conditions are shown on the x-axis. Successfully changed reactions are defined as those whose flux changed from the reference in the desired direction by a minimum threshold of significance (successful: dark blue; unsuccessful: light blue). The metabolic pathways for these reactions are shown beneath the clustering dendrogram at the top.

Discussion

Research on *C. difficile* toxins has been extensive and wide-ranging, investigating biochemistry and structure, mechanisms of action, host and microbiome interactions, nutritional environments, genetic and metabolic regulation, and more[6,10–12,20,30]. Of these, the link between metabolism and regulation is particularly salient because of its therapeutic potential for virulence attenuation. Many *in vitro* studies have shown that *C. difficile* toxin production can be manipulated through its environment[13–16,18,20,31,32]. Additionally, multiple mechanisms of intracellular toxin regulation have been shown, such as cyclic di-GMP, carbon catabolite repression via *ccpA*, and nutritional limitation via *codY*[13,33,34]. However, while all these studies have elucidated various facets of *C. difficile* toxin regulation under specific conditions, our understanding of the relationship between metabolism and

toxin regulation remains fragmented. To address this, we conducted a systems analysis in which we interrogated an array of environmental conditions and toxin transcript levels with the goal of defining the metabolic state of *C. difficile* in these conditions.

The role of arginine in *C. difficile* toxin production is contradictory. Early investigation of *C. difficile* grown in minimal defined media supplemented with arginine found that increasing arginine concentrations resulted in decreased toxin production and enhanced growth[31,32]. Conversely, a study using phenotype microarrays found that arginine, and arginine dipeptides in particular, induce toxin production[26]. When we simulated growth under a subset of PM conditions, we did not find any correlations between metabolic flux and the PM toxin data. The only model constraints in this simulation were growth limits rather than intracellular limits as for the RIPTiDe models; these constraints, combined with the differences in which strain was profiled with the PM experiments and which strains were modeled with the analysis presented here, may have contributed to this lack of correlation. Cecal metabolomics of mice infected with wild-type (inflammatory) and toxin-deficient (non-inflammatory) *C. difficile* strains show that metabolic pathways for arginine and ornithine in the gut microbial community are more active in the non-inflammatory state[35]. This same study also linked ornithine metabolism in *C. difficile* with reduced inflammation during CDI. In our analysis, an arginine/ornithine transport reaction was identified as an important differentiator between toxin states by Random Forest and is particularly active in low toxin states across strains (Fig 2.2B). Shadow price analysis of these reactions found it was highly sensitive to fatty acids or large polymers primarily involved in cell wall synthesis (Fig 2.3B). Fatty acids have been shown to regulate arginine and ornithine metabolism in other organisms[36–38]. This potential regulatory interaction between fatty acids and arginine in *C. difficile* could explain the contradictory results for the impact of arginine and ornithine metabolism on toxin production.

Toxin production in *C. difficile* is clearly linked to its nutritional status. It has recently been shown that *C. difficile* grown in cooperative community environments has significantly lower toxin production than when it is grown in competitive communities[21]. We used Random Forest to attempt to identify any underlying metabolic patterns between toxin states but to understand better the intracellular metabolic switches necessary for transitioning from a high to low toxin production state we used mMTA. The results from mMTA give us two types of information. First, the metabolic reactions whose activity is important for this transition and second, the metabolic reactions that can be modulated to induce these metabolic changes. We found that within the metabolic network model corresponding to *C. difficile* grown in BHIS with high toxin production, flux through 86% of the reactions identified as important for transitioning to a low toxin state could be modulated by at least one reaction knockout. With the Random Forest analysis we performed, we found that the 20 reactions with the highest Gini scores are heavily involved in energy metabolism (e.g., arginine/ornithine, glycine, glutamate, aspartate, sucrose, glucose, and N-acetyl-D-mannosamine) as well as nucleotide metabolism (e.g., guanosine and uridine). Similar patterns of important metabolic pathways are replicated in the mMTA results; reactions whose flux can be changed to mimic a low toxin state fall into carbohydrate, amino acid, and nucleotide metabolism categories (Fig 2.4). The mMTA results also predict that these reactions can be metabolically modulated via knockouts of specific reactions (as indicated on the y-axis in Fig 2.4). The reaction knockouts with the highest transformation scores were frequently key reactions in alternative energy-generating pathways such as carbon metabolism and Stickland fermentation of leucine and valine (Fig 2.4). When these reactions were knocked out, Stickland fermentation of isoleucine increased (Supplemental Data 2.3).

Isoleucine is another metabolite whose role in toxin production in *C. difficile* is unclear. Isoleucine activates CodY which represses toxin production (Fig 2.1)[39]; it is reasonable to hypothesize that conditions supplemented with isoleucine would have low toxin production. However, defined media experiments show the opposite effect. *C. difficile* (VPI 10463) grown in a minimal media supplemented with isoleucine demonstrated increased TcdA production[32]. Another study growing *C. difficile* (ATCC 9689) in phenotype microarrays found that isoleucine induced middle-level toxin production[26]. It may be that preferential fermentation of isoleucine throughout the exponential growth phase depletes stores of bioavailable isoleucine for CodY activation in stationary phase, resulting in CodY deactivation and increased toxin production. While our mMTA results show that isoleucine fermentation was maximized in the target low toxin state, perhaps the driving differential feature between states is carbohydrate metabolism (Fig S2.7). In the reference state, increased glucose metabolism is likely driving the high toxin transcript levels[14]; in the target state, suppression of glucose metabolism is accompanied by an increase in isoleucine fermentation to maintain energy generation. In the short-term, this increased isoleucine

fermentation likely results in increased uptake and availability of isoleucine which is sufficient for CodY activation and suppression of toxin transcription.

We used *C. difficile* GENREs and publicly available transcriptomic data for our analyses, applying RIPTiDe for data integration. RIPTiDe uses genetic evidence from the transcriptomic data as weights indicating the likelihood that a reaction will carry flux and to what extent [24]. It additionally prunes reactions for which there is no evidence or that do not pass a minimum flux threshold. This approach creates a contextualized model for a specific environmental and metabolic state. However, it is possible that this method may have inadvertently limited the results of our mMTA analysis. The flux bounds of reactions in a RIPTiDe model are set based on genetic evidence and are often quite restrictive because the goal of RIPTiDe is to describe the metabolic state of an organism in a specific condition. Therefore, any significant departure from the reference state flux necessary for a reaction to achieve a target state may not even be feasible due to preset flux bounds. However, we were able to successfully transform 86% of the reactions targeted for change, and therefore do not consider any potential limitation from application of RIPTiDe as having a significant impact on our mMTA results.

We additionally used RIPTiDe to sample the flux distributions for each model which we then used for our Random Forest analysis. The flux samples for each condition are highly correlated in part due to the RIPTiDe restrictions discussed above; this characteristic could lead to overfitting in a Random Forest model because the model would be able to learn what condition a sample is from and then use this information to infer the toxin level. To prevent this, we used a random stratified group sampling approach for splitting the data into train-test sets for Random Forest (Methods). This approach ensures first that there is an equal (or near-equal) ratio of low and high toxin conditions in the train and test sets and second that all the flux samples from a single RIPTiDe model are used in either the train or test set. The classifier had a mean accuracy of 95% which suggests model over-fitting despite the feature selection and sampling approaches we took. A closer look at the features with the highest Gini scores shows that there is a greater difference in flux between strains than in flux between toxin states (Fig 2.2). While the classifier is not learning what condition a sample is from, it may be learning what strain it is and then predicting the toxin state based on strain-specific criteria. This result highlights the importance in accounting for strain differences when interrogating toxin production in *C. difficile*.

The reactions from the Random Forest Classifier with the highest Gini scores were analyzed in a shadow price analysis. The results of this analysis showed that arginine and ornithine transport as well as an aspartate fumurate oxidoreductase were highly sensitive to fatty acids and large polymers. However, two metabolites that also commonly occurred as limiting (shadow price < -5) were “Protein biosynthesis” and “Cell Wall Polymer” (Fig 2.3, Fig S2.5). These metabolites are not true biological metabolites but rather *in silico* substitutions for high-level cellular processes just upstream of the biomass reaction within the model. While the biomass reaction is not set as the objective function in the shadow pricing analysis, there may be underlying biases that drive flux towards biomass production. Regulation via these two “metabolites” may indicate an intracellular sensing mechanism or be used as proxy for cell status but it is also possible that they are merely modeling artefacts.

In conclusion, we performed a systems analysis of *C. difficile* metabolism under different growth conditions, paired with the associated toxin transcript level to define the relationship between metabolism and toxin production. These toxins are essential in establishing a nutritional niche for *C. difficile* and can cause extensive damage in the host colon. CDI is most effectively resolved through fecal microbiota transplants (FMTs); however, FMTs are typically only prescribed for patients with severe or recurrent cases of CDI[40]. Using microbial engineering to design probiotic communities that can be offered as a non-invasive CDI therapeutic is a major advancement already underway within the field[21,41,42]. An important step in designing these therapeutic communities is identifying reactions or pathways associated with high and low toxin production and understanding how those reactions change as a function of the environment, resulting in specific toxin-associated phenotypes. Future research investigating questions include accounting for the relationship between regulatory networks and metabolism in *C. difficile* toxin production as we know that toxin responses are in part the effect of global regulators. Additionally, modeling the effect of toxin activity once it is released from *C. difficile* could also help guide selection of members of a microbial community. Research in these areas will provide foundational understanding of *C. difficile* biology that will enable intentional and specific therapeutic community design.

Methods

Processing the RNA-sequencing data: We compiled transcriptome count matrices from seven publicly available RNA-sequencing studies of *C. difficile* covering two strains (630 and R20291) and 16 unique conditions (Supplemental Table 1). Raw count matrices were normalized using the reads per million (RPM) formula

$$\frac{\text{Number of reads mapped to gene} \times 10^6}{\text{Total number of mapped reads}}.$$

RPM normalized toxin gene transcripts from all studies were grouped by condition, averaged, and binned into high and low categories by *tcdA* medians (low < median < high) (Fig S2.1). We used medians of *tcdA* rather than *tcdB* because *tcdB* had low to no expression across all conditions with the exception of the tryptone yeast conditions (TY and TY + Cysteine).

Genome-scale metabolic models: Previously published *C. difficile* genome-scale metabolic network reconstructions (GENREs) of strains 630 and R20291 (referred to as iCdG709 and iCdR703, respectively) were used for all of the modeling analyses [22]. We created a total of 16 contextualized metabolic models of *C. difficile* using RIPTiDe and the processed transcriptomic read matrices from publicly available RNA-Seq studies (Table S2.1) [24].

Random Forest analysis of flux samples: For each RIPTiDe model, we optimized for biomass production and then sampled (n=500) flux distributions of the entire feasible steady-state solution space using RIPTiDe. We then randomly down sampled to 100 flux samples per condition and performed a principal component analysis (PCA) of sampled flux distributions for all the models.

To identify the most important reactions in differentiating between toxin production (low, high) we ran a Random Forest classifier with 500 trees on the down-sampled flux sampling data (100 flux samples per RIPTiDe model). We reduced the feature space by selecting features with a near-zero variance (NZV) > 0.005 and an absolute Pearson's correlation coefficient < 0.8. We used random stratified group K-fold (k=5) cross validation to check the classifier (Fig S2.3). Using a stratified group k-fold to split the data into train and test sets ensures first that there is an equal (or near-equal) ratio of low and high toxin conditions in the train and test sets and second that all the flux samples from a single RIPTiDe model are used in either the train or test set. We used this approach to prevent the classifier from learning which RIPTiDe model the flux sample came from and substituting that information to infer toxin level; while there are many flux samples per RIPTiDe model, these flux samples tend to be highly correlated. For model predictions, the data was split using a random stratified group split as described for the cross validation. The classifier was then trained on 75% of the data and tested on the remaining 25%.

Following classifier testing, we ranked the features by their Gini score and selected the top 20 most important features for model predictions. For these 20 reactions, we calculated the median flux value for each condition, normalized, and visualized using a heatmap (Fig 2.2B). We tracked the flow of flux through these reactions to create a human readable metabolic map (Fig 2.2A) as well as an Escher metabolic map [43] with the GENRE IDs for the reactions and metabolites (Fig S2.4). To investigate metabolism of *C. difficile* that literature indicates can impact toxin production, we compiled lists of identifiers for reactions within a specific metabolic pathway. We did this analysis for three metabolic processes: Stickland fermentation, ATP production, and redox reactions. We filtered and processed the flux sampling data in the same way as for the Random Forest results.

Shadow pricing analysis: The 20 reactions with the highest Gini scores from the Random Forest analysis were iteratively set as the objective function for each RIPTiDe model. This model was then optimized and the corresponding shadow price for the FBA solution was saved. For each metabolite and each objective function, we calculated the median shadow price and range across all RIPTiDe models. We summarized the shadow pricing results for each objective function (OF) across all RIPTiDe models using the following metrics: fraction of RIPTiDe models able to carry flux for that OF, total number of metabolites that increase flux through the OF (median shadow price < -0.1, range < 2), total number of metabolites that decrease flux through the OF (median shadow price > 0, range < 2), total number of metabolites whose shadow price varied across RIPTiDe models (range > 2). We then plotted the metabolites with a median shadow price < -5 and a range < 2 for the 9 objective functions that had metabolites in this category (Fig 2.3B, Fig S2.5).

Metabolic transformation algorithm (MTA): The goal of MTA is to identify perturbations that transform a metabolic network from a reference state (e.g., diseased) to a target state (e.g., healthy)[25]. In our case, we used the MTA to find reaction knockouts that lead to transformation of high toxin states to low toxin states in *C. difficile*. The generic MTA is comprised of four distinct steps which are briefly: (1) calculate a flux solution for the reference state, v_{ref}, (2) identify reactions that are changed in the forward (R_F) or backward (R_B) direction and unchanged (R_S) between the reference and target state, (3) solve the MIQP optimization problem formulated to minimize change in R_S and maximize change in R_F and R_B, and (4) calculate a transformation score (TS) to quantify the success of each reaction knockout in transforming the reference state to the target state. To successfully apply MTA to our problem, we made changes to the original formulation at each step, resulting in a modified MTA (mMTA, described below) compatible with COBRApy tools.

Step 1: We created two contextualized metabolic models (reference and target) using RIPTiDe with gene expression data and then generated 500 flux samples using RIPTiDe[24]. Because the mean (or median) of the flux samples is not a mass-balanced solution, setting it as v_{ref} can lead to infeasible MIQP solutions downstream. Therefore, we used a Bray-Curtis non-parametric multidimensional scaling (NMDS) to reduce the flux samples to a two-dimensional space, then calculated the centroid of the flux sampling distribution, and finally calculated the point closest to the centroid and set this flux sample as v_{ref} (Fig S2.6A).

Step 2: We determine significantly changed and unchanged reactions by using a Mann-Whitney U test with a Bonferroni multiple tests correction. We categorize all reactions into three sets: statistically insignificant reactions (R_S), and statistically significant reactions which must change in the forward (R_F) or backward (R_B) direction in order to match the target state. The threshold for statistical significance is an adjusted p-value < 0.05.

Step 3: The goal of the MIQP problem is to minimize changes in flux for reactions in R_S and maximize changes in flux for reactions in R_F and R_B in the intended direction. We implemented the MIQP formulation as it was set out for the gMTA:

1) $\min_{v,y}((1-\alpha)\sum_{i\in R_S}(v_i^{ref}-v_i)^2 + \frac{\alpha}{2}\sum_{i\in R_F} y_i + \frac{\alpha}{2}\sum_{i\in R_B} y_i)$

$s.t.$

2) $S \cdot v = 0$
3) $v_{min} \leq v \leq v_{max}$
4) $v_j = 0$
5) $v_i - y_i^F\left(v_i^{ref} + \varepsilon_i\right) - y_i v_i^{min} \geq 0, i \in R_F$
6) $y_i^F + y_i = 1, i \in R_F$
7) $v_i - y_i^B\left(v_i^{ref} - \varepsilon_i\right) - y_i v_i^{max} \geq 0, i \in R_B$
8) $y_i^B + y_i = 1, i \in R_B$
9) $y_i, y_i^F, y_i^B \in \{0,1\}$

Mass balance constraints, thermodynamic constraints, and the reaction knockout perturbation are enforced in equations 2, 3, and 4 respectively. The demands for changed reactions are represented in equations 5-8, such that the Boolean variables y_i, y_i^F, y_i^B indicate whether a forward reaction (R_F) either increases by more than ε with respect to v^{ref} or maintains a preset flux minimum and whether a backward reaction (R_B) either decreases by more than ε with respect to v^{ref} or maintains a preset flux maximum. ε is the vector of thresholds used to determine if flux changes are statistically significant ($p < 0.05$) and was calculated using a one-sided T-test with a 95% confidence interval, such that $\varepsilon = t\frac{S}{\sqrt{n}}$ where t is the critical value, S is the standard deviation, and n is the number of flux samples. We set α = 0.66 as in the original formulation.

Step 4: We categorize forward and backward reactions as successful if $v_{RF} > (v_{RF}^{ref} + \varepsilon)$ or if $v_{RB} < (v_{RB}^{ref} - \varepsilon)$, respectively. Next, to quantify how well each reaction knock-out transformed the reference state to the target state, we calculated the transformation score (TS) as formulated for the gMTA:

10) $$\frac{\sum_{i\in R_{Success}} abs[\left(v_i^{ref} - v_i^{res}\right]] - \sum_{i\in R_{unsuccess}} abs[(v_i^{ref} - v_i^{res})]}{\sum_{i\in R_S} abs[(v_i^{ref} - v_i^{res})]}$$

Ranking reaction knockouts by the TS provides a helpful metric for evaluating the flux solution from each knockout while taking the MIQP objective value would result in reaction knockouts with varying degrees of transformational success being set as equivalent (Fig S2.6B).

Phenotype Microarray (PM) data integration and analysis: We used public data from a PM study that measured toxin production of *C. difficile* type strain ATCC 9689 when grown in each condition[26]. The toxin concentrations from this study were calculated by comparing the amount of dye reduction in cell cytotoxicity assays to a standard curve of toxin concentrations[26]. The authors defined toxins as low (<42 ng/uL), mid (42-420 ng/uL), or high (>420 ng/uL) and we used the same categories in this analysis. The dataset from this PM study included 652 unique growth conditions. The GENREs iCdG709 and iCdR703 contain 171 unique extracellular metabolites, 65 of which overlapped with metabolites from the PM dataset (Fig S2.8A). We constrained the GENREs to minimal media conditions and iteratively added one of the 65 overlapping metabolites and simulated flux while optimizing for biomass. We normalized the flux sampling data using min-max normalization and then removed reactions with variance < 0.05. This step trimmed the flux data from 1323 reactions to 67 reactions. Next, we calculated the Pearson's correlation between each reaction flux vector and the PM toxin data. None of the reactions were correlated with toxin production. Finally, we visualized the absolute flux data for each of the 65 simulated PM conditions (Fig S2.8C).

Data availability

The transcriptomic data was retrieved from public databases; the GEO IDs for each study can be found in Supplemental Table 1. The flux sampling, shadow pricing, and mMTA data are shared in Supplemental Data Files 2.1, 2.2, and 2.3 respectively. The scripts used to generate and analyze the data are available on GitHub (https://github.com/dap5mb/cdToxinAnalysis).

References

1. Bouwknegt, M., van Dorp, S. & Kuijper, E. Burden of Clostridium difficile infection in the United States. *N Engl J Med* **372**, 2368 (2015).
2. Zhang, S. *et al.* Cost of hospital management of Clostridium difficile infection in United States—a meta-analysis and modelling study. *BMC Infect Dis* **16**, 447 (2016).
3. Ross, C. L., Spinler, J. K. & Savidge, T. C. Structural and functional changes within the gut microbiota and susceptibility to Clostridium difficile infection. *Anaerobe* **41**, 37–43 (2016).
4. Olsen, M. A., Yan, Y., Reske, K. A., Zilberberg, M. D. & Dubberke, E. R. Recurrent Clostridium difficile infection is associated with increased mortality. *Clin Microbiol Infect* **21**, 164–170 (2015).
5. Chandrasekaran, R. & Lacy, D. B. The role of toxins in Clostridium difficile infection. *FEMS Microbiol Rev* **41**, 723–750 (2017).
6. Voth, D. E. & Ballard, J. D. Clostridium difficile toxins: mechanism of action and role in disease. *Clin Microbiol Rev* **18**, 247–263 (2005).
7. Vedantam, G. *et al.* Clostridium difficile infection: toxins and non-toxin virulence factors, and their contributions to disease establishment and host response. *Gut Microbes* **3**, 121–134 (2012).
8. D'Auria, K. M. *et al.* Systems analysis of the transcriptional response of human ileocecal epithelial cells to Clostridium difficile toxins and effects on cell cycle control. *BMC Syst Biol* **6**, 2 (2012).
9. D'Auria, K. M. *et al.* In vivo physiological and transcriptional profiling reveals host responses to Clostridium difficile toxin A and toxin B. *Infect Immun* **81**, 3814–3824 (2013).
10. D'Auria, K. M. *et al.* High temporal resolution of glucosyltransferase dependent and independent effects of Clostridium difficile toxins across multiple cell types. *BMC Microbiol* **15**, 7 (2015).
11. Aktories, K., Schwan, C. & Jank, T. Clostridium difficile Toxin Biology. *Annu Rev Microbiol* **71**, 281–307 (2017).
12. Martin-Verstraete, I., Peltier, J. & Dupuy, B. The Regulatory Networks That Control Clostridium difficile Toxin Synthesis. *Toxins (Basel)* **8**, E153 (2016).
13. Antunes, A., Martin-Verstraete, I. & Dupuy, B. CcpA-mediated repression of Clostridium difficile toxin gene expression: C. difficile toxin regulation by CcpA. *Molecular Microbiology* **79**, 882–899 (2011).
14. Hofmann, J. D. *et al.* Influence of L-lactate and low glucose concentrations on the metabolism and the toxin formation of Clostridioides difficile. *PLoS One* **16**, e0244988 (2021).
15. Karlsson, S., Burman, L. G. & Åkerlund, T. Suppression of toxin production in Clostridium difficile VPI 10463 by amino acids. *Microbiology (Reading)* **145 (Pt 7)**, 1683–1693 (1999).

16. Karlsson, S., Lindberg, A., Norin, E., Burman, L. G. & Akerlund, T. Toxins, butyric acid, and other short-chain fatty acids are coordinately expressed and down-regulated by cysteine in Clostridium difficile. *Infect Immun* **68**, 5881–5888 (2000).
17. Neumann-Schaal, M., Jahn, D. & Schmidt-Hohagen, K. Metabolism the Difficile Way: The Key to the Success of the Pathogen Clostridioides difficile. *Front. Microbiol.* **10**, 219 (2019).
18. Yamakawa, K., Kamiya, S., Meng, X. Q., Karasawa, T. & Nakamura, S. Toxin production by Clostridium difficile in a defined medium with limited amino acids. *J Med Microbiol* **41**, 319–323 (1994).
19. Bouillaut, L., Self, W. T. & Sonenshein, A. L. Proline-dependent regulation of Clostridium difficile Stickland metabolism. *J Bacteriol* **195**, 844–854 (2013).
20. Carlucci, C. *et al.* Effects of defined gut microbial ecosystem components on virulence determinants of Clostridioides difficile. *Sci Rep* **9**, 885 (2019).
21. Jenior, M. L. *et al.* Systems-ecology designed bacterial consortium protects from severe Clostridioides difficile infection. *Under Review* (2022).
22. Jenior, M. L. *et al.* Novel Drivers of Virulence in Clostridioides difficile Identified via Context-Specific Metabolic Network Analysis. *mSystems* **6**, e0091921 (2021).
23. Oberhardt, M. A., Chavali, A. K. & Papin, J. A. Flux Balance Analysis: Interrogating Genome-Scale Metabolic Networks. in *Systems Biology* (ed. Maly, I. V.) vol. 500 61–80 (Humana Press, Totowa, NJ, 2009).
24. Jenior, M. L., Moutinho, T. J., Dougherty, B. V. & Papin, J. A. Transcriptome-guided parsimonious flux analysis improves predictions with metabolic networks in complex environments. *PLoS Comput Biol* **16**, e1007099 (2020).
25. Yizhak, K., Gabay, O., Cohen, H. & Ruppin, E. Model-based identification of drug targets that revert disrupted metabolism and its application to ageing. *Nat Commun* **4**, 2632 (2013).
26. Lei, X.-H. & Bochner, B. R. Using Phenotype MicroArrays to Determine Culture Conditions That Induce or Repress Toxin Production by Clostridium difficile and Other Microorganisms. *PLoS ONE* **8**, e56545 (2013).
27. Dubois, T. *et al.* Control of Clostridium difficile Physiopathology in Response to Cysteine Availability. *Infect Immun* **84**, 2389–2405 (2016).
28. Reznik, E., Mehta, P. & Segrè, D. Flux Imbalance Analysis and the Sensitivity of Cellular Growth to Changes in Metabolite Pools. *PLoS Comput Biol* **9**, e1003195 (2013).
29. Tervo, C. J. & Reed, J. L. Expanding Metabolic Engineering Algorithms Using Feasible Space and Shadow Price Constraint Modules. *Metab Eng Commun* **1**, 1–11 (2014).
30. Fletcher, J. R. *et al.* Clostridioides difficile exploits toxin-mediated inflammation to alter the host nutritional landscape and exclude competitors from the gut microbiota. *Nat Commun* **12**, 462 (2021).
31. Karasawa, T., Maegawa, T., Nojiri, T., Yamakawa, K. & Nakamura, S. Effect of Arginine on Toxin Production by *Clostridium difficile* in Defined Medium. *Microbiology and Immunology* **41**, 581–585 (1997).
32. Ikeda, D. *et al.* Effect of Isoleucine on Toxin Production by Clostridium difficile in a Defined Medium. *Zentralblatt für Bakteriologie* **287**, 375–386 (1998).
33. McKee, R. W., Mangalea, M. R., Purcell, E. B., Borchardt, E. K. & Tamayo, R. The second messenger cyclic Di-GMP regulates Clostridium difficile toxin production by controlling expression of sigD. *J Bacteriol* **195**, 5174–5185 (2013).
34. Dineen, S. S., McBride, S. M. & Sonenshein, A. L. Integration of metabolism and virulence by Clostridium difficile CodY. *J Bacteriol* **192**, 5350–5362 (2010).
35. Pruss, K. M. *et al.* Oxidative ornithine metabolism supports non-inflammatory C. difficile colonization. *Nat Metab* **4**, 19–28 (2022).
36. Guelzim, N. *et al.* Regulation of arginine metabolism by dietary fatty acids: involvement of PPARα. *Proc. Nutr. Soc.* **67**, E203 (2008).
37. Li, S. *et al.* Dietary L-arginine supplementation reduces lipid accretion by regulating fatty acid metabolism in Nile tilapia (Oreochromis niloticus). *J Animal Sci Biotechnol* **11**, 82 (2020).
38. Jobgen, W. S., Fried, S. K., Fu, W. J., Meininger, C. J. & Wu, G. Regulatory role for the arginine-nitric oxide pathway in metabolism of energy substrates. *J Nutr Biochem* **17**, 571–588 (2006).
39. Sonenshein, A. L. CodY, a global regulator of stationary phase and virulence in Gram-positive bacteria. *Current Opinion in Microbiology* **8**, 203–207 (2005).
40. Baunwall, S. M. D. *et al.* Faecal microbiota transplantation for first or second Clostridioides difficile infection (EarlyFMT): a randomised, double-blind, placebo-controlled trial. *The Lancet Gastroenterology & Hepatology* **7**, 1083–1091 (2022).

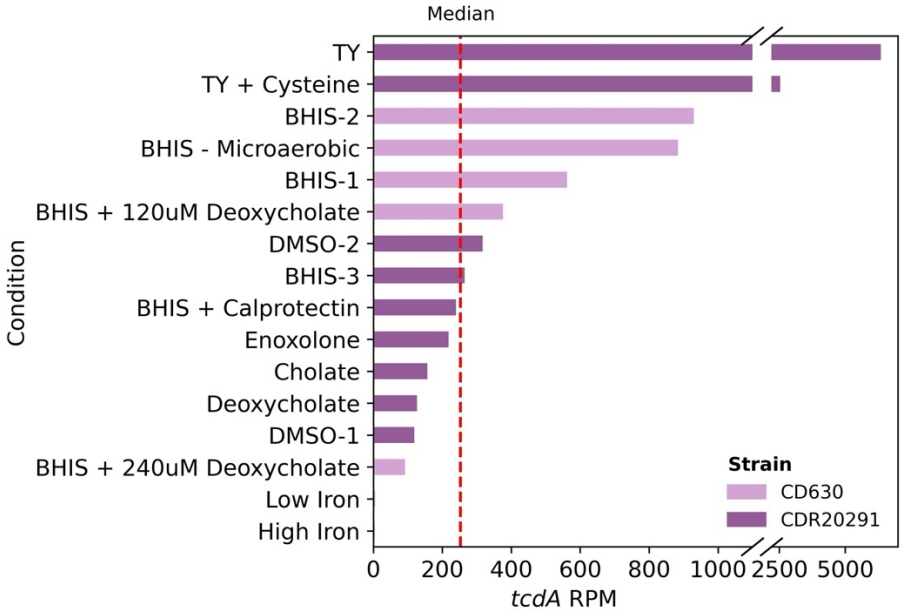
41. Smith, A. B. *et al.* Enterococci enhance Clostridioides difficile pathogenesis. *Nature* **611**, 780–786 (2022).
42. Bohmann, N. *et al. Microbial Community-Scale Metabolic Modeling Predicts Personalized Short-Chain-Fatty-Acid Production Profiles in the Human Gut*. http://biorxiv.org/lookup/doi/10.1101/2023.02.28.530516 (2023) doi:10.1101/2023.02.28.530516.
43. King, Z. A. *et al.* Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways. *PLoS Comput Biol* **11**, e1004321 (2015).

Supplemental Tables

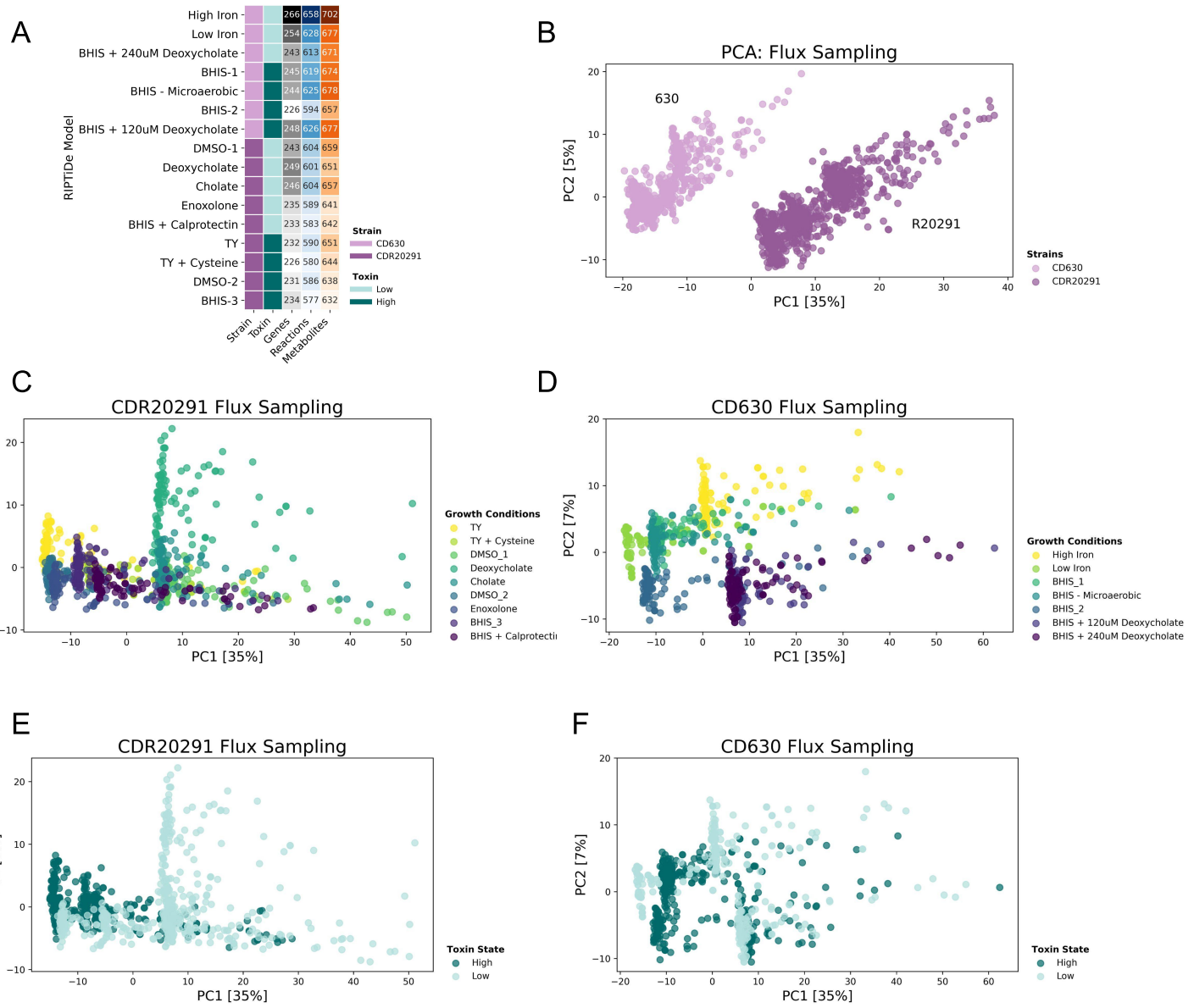
GEO ID	Strain	Growth Condition (RIPTiDe Name)	Replicates	Authors	Year
GSE173804	630	BHIS – Anaerobic (BHIS-1)	3	Weiss A, et al.	2021
		BHIS – microaerobic	3		
GSE120189	630	CDMM + low iron (Low Iron)	3	Berges M, et al.	2018
		CDMM + high iron (High Iron)	3		
GSE165116	630	BHIS (BHIS-2)	3	Monot M, et al.	2021
		BHIS + 120 uM DCA	3		
		BHIS + 240 uM DCA	3		
GSE135912	R20291	BHIS (BHIS-3)	3	Lopez CA, et al.	2019
		BHIS + Calprotectin (Calprotectin)	3		
GSE199109	R20291	BHIS + DMSO (DMSO-1)	3	Mareddy RKR, et al.	2022
		BHIS + Enoxolone (Enoxolone)	3		
GSE107961	R20291	TY	2	Gu H, et al.	2018
		TY + 5 mM Cysteine	2		
GSE86152	R20291	5 mM DMSO (DMSO-1)	4	Sorg J, Monot M	2019
		0.5 mM DCA (Deoxycholate)	4		
		5 mM Cholate (Cholate)	4		

Supplemental Table 2.1. Public RNA-sequencing datasets. BHIS(G): Brain-Heart Infusion Supplemented (Glucose), Cd: *C. difficile*, CDMM: *C. difficile* Minimal Media, DCA: Deoxycholate, DMSO: Dimethyl Sulfoxide, GEO ID: Gene Expression Omnibus Identifier, TY: Tryptone Yeast. Alternate identifiers for RIPTiDe models with similar growth conditions are indicated in parentheses in the Growth Condition column when applicable; these identifiers are used for all analyses.

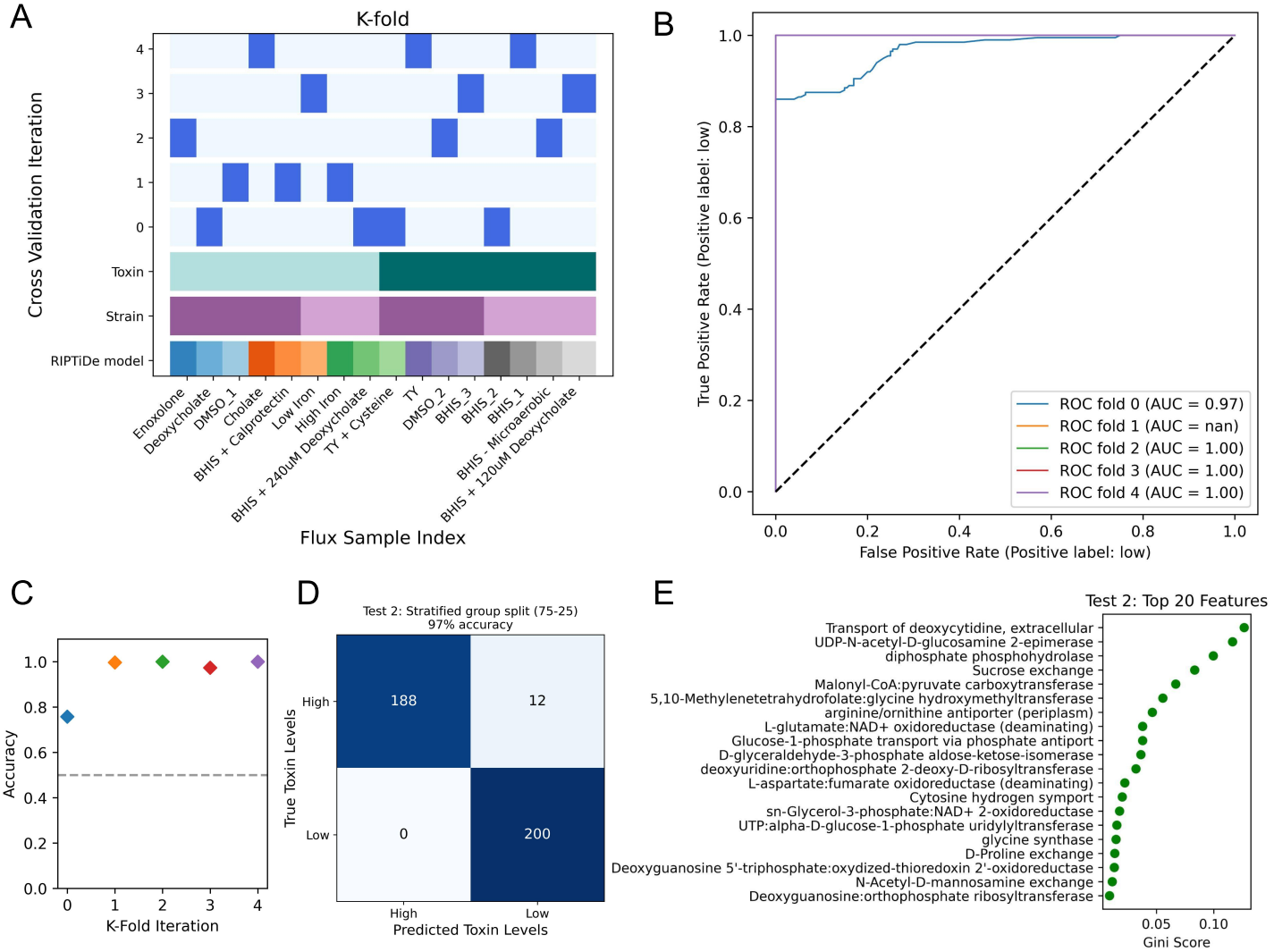
Supplemental Figures



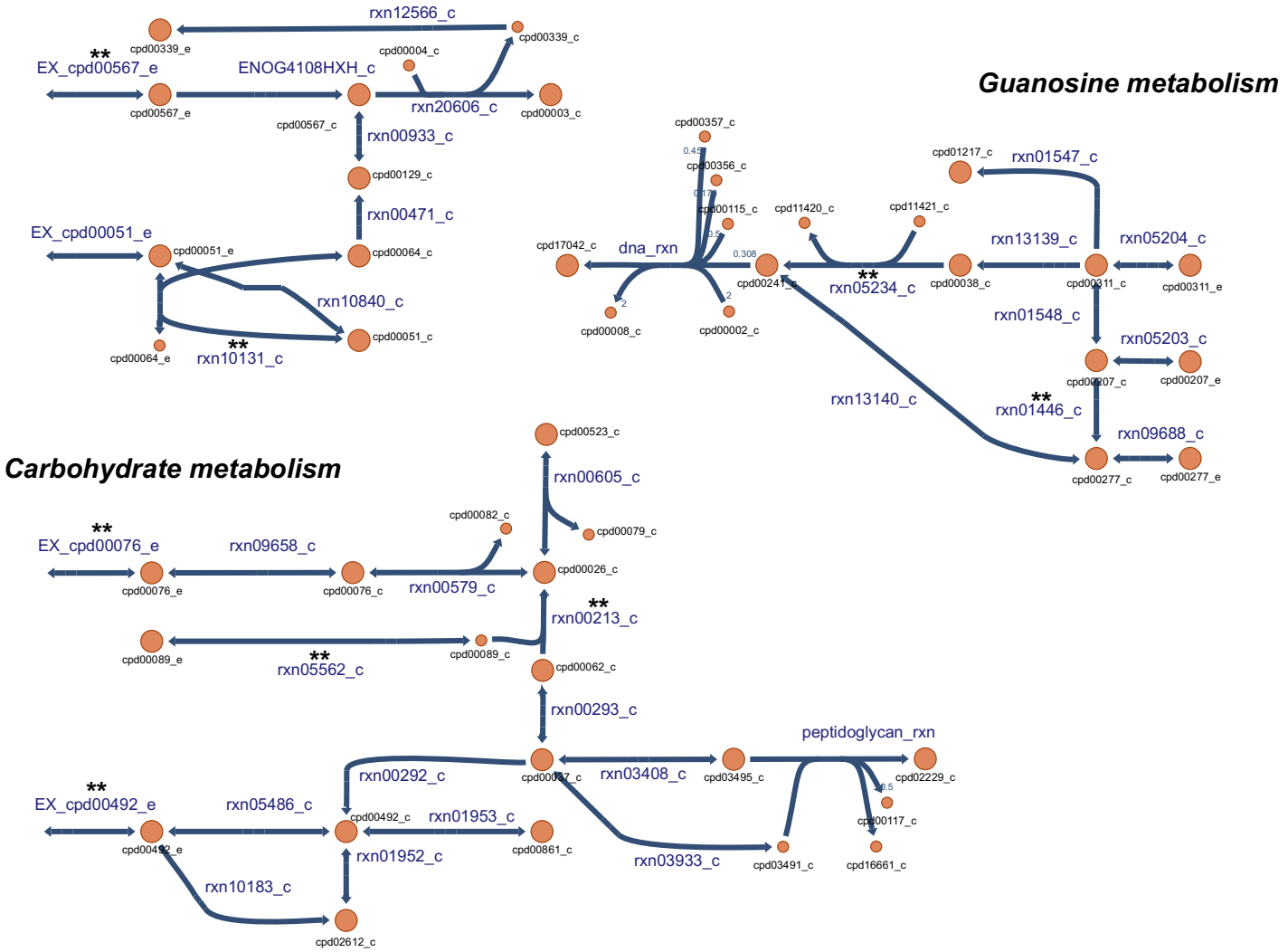
Supplemental Figure 2.1. Toxin transcript counts across conditions. Toxin transcript counts quantified as reads per million (RPM) are shown for all conditions included in the study (see Supplemental Table 1 for more details). Conditions were binned based on median *tcdA* transcript levels across all conditions and labeled as low (< median) or high (> median).



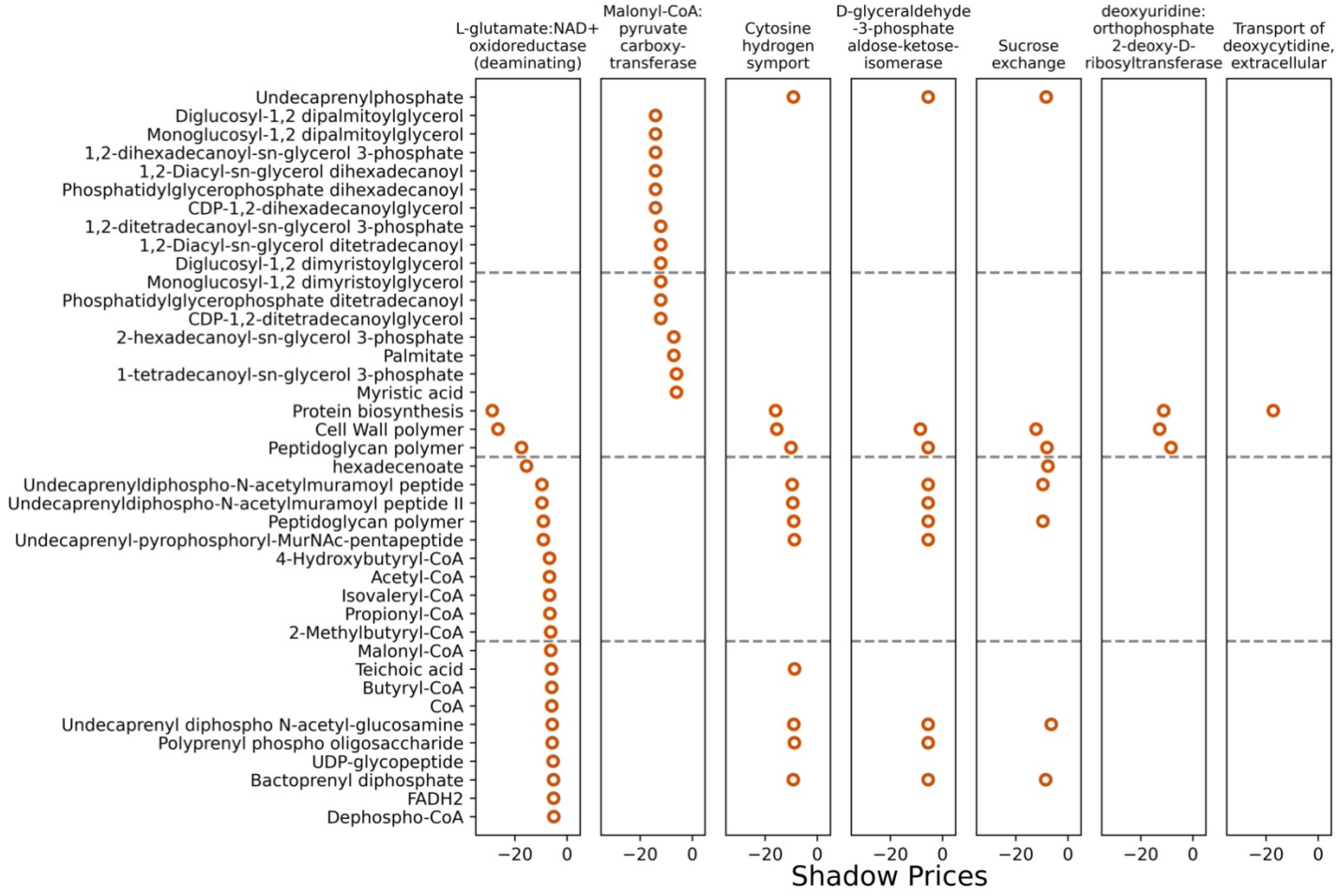
Supplemental Figure 2.2. PCA of flux sampling of RIPTiDe-contextualized iCdG709 and iCdR703 models. (A) Summary table of the RIPTiDe-contextualized models including the strain, toxin production level, and number of genes, reactions, and metabolites. (B) The iCdG709 (CD630, light purple) and iCdR703 (CDR20291, dark purple) *C. difficile* models were contextualized with transcriptomic data (Supplemental Table 1) and flux distributions were sampled (n=500) using RIPTiDe. The flux sampling for each model was randomly down-sampled to 100 flux samples and PCA was performed for all the models together (B) and by strain (C–F).



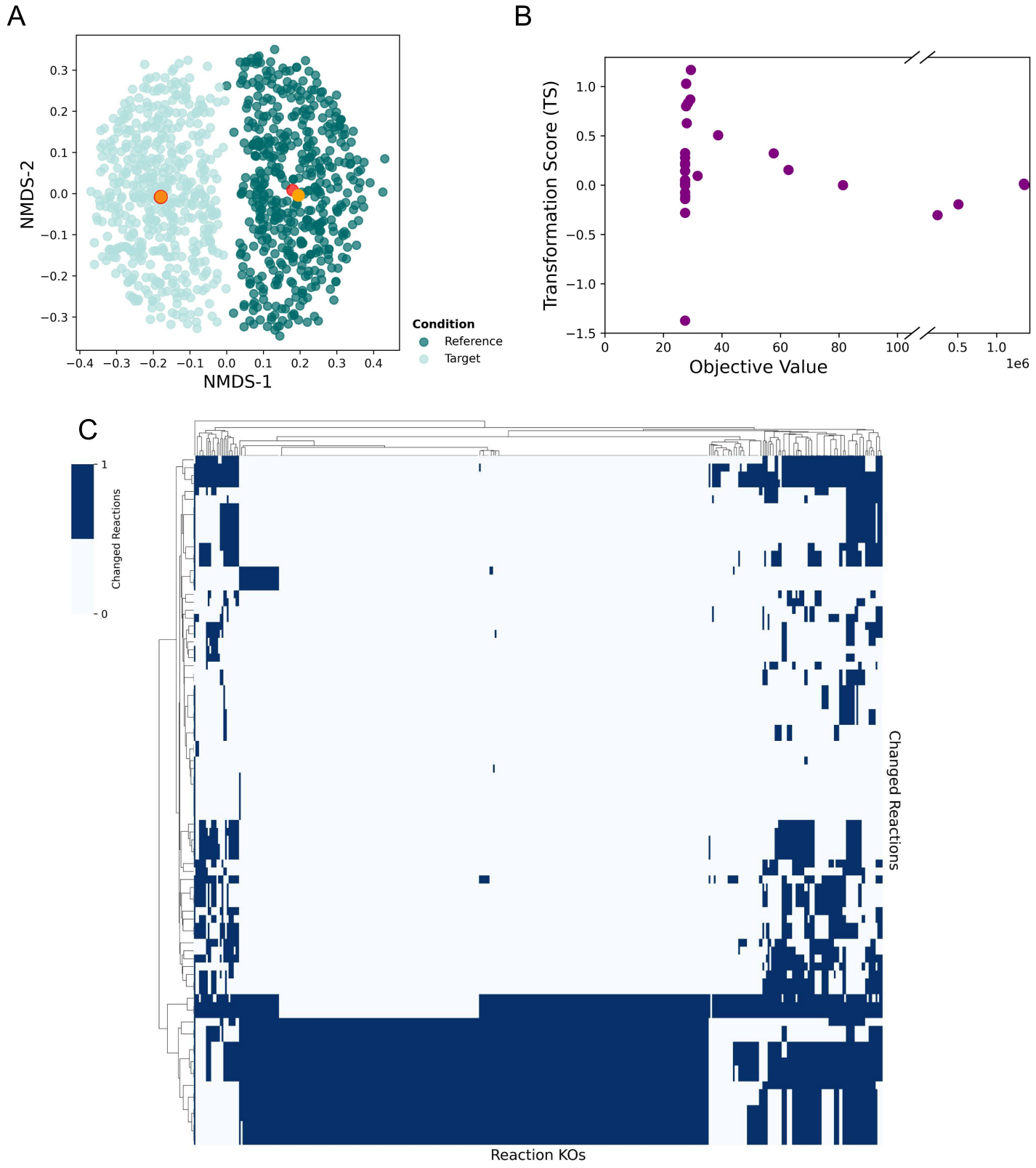
Supplemental Figure 2.3. Random Forest validation metrics. (A) Visualization of the random stratified group k-fold splits used for cross validation of the Random Forest classifier. (B-C) K-fold cross validation (k=5) of the Random Forest classifier testing ROC (B) and accuracy (C), with an average accuracy of 95% in cross validation. (D) Confusion matrix for model predictions with train and test sets selected in a 75-25 ratio using random stratified group splits. The model trained on this set had a 97% accuracy. (E) The top 20 features for model predictions by Gini score.



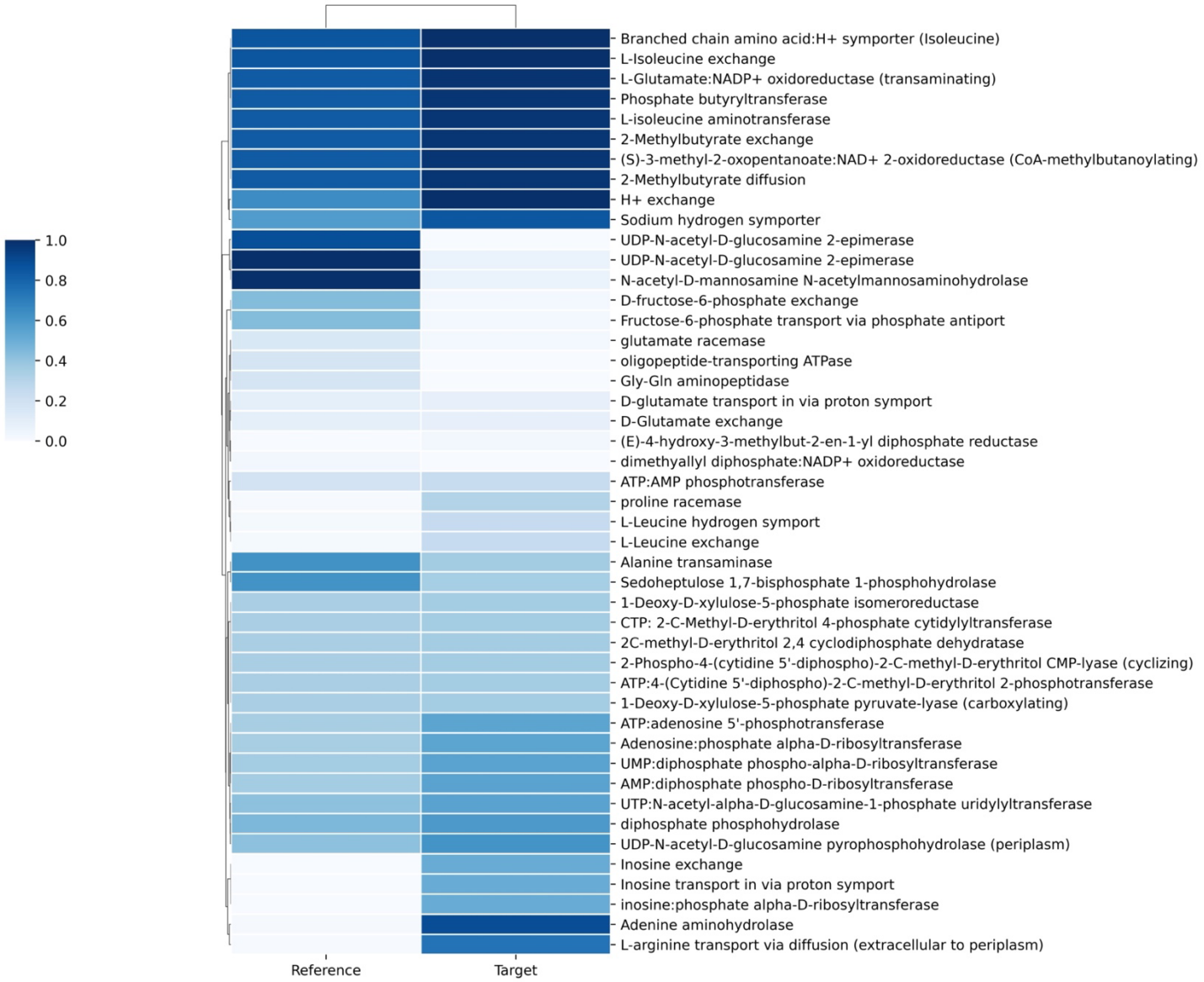
Supplemental Figure 2.4. Escher metabolic maps. Metabolic context for reactions from the Random Forest analysis labeled with the reaction and model IDs from the GENREs iCdG709 and iCdR703.



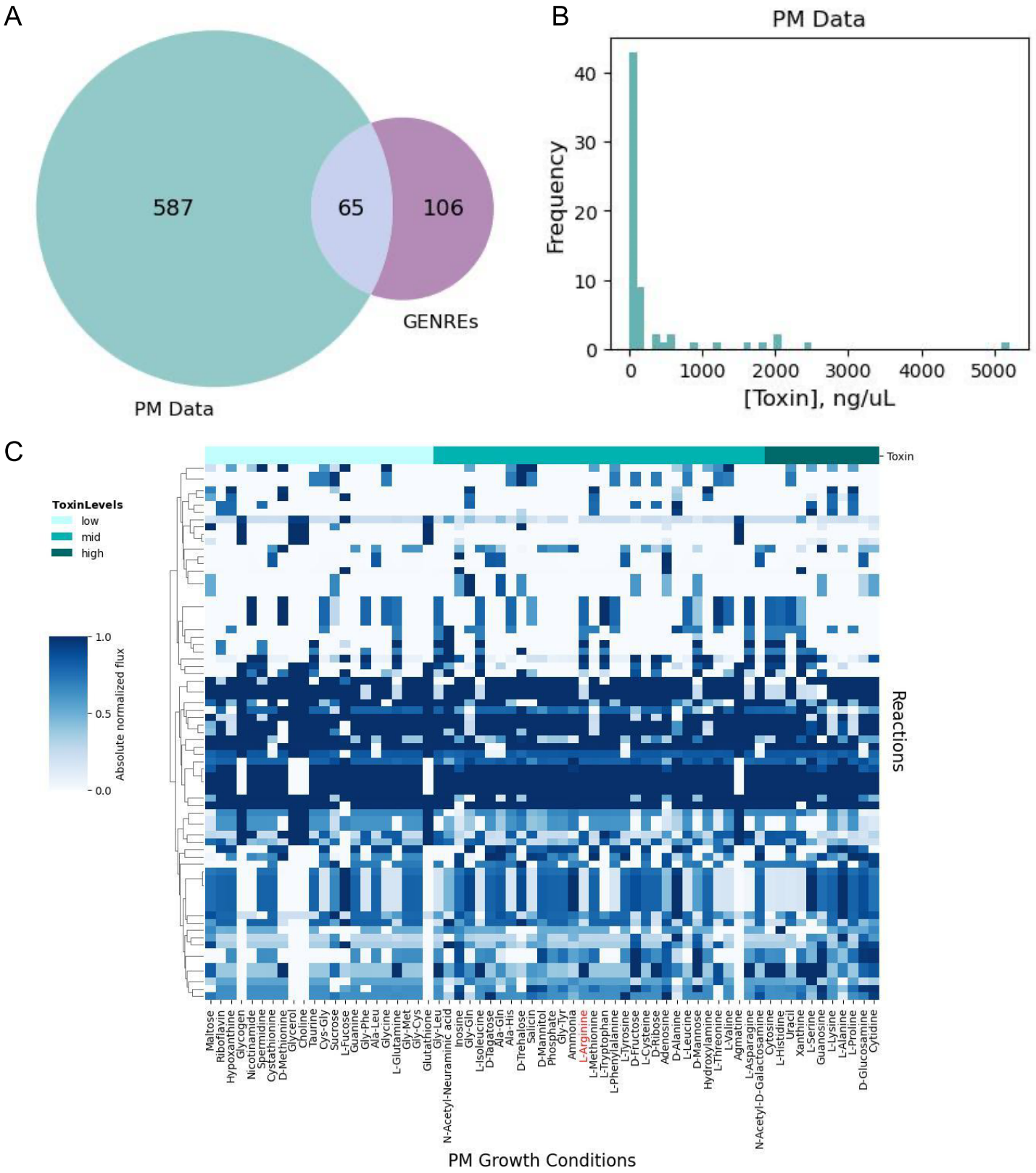
Supplemental Figure 2.5. Shadow prices of metabolites that decrease flux through reactions from Random Forest. For each objective function (OF) listed in Figure 2.3A, the metabolites categorized as decreasing and with a shadow price < -5 are shown.



Supplemental Figure 2.6. MTA calculations for centroids and transformation scores (TS). (A) Bray-Curtis NMDS of flux sampling results for iCdG709 contextualized for BHIS + DCA 240 uM (target, low toxin, light teal) and BHIS (reference, high toxin, dark teal) was used to calculate the centroids (red) and the flux sample closest to the centroid (orange) for each model. (B) The MIQP objective value verses the TS calculated using equation 10 demonstrates the utility of the TS in ranking flux solutions with a similar objective-value based on success of the flux solution in transforming reactions to the target state. (C) Successfully changed reactions for each reaction knockout. Successful (dark blue), unsuccessful (light blue).



Supplemental Figure 2.7. Comparison of metabolic flux through reactions in the Reference and Target state.



Supplemental Figure 2.8. Phenotype microarray (PM) simulation and analysis. (A) Venn diagram showing the overlap of unique metabolites from the PM dataset and the extracellular metabolites from the GENREs. (B) The toxin concentration distribution for the 65 overlapping growth conditions from panel (A). (C) Simulated reaction flux through each *in silico* PM condition (n=65). The flux data was min-max normalized and reactions with flux variance across all conditions < 0.05 were removed and the absolute flux value of the remaining reactions was visualized. The PM growth conditions are sorted by their toxin category. Toxin categories were defined as low (<42 ng/uL), mid (42-420 ng/uL), and high (>420 ng/uL) as in Lei, XH and Bochner, BR (2013).

Chapter 3: Nutritional, commensal, and genetic drivers of metabolic phenotypes in *C. difficile*

Authors
Deborah A. Powers[1], Matthew Jenior[2], Jason A. Papin[1,2]

1. Biochemistry and Molecular Genetics, School of Medicine, University of Virginia, Charlottesville, Virginia, United States of America
2. Biomedical Engineering, School of Engineering, University of Virginia, Charlottesville, Virginia, United States of America

Abstract
The primary risk factor for *C. difficile* infection (CDI) is broad-spectrum antibiotic use which clears the intestinal microbiome leaving an open niche for *C. difficile* to colonize. During infection, *C. difficile* metabolism is highly responsive to its nutritional environment. For example, in mice with CDI, a high protein diet exacerbates disease severity while a high carbohydrate, high fiber diet ameliorates disease. However, *C. difficile* is a genetically and metabolically diverse species with often significant differences between strains. To investigate how *C. difficile* metabolism differs between strains as a function of diet, we reconstructed genome-scale metabolic models of 105 *C. difficile* strains from publicly available, whole-genome sequencing data and then simulated metabolic flux for each model under three unique diet conditions. We found that strains cluster in two metabolically distinct groups by diet; these groups can be differentiated from each other by flux through alternate energy pathways and redox reactions. We also simulated growth of the *C. difficile* strains in the presence of 8 different commensals to understand the impact of diet on common microbiome interactions in the large intestine. We found that *C. difficile*-commensal interactions were more impacted by diet than differences in *C. difficile* strain. Overall, the diverse metabolic responses of both *C. difficile* strains and commensals to different diet conditions has therapeutic implications, particularly for fecal microbiota transplants where differences in metabolism could impact the effectiveness of treatment and as an important parameter to incorporate in the design of therapeutic synthetic bacterial communities for CDI.

Introduction

C. difficile infection (CDI) is a gastrointestinal infection caused by broad-spectrum antibiotic use which clears the intestinal microbiome leaving an open niche for *C. difficile* to colonize[1,2]. The standard treatment for CDI is broad-spectrum antibiotics, such as vancomycin or metronidazole[3]. These antibiotics are effective in resolving CDI initially; however, because they are broad-spectrum antibiotics, they continue clearing the intestinal microbiome of its normal flora. In approximately 30% of patients the loss of a healthy microbiome leads to *C. difficile* recolonizing the intestine, resulting in recurrent CDI[4]. Patients with recurrent CDI are treated with a fecal microbiota transplant (FMT) which effectively resolves disease in 90% of cases[5,6]. FMT is effective because it restores the healthy state of the microbiome and provides colonization resistance to *C. difficile*[5,7,8]. However, inherent features of FMTs prevent it from being the standard of care in CDI, such as the invasive procedure, the risk of pathogen transfer from the donor to the recipient, and scalability[8,9]. To overcome the limitations of FMTs, significant research efforts are being made to engineer synthetic microbial therapeutics[10–15]. These efforts must address the complex interactions between *C. difficile*, the microbiome, the host, and the nutritional environment.

Diet affects CDI outcomes. Mouse models of CDI given low protein diets had decreased *C. difficile* growth and decreased toxin production[16]. Similarly, mice with CDI given a high protein diet had increased mortality[17].
A high protein diet increases the availability of amino acids, a primary nutritional resource for energy generation in *C. difficile* via Stickland fermentation[18,19]. Additionally, mice given either high fat/high protein or high fat/low fiber diets had increased mortality[20,21]. High fat diets elevate primary bile acids (BAs) to digest fat; however, primary BAs also promote *C. difficile* spore germination[1,22]. Conversely, secondary BAs are decreased in high fat diets which reflects low microbiome diversity; secondary BAs have also been shown to protect against CDI[1,22,23]. Diet can also improve CDI outcomes. High carbohydrate and high fiber diets in mice protected against CDI[20,24,25]. This protective effect was non-specific and independent of the carbohydrate composition used [24,25]. Complex and simple carbohydrates such as inulin or pectin are known as microbiota-accessible carbohydrates (MACs)[26,27]. MACs expand microbial diversity by providing easily metabolized carbohydrates and protect the epithelial barrier through SCFAs, a downstream byproduct of MACs[26–28]. The nutritional landscape of the large intestine modulates *C. difficile* through a multitude of biological systems and metabolic pathways. However, the role of diet on different *C. difficile* strains and on the efficacy of bacterial therapeutics in the treatment of CDI remains unknown.

C. difficile has a dynamic, mosaic genome. Its genome is approximately 4.1-4.3 Mb and contains a high proportion of mobile genetic elements[29–31]. *C. difficile* has an ultra-low level of genome conservation; the core genome may be as low as 16% conserved[31]. Furthermore, there are large phylogenetic distances between strains, which makes species classification difficult[29]. *C. difficile* strain relatedness is determined in several ways including whole-genome sequencing, multi-locus sequence typing, and sequencing single regions within the genome (such as PCR ribotyping and toxinotyping). Several pan-genome analyses have been performed to identify the relationship between the genetic and phenotypic profiles of *C. difficile* strains[32–35]. These studies have found that carbohydrate metabolism in particular is highly variable across the pangenome and that metabolic variability does not align with any current strain typing systems[32,36]. However, it is unknown how these genotypic and phenotypic differences influence responses to diet or how they might impact microbiome dynamics.

To investigate the effect of diet on *C. difficile* strains and on *C. difficile*-commensal interactions, we conducted a systematic analysis using genome-scale metabolic models of 105 *C. difficile* strains. We simulated metabolism in these models under three diet conditions. We found that *C. difficile* strains cluster into two groups based on flux through energy-generating pathways, specifically the electron bifurcating ferredoxin reduction pathway. Furthermore, we examined whether diet influenced metabolic interactions between *C. difficile* and gut commensals. We found that metabolic interactions between *C. difficile* and commensals increased as the diet became more restrictive. Additionally, in *C. difficile*-commensal interaction simulations that improved *C. difficile* growth, *C. difficile* had increased uptake of metabolites associated with Stickland fermentation. These results uncover important metabolic phenotypes to consider when designing a bacterial therapeutic that will have robust efficacy regardless of the *C. difficile* strain causing disease or the diet of the host.

Results

Workflow validation

C. difficile is metabolically versatile. In mouse models of CDI, hosts worsen under high protein, low carbohydrate diet conditions and improve under low protein, high carbohydrate diet conditions[16,17,20,21,24,25]. However, the underlying metabolic and ecological principles driving these responses are still undefined. To better understand the metabolism of *C. difficile* in response to different diets, we designed a metabolic modeling simulation to systematically investigate the range of possible metabolic responses to diet as a function of strain. We defined two *in silico* media based on mouse diet formulas[20]: high-protein, low-carbohydrate (HPLC) and low-protein, high-carbohydrate (LPHC). We additionally defined a rich media control based on brain heart infusion (BHI) media. Our standard workflow is to reconstruct metabolic models of *C. difficile* strains, gapfill to BHI, constrain to the 3 media (BHI, HPLC, LPHC), sample the flux distributions, and finally analyze the results (Fig S3.1A).

To test if the computational modeling approach could capture known metabolic differences between *C. difficile* strains, we built genome-scale metabolic models (GEMs) for three *C. difficile* strains (DSM 102860, DSM 28666, and DSM 29629) which have unique fermentation profiles[36]. A primary component analysis (PCA) of the metabolic flux distributions (n=50) for each strain model under each *in silico* diet condition separates DSM 28666 from the other two *C. difficile* strains (Fig S3.1B); this separation aligns with the previously reported metabolic differences between these *C. difficile* strains based on hierarchical clustering of fermentation profiles. To investigate the metabolism driving this separation between strains, we applied a Random Forest classifier. Flux through the 15 reactions with the highest Gini Index shows increased activity in carbohydrate metabolism (glucose and sucrose transport reactions) and threonine transport in DSM 28666 relative to the other two strains (Fig S3.1C). This threonine activity in DSM 28666 corresponds with the increased 2-aminobutanoate and 2-hydroxybutanoate (threonine oxidative Stickland fermentation products) shown in Riedel, T. 2017[36]. Our approach successfully recapitulated overall separation by strain in addition to identifying strain-specific metabolism driving these differences giving us confidence to apply it to our primary question.

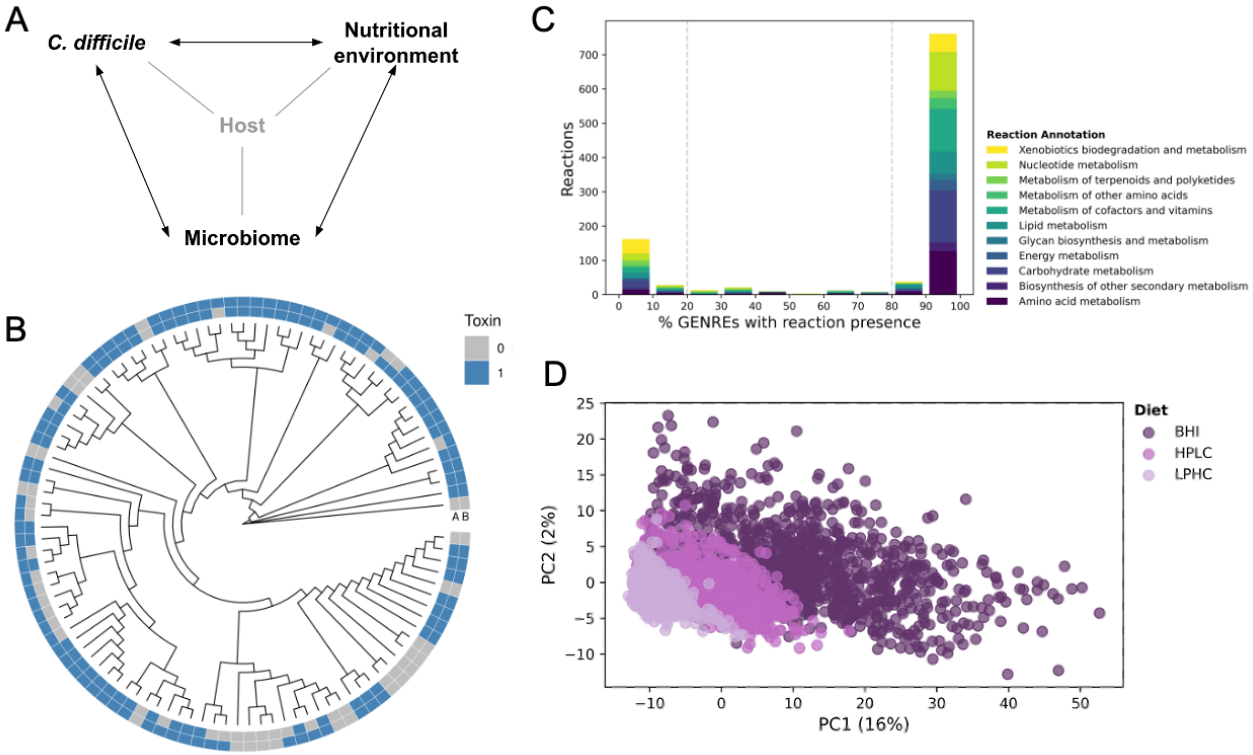


Figure 3.1. *C. difficile* strain library. (A) Diagram of the variables and interactions that occur during CDI. **(B)** A phylogenetic tree of the 105 C. *difficile* strains used in this analysis. The presence or absence of the TcdA and TcdB proteins in each strain is indicated by blue or grey, respectively. **(C)** KEGG annotations for metabolic reactions are binned by inclusion frequency in the 105 *C. difficile* GEMs. Reactions are classified as unique, accessory, or core based on the percent of GENREs with the reaction: 1-20%, 20-80%, 80-100% respectively. **(D)** PCA of flux distributions (n=10) from 105 *C. difficile* strain GENREs. The standard deviation from the cluster centroid was calculated by diet for BHI (13.27), HPLC (6.22), and LPHC (2.93). The median distance between centroids significantly differed for all diet conditions by Kruskal-Wallis test statistic (p = 2.58e-307).

Metabolic variation between *C. difficile* strains in response to diet

To explore the range of possible metabolic responses to diet as a function of strain, we created a panel of 105 *C. difficile* strains using whole-genome sequences from the BV-BRC, covering a range of phylogenetic relatedness (Fig 3.1B). The metabolic models reflect this phylogeny: across the models, 57% of the metabolic reactions are conserved and 13% are unique (Fig 3.1C). We constrained each *C. difficile* model to each media condition (BHI, HPLC, LPHC) and sampled the flux distribution space. PCA of the flux samples (n=10) from the 105 *C. difficile* strain models in three conditions shows strong clustering by diet (Fig 3.1D). Nutritionally, the diets progress from a rich to restrictive formula from BHI to HPLC to LPHC, with LPHC being the most restricted diet. Metabolically, the standard deviation within diet clusters decreases with increasing nutritional restriction, indicating a reliance on core metabolic functions.

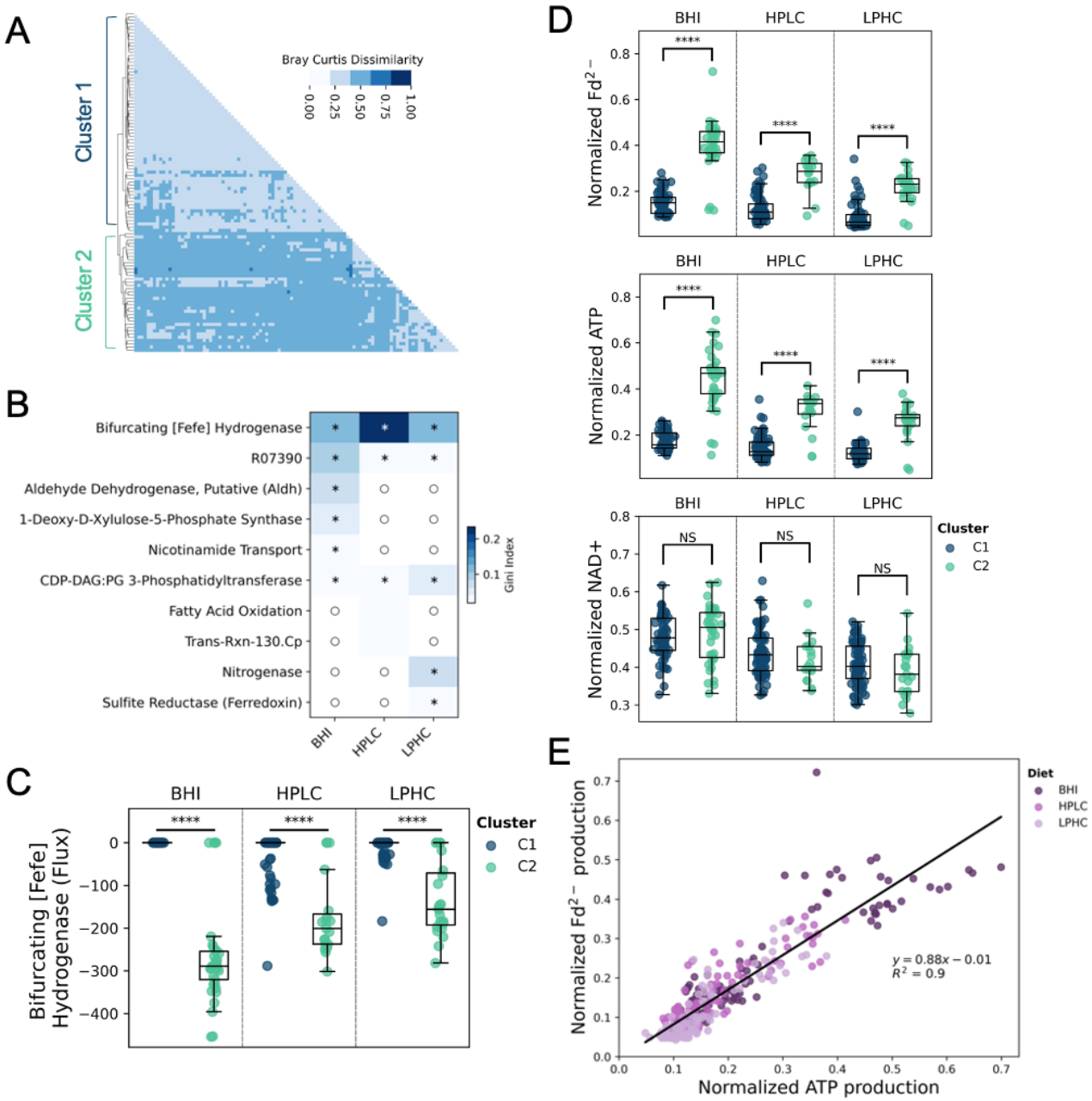


Figure 3.2. C. difficile strains show differences in energy metabolism. (A) Pair-wise Bray-Curtis dissimilarity index for flux samples from 105 C. difficile strains constrained to BHI divides strains into similar (cluster 1) and dissimilar (cluster 2) groups. **(B)** The top five Random Forest features by Gini index for classifying models as cluster 1 or 2 are shown by diet. Asterisk indicates that a significant difference in reaction flux between clusters 1 and 2 by Mann-Whitney U test; circle indicates that the reaction was not included in the top 20 features in the Random Forest analysis. **(C)** Distribution of the median flux by strain through bifurcating [FeFe] hydrogenase for each cluster and diet. **(D)** Normalized sum of flux through all producing reactions for reduced ferredoxin (Fd^{2-}), ATP, and NAD^{+}; significance

tested using Mann-Whitney U test. **(E)** Normalized flux production of ATP vs Fd^{2-} has a Pearson's correlation of 0.9. BHI: brain heart infusion; HPLC: high protein, low carbohydrate; LPHC: low protein, high carbohydrate.

To investigate how strain differences contribute to metabolic responses, we calculated pair-wise Bray-Curtis dissimilarity indices using flux samples from the *C. difficile* strain panel under each diet condition. Hierarchical clustering of the indices separates the strains into two clusters that are either similar (Bray-Curtis index < 0.4) or dissimilar (Bray-Curtis index > 0.4) (Fig 3.2A, Fig S3.2). It is interesting to note that in all diets, Cluster 2 is not only dissimilar to Cluster 1, but also dissimilar to itself (Fig 3.2A, Fig S3.2). Furthermore, the metabolic phenotypes of *C. difficile* strains were most similar in HPLC (Fig S3.2A). This convergence on a similar metabolic phenotype in HPLC is likely due to increased reliance on highly conserved Stickland fermentation pathways in a restrictive but high protein condition.

Next, we ran a Random Forest classifier for each diet condition to identify the metabolic features driving the differences between the clusters. We took the reactions with the top five Gini indices from each diet and tested whether the flux for those reactions was significantly different between clusters using a Mann-Whitney U test (Fig 3.2B). Bifurcating [FeFe] hydrogenase was the most important feature in differentiating between clusters in all diet conditions and the median flux by strain through this reaction was significantly more active in Cluster 2 (Fig 3.2B-C). The bifurcating [FeFe] hydrogenase belongs to a family of enzymes that release energy from H_2 through redox reactions and is particularly important in anaerobic bacteria[37].

Because the bifurcating [FeFe] hydrogenase reaction was flagged as important in the Random Forest analysis, we decided to measure the metabolite-producing flux for energy-related metabolites. Briefly, the flux through all reactions that produce the metabolite of interest is multiplied by the relevant coefficient factor then summed and normalized across strains. This calculation provides us with a total production value for a metabolite that can be compared across models. From this calculation, we see that the production of reduced ferredoxin (Fd^{2-}) and ATP is significantly increased in Cluster 2 while NAD^+ is not significantly different between the two clusters (Fig 3.2D). Additionally, the Fd^{2-} and ATP are positively correlated with a Pearson's correlation of 0.9 (Fig 3.2E). Bifurcating [FeFe] hydrogenase reaction uses ferredoxin as an electron acceptor[18,37,38]. The reduced Fd2- is then able to donate these electrons to the electron transport chain to generate ATP[18,37–39]. This simulation data suggests that this pathway for energy generation is more active in Cluster 2, potentially providing a metabolic advantage under different diet conditions.

Community cross-feeding dynamics change based on diet

To investigate pairwise interactions between *C. difficile* strains and gut commensals, we used the Metabolic CrossTalk (MetCT) algorithm[40]. MetCT can be used to simulate either the mutualistic or competitive potential of a given pair of models; it provides three types of data for each pairwise simulation: the overall change in biomass for each model when grown together, a competitive and/or cooperative index, and an exchange flux vector for each model. Using MetCT, we ran a pairwise mutualism simulation between each *C. difficile* strain model and eight gut commensals. For each commensal, we averaged the final biomass of the *C. difficile* models to determine if growth with a commensal improved or impaired *C. difficile* biomass (Fig 3.3A). Under BHI conditions, *C. difficile* growth does not change regardless of the commensal pair. However, in HPLC, growth with *S. thermophilus* and *E. coli* significantly increases *C. difficile* biomass. These commensals as well as *B. producta* also increase growth in LPHC conditions.

To determine which metabolic features drive changes in C. difficile biomass with specific commensal partners under specific diet conditions, we analyzed the exchange fluxes from the pairwise mutualism simulations between the *C. difficile* models and the commensals (Fig 3.3B). Here we see that an increase in *C. difficile* biomass is accompanied by an increase in flux through Stickland fermentation metabolites aspartate and glutamate as well as increased flux through water and nitrite/nitrate exchanges. This suggests a beneficial metabolic interaction allowing *C. difficile* to boost energy generation via Stickland fermentation.

To experimentally test these *C. difficile*-commensal interactions, we grew *C. difficile* in co-culture with *E. coli*, *B. producta*, and *S. thermophilus* using Cerillo duet plates (Fig 3.3C, Fig S3.3). The Cerillo duets feature a semi-permeable membrane that allows metabolite exchange between compartments but not bacterial migration. This system allows us to measure the individual growth of two species with the same nutritional resources. We used three media conditions: BHI, P40, and P10; P40 and P10 are *in vitro* formulations of the *in silico* HPLC and

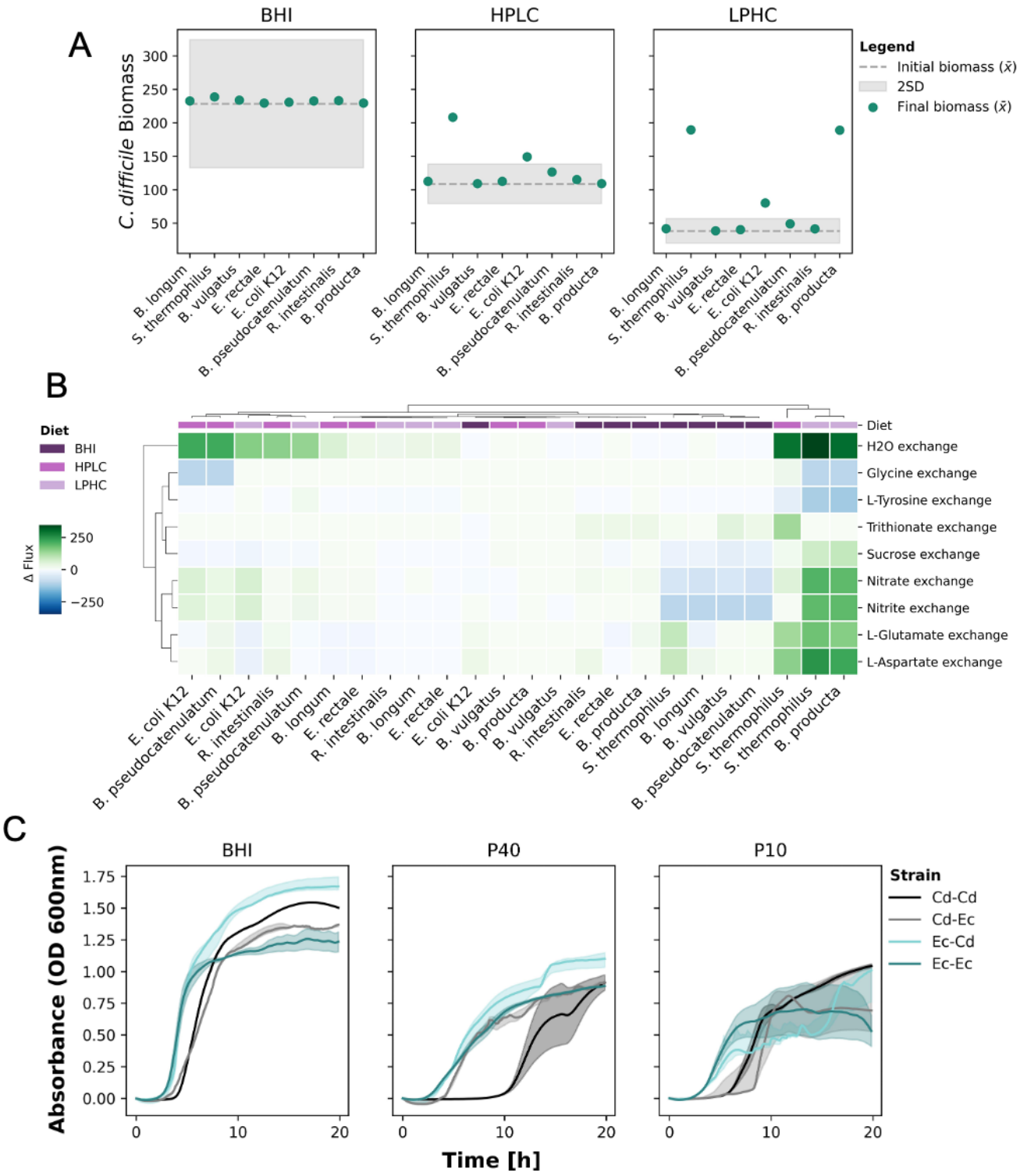


Figure 3.3. C. difficile-commensal metabolic interactions shift based on the nutritional background. (A) The average biomass ± 2 standard deviations (SD) of all the C. difficile models (n=105) are depicted as the grey dashed line and the grey box. The average biomass of all the C. difficile models when placed in a mutualistic simulation with a commensal is shown in green. **(B)** The median change in flux through exchange reactions from the pairwise mutualism simulations between the C. difficile models and the commensals. **(C)** Median optical density (600 nm) and IQR of C. difficile growth with itself (n=3, black) or with E. coli (n=3, grey) and of E. coli growth with itself (n=3, dark teal) or with C. difficile (n=3, light teal) in three media conditions (BHI, P40, and P10).

LPHC, respectively. *C. difficile* growth in P40 was significantly greater in co-cultures with *E. coli* (n=3) compared to the control (*C. difficile* grown with itself) (Fig 3.3C). Similarly, *E. coli* growth also increased in P40 co-culture with *C. difficile* compared to growth with itself. However, in P10 medium, *E. coli* and *C. difficile* co-culture did not

provide a growth benefit to either species. These *in vitro* results partially validate the *in-silico* predictions from the MetCT analysis (Fig 3.3A). Co-culture of *C. difficile* and *B. producta* showed a slight growth benefit to *C. difficile* in P10 but not in P40, which matches the MetCT predictions (Fig S3.3A). Finally, *C. difficile* had a growth benefit in co-culture with *S. thermophilus* in all media conditions (Fig S3.3B). However, in the case of *S. thermophilus* co- culture, large differences in the lag phase between *C. difficile* and *S. thermophilus* may be a confounding variable. To control for differences in growth dynamics, we did an additional spent-media experiment with *S. thermophilus*. Briefly, *S. thermophilus* was grown in the three media conditions for 24 hours; *C. difficile* growth was then measured on the filter-sterilized spent media (Fig S3.3C). We found that *C. difficile* also demonstrates a growth benefit in *S. thermophilus* spent media compared to controls (*C. difficile* grown in *C. difficile* spent media).

Discussion

Using an ensemble of 105 *C. difficile* strains, we found that *C. difficile* strains could be grouped into two clusters based on simulated flux production for ferredoxin and ATP. Furthermore, ferredoxin and ATP production are positively correlated and increase with increasing nutritional resources. We also found that *C. difficile* biomass increased in simulations with *S. thermophilus*, *E. coli*, and *B. producta*. In these simulations, *C. difficile* flux through glutamate and aspartate exchanges are increased as well as nitrite and nitrates. Glutamate and aspartate are important precursors in Stickland fermentation. Additionally, as diet became more restrictive, the interactions between the commensals and *C. difficile* became more pronounced indicating a higher degree of cross-feeding.

Our goal was to investigate the effect of diet on *C. difficile* strains to see if wide genetic variability resulted in phenotypic variability. We modified defined mouse diets to create both *in silico* and *in vitro* growth media for testing. These three defined diets allowed us to precisely modulate the intake of a model/organism and measure the resulting metabolic phenotype. However, these diets only begin to scratch the surface of the range of possible outcomes as the human diet is considerably more varied and imprecise. Future work including both a greater variety of diets and more complex components, such as dietary fats and MACs, will be an important next step.

C. difficile growth increased in simulations with *S. thermophilus*, possibly due to resource sharing of Stickland fermentation precursors (Fig 3). *In vitro* experiments of *C. difficile* grown in spent media from *S. thermophilus* validated this growth benefit in C. difficile when the spent media was pH adjusted. *S. thermophilus* is one of the members of a therapeutic microbial consortia that successfully resolved CDI in mice[40]. *S. thermophilus* creates an acidic environment through the production of lactic acid, leading to unfavorable growth conditions for *C. difficile.*

Importantly, these results showed that diet impacts *C. difficile* metabolism and interactions with the microbiome, more than strain differences. Therefore, the ecological principles that modulate *C. difficile* through diet and the surrounding microbiome could be manipulated to resolve CDI regardless of the infecting strain. Future research efforts should continue to investigate the efficacy of different bacterial consortia under a variety of diet conditions. The premise that ecological controls can be applied to resolve CDI is an exciting frontier for infectious diseases in general. So long as the ecological principles of control for a disease are understood, there could be innumerable bacterial combinations that replicate the necessary ecological and metabolic functions. Finally, ecological control of infection provides a timely alternative to antibiotic treatments and the growing problem of antibiotic resistance in pathogens.

Methods

Metabolic model reconstruction: Whole-genome sequences of 105 *C. difficile* strains were selected from the BV-BRC database by filtering for sequences that were “complete” and “good”. These sequences were used as the input for the Reconstructor algorithm to build genome-scale metabolic models[41]. All the models were gap-filled to a rich, *in-silico* BHI media using Reconstructor. The same reconstruction method was applied to 8 commensals: *B. producta*, *B. vulgatus*, *B. pseudocatenulatum*, *E. rectale*, *B. longum*, *E. coli K-12*, *R. intestinalis*, and *S. thermophilus*. Additionally, whole-genome sequences for 3 *C. difficile* strains profiled in Riedel, T., *et al*., 2017 (DSM 102860, DSM 28666, and DSM 29629) were selected from the BV-BRC database and used to reconstruct metabolic models for the purpose of validating the computational approach.

Model analyses: We simulated flux under three diet conditions: BHI; high protein, low carbohydrate (HPLC); and low protein, high carbohydrate (LPHC). For BHI, we set the lower bounds of the exchange reactions for metabolites in the *in-silico* BHI media to -1000 and the lower bounds of the remaining exchanges to 0. For HPLC and LPHC, the bounds for the BHI exchanges were adjusted to represent proportions from published mouse diets for high and low protein diets[20]. We generated flux samples (n=500) for each *C. difficile* model (n=105) under each diet condition (n=3) using Gapsplit[42].

Flux samples were randomly down-sampled (n=50) for each diet condition for PCA. The centroid and standard deviation was calculated for each diet cluster from the PCA results; a Kruskal-Wallis test was used to determine if the groups were significantly different. Pairwise Bray-Curtis dissimilarities between *C. difficile* strains were calculated using the median flux values from the flux samples. These Bray-Curtis dissimilarities were clustered hierarchically, and a Random Forest classifier was applied to the two primary clusters to determine which metabolic features differentiated between them. The flux samples were down-sampled (n=50) before applying the Random Forest classifier.

To calculate metabolite flux production, all reactions producing or consuming a metabolite of interest were listed. Two vectors were then extracted from this reaction list: the metabolite coefficient vector and the reaction flux vector. These vectors were then multiplied to determine the production and consumption of the metabolite of interest for each reaction. A positive sign indicates metabolite production, and a negative sign indicates metabolite consumption. Metabolite production was then summed across all the associated reactions and normalized across the models by diet.

Metabolic Cross Talk (MetCT) algorithm: MetCT simulates metabolic interactions between bacterial GEMs using two main functions: mutualism and competition. In both functions, MetCT sets a minimum growth threshold and calculates the minimal media required to meet this threshold for each GEM. MetCT then takes the union of the minimal media from each GEM to create a shared media. The mutualism function measures the metabolite production of each model using flux variability analysis (FVA). If a metabolite is produced by the first GEM, then the lower bound of the corresponding exchange reaction in the second GEM is decreased by a step. These steps are repeated until there are no new additions to the shared media. The competitive function measures metabolite consumption using FVA. If both GEMs consume a metabolite, then the lower bound of the corresponding exchange reaction is increased in both GEMs by a pre-set fraction. These steps are repeated until one or both GEMs cannot grow. For both the mutualism and competition simulations, MetCT returns three levels of data: overall change in biomass for each model, a mutualism or competition index, and a data frame of the metabolites produced or consumed in the simulation.

Growth experiments: Three *C. difficile* strains (R20291, Cd10, and Cd4) and three commensal strains (*Escherichia coli K-12*, *Bacteroides producta*, and *Streptococcus thermophilus*) were used in these experiments. All strains were grown anaerobically overnight at 37C in BHIS (brain heart infusion supplemented with 5% yeast extract). Overnight cultures were then diluted 1:2 in fresh BHIS and incubated for 2 hours. The 2-hour cultures were then inoculated to the three media: BHIS, P10 (10 mg/mL casaminos, 5 mg/mL sucrose, 5 mg/mL glucose, 1% cysteine, 1% tryptophan, 1% ATCC vitamins, 1% ATCC minerals) and P40 (40 mg/mL casaminos, 0.2 mg/mL sucrose, 0.2 mg/mL glucose, 1% cysteine, 1% tryptophan, 1% ATCC vitamins, 1% ATCC minerals). Growth was measured at OD 600 every 10 minutes for 20-24 hours using a Cerillo Alto plate reader while being shaken continuously at 250 rpm. Statistical differences between curves was evaluated using area under the curve (AUC).

Data Availability Statement

Publicly available whole-genome sequencing data from the BV-BRC was used to reconstruct all the metabolic models in this study; the genome IDs are listed in Supplemental File 1. The code used for modeling simulations and all analyses is available on GitHub: https://github.com/dap5mb/CdInteractions.

Acknowledgements

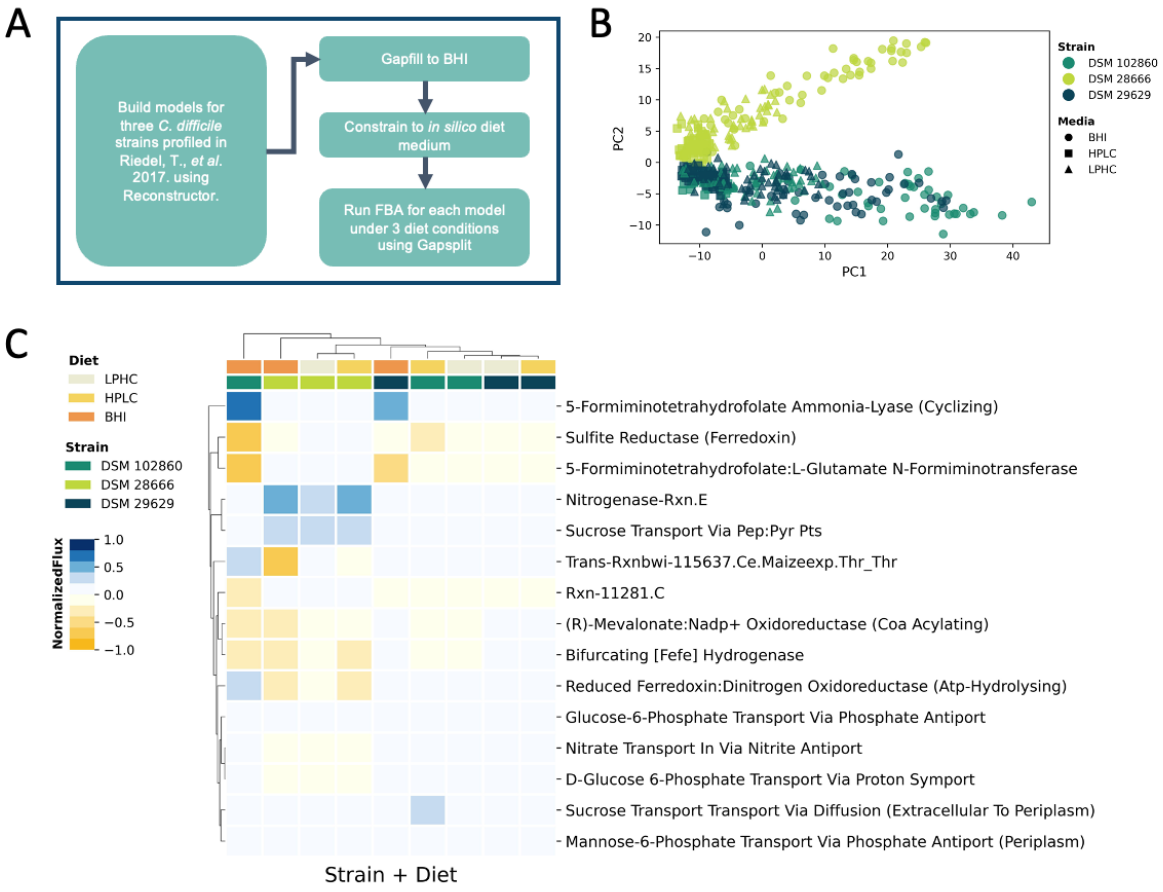
The *C. difficile* strains Cd4 and Cd10 were kindly provided by Joseph Zackular, PhD at the University of Pennsylvania.

References

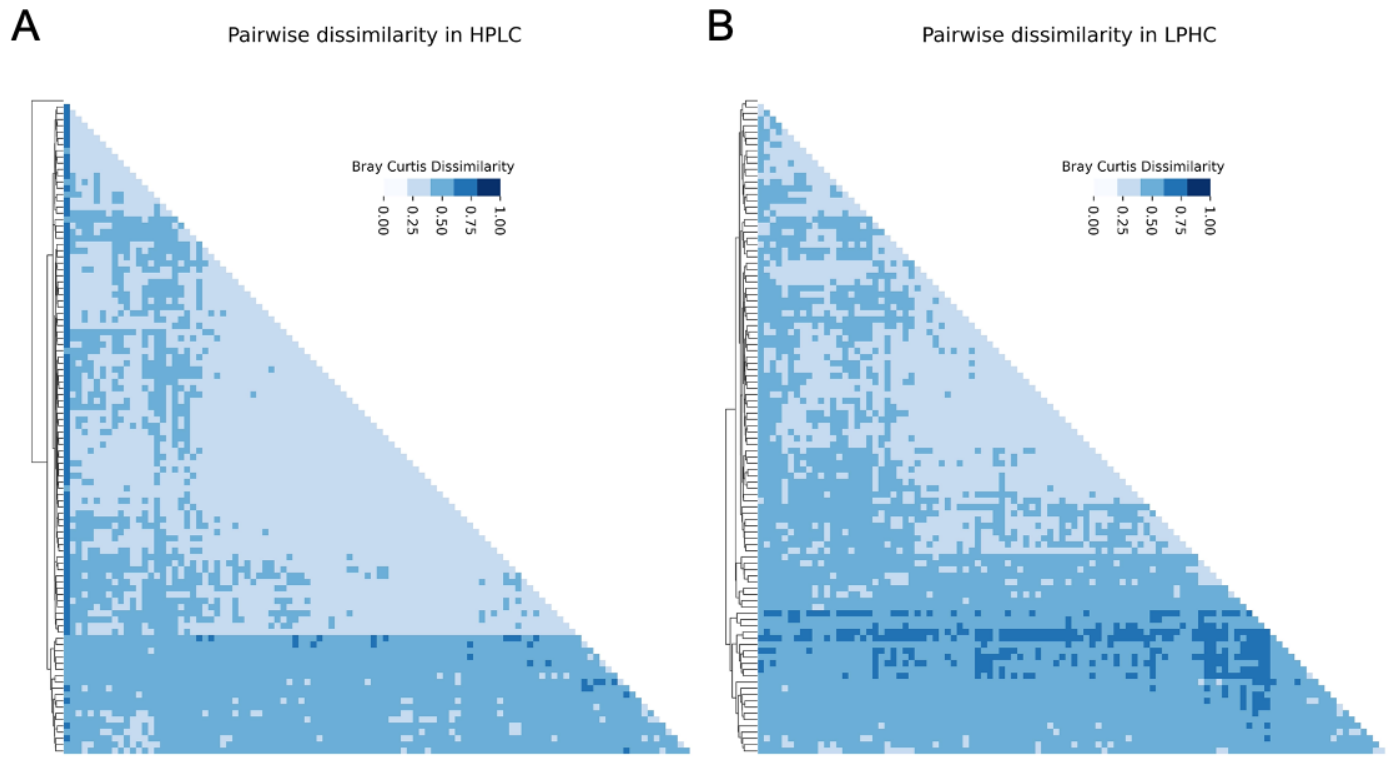
1. Theriot, C. M. & Young, V. B. Interactions Between the Gastrointestinal Microbiome and Clostridium difficile. *Annu Rev Microbiol* **69**, 445–461 (2015).
2. Sehgal, K. & Khanna, S. Gut microbiome and *Clostridioides difficile* infection: a closer look at the microscopic interface. *Therap Adv Gastroenterol* **14**, 175628482199473 (2021).
3. Al-Jashaami, L. S. & DuPont, H. L. Management of Clostridium difficile Infection. *Gastroenterol Hepatol (N Y)* **12**, 609–616 (2016).
4. Song, J. H. & Kim, Y. S. Recurrent Clostridium difficile Infection: Risk Factors, Treatment, and Prevention. *Gut Liver* **13**, 16–24 (2019).
5. Baktash, A. *et al.* Mechanistic Insights in the Success of Fecal Microbiota Transplants for the Treatment of Clostridium difficile Infections. *Front Microbiol* **9**, 1242 (2018).
6. Baunwall, S. M. D. *et al.* Faecal microbiota transplantation for first or second Clostridioides difficile infection (EarlyFMT): a randomised, double-blind, placebo-controlled trial. *The Lancet Gastroenterology & Hepatology* **7**, 1083–1091 (2022).
7. Kamiya, S. Microbial ecology between Clostridioides difficile and gut microbiota. *Biosci Microbiota Food Health* **42**, 229–235 (2023).
8. Hitch, T. C. A. *et al.* Microbiome-based interventions to modulate gut ecology and the immune system. *Mucosal Immunology* **15**, 1095–1113 (2022).
9. Wang, S. *et al.* Systematic Review: Adverse Events of Fecal Microbiota Transplantation. *PLoS One* **11**, e0161174 (2016).
10. Carlucci, C. *et al.* Effects of defined gut microbial ecosystem components on virulence determinants of Clostridioides difficile. *Sci Rep* **9**, 885 (2019).
11. Staley, C. *et al.* Successful Resolution of Recurrent Clostridium difficile Infection using Freeze-Dried, Encapsulated Fecal Microbiota; Pragmatic Cohort Study. *Am J Gastroenterol* **112**, 940–947 (2017).
12. Khanna, S. *et al.* A Novel Microbiome Therapeutic Increases Gut Microbial Diversity and Prevents Recurrent *Clostridium difficile* Infection. *J Infect Dis.* **214**, 173–181 (2016).
13. Furuichi, M. *et al.* Commensal consortia decolonize Enterobacteriaceae via ecological control. *Nature* (2024) doi:10.1038/s41586-024-07960-6.
14. Lawley, T. D. *et al.* Targeted Restoration of the Intestinal Microbiota with a Simple, Defined Bacteriotherapy Resolves Relapsing Clostridium difficile Disease in Mice. *PLoS Pathog* **8**, e1002995 (2012).
15. Lesniak, N. A. *et al.* The Gut Bacterial Community Potentiates Clostridioides difficile Infection Severity. *mBio* **13**, e01183-22 (2022).
16. Moore, J. H. *et al.* Defined Nutrient Diets Alter Susceptibility to Clostridium difficile Associated Disease in a Murine Model. *PLoS ONE* **10**, e0131829 (2015).
17. Yakabe, K. *et al.* Dietary-protein sources modulate host susceptibility to Clostridioides difficile infection through the gut microbiota. *Cell Reports* **40**, 111332 (2022).
18. Pavao, A. *et al.* Reconsidering the in vivo functions of Clostridial Stickland amino acid fermentations. *Anaerobe* **76**, 102600 (2022).
19. Fletcher, J. R. *et al.* Clostridioides difficile exploits toxin-mediated inflammation to alter the host nutritional landscape and exclude competitors from the gut microbiota. *Nat Commun* **12**, 462 (2021).
20. Mefferd, C. C. *et al.* A High-Fat/High-Protein, Atkins-Type Diet Exacerbates Clostridioides (Clostridium) difficile Infection in Mice, whereas a High-Carbohydrate Diet Protects. *mSystems* **5**, e00765-19 (2020).
21. Hazleton, K. Z. *et al.* Dietary fat promotes antibiotic-induced Clostridioides difficile mortality in mice. *npj Biofilms Microbiomes* **8**, 15 (2022).
22. Ticho, A. L., Malhotra, P., Dudeja, P. K., Gill, R. K. & Alrefai, W. A. Intestinal Absorption of Bile Acids in Health and Disease. *Compr Physiol* **10**, 21–56 (2019).
23. Collins, S. L., Stine, J. G., Bisanz, J. E., Okafor, C. D. & Patterson, A. D. Bile acids and the gut microbiota: metabolic interactions and impacts on disease. *Nat Rev Microbiol* **21**, 236–247 (2023).
24. Hryckowian, A. J. *et al.* Microbiota-accessible carbohydrates suppress Clostridium difficile infection in a murine model. *Nat Microbiol* **3**, 662–669 (2018).
25. Wu, Z. *et al.* The impact of dietary fibers on Clostridioides difficile infection in a mouse model. *Front. Cell. Infect. Microbiol.* **12**, 1028267 (2022).
26. Makki, K., Deehan, E. C., Walter, J. & Bäckhed, F. The Impact of Dietary Fiber on Gut Microbiota in Host Health and Disease. *Cell Host & Microbe* **23**, 705–715 (2018).
27. Deehan, E. C. *et al.* Modulation of the Gastrointestinal Microbiome with Nondigestible Fermentable Carbohydrates To Improve Human Health. *Microbiol Spectr* **5**, 5.5.04 (2017).

28. Morrison, D. J. & Preston, T. Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism. *Gut Microbes* **7**, 189–200 (2016).
29. Knight, D. R. *et al.* Major genetic discontinuity and novel toxigenic species in Clostridioides difficile taxonomy. *Elife* **10**, e64325 (2021).
30. Knight, D. R., Elliott, B., Chang, B. J., Perkins, T. T. & Riley, T. V. Diversity and Evolution in the Genome of Clostridium difficile. *Clin Microbiol Rev* **28**, 721–741 (2015).
31. Scaria, J. *et al.* Analysis of Ultra Low Genome Conservation in Clostridium difficile. *PLoS ONE* **5**, e15147 (2010).
32. Norsigian, C. J. *et al.* Systems biology analysis of the Clostridioides difficile core-genome contextualizes microenvironmental evolutionary pressures leading to genotypic and phenotypic divergence. *npj Syst Biol Appl* **6**, 31 (2020).
33. Norsigian, C. J. *et al.* Systems biology approach to functionally assess the *Clostridioides difficile* pangenome reveals genetic diversity with discriminatory power. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2119396119 (2022).
34. Kulecka, M. *et al.* Higher genome variability within metabolism genes associates with recurrent Clostridium difficile infection. *BMC Microbiol* **21**, 36 (2021).
35. Scaria, J. *et al.* Comparative nutritional and chemical phenome of Clostridium difficile isolates determined using phenotype microarrays. *Int J Infect Dis* **27**, 20–25 (2014).
36. Riedel, T. *et al.* High metabolic versatility of different toxigenic and non-toxigenic Clostridioides difficile isolates. *International Journal of Medical Microbiology* **307**, 311–320 (2017).
37. Katsyv, A. *et al.* Molecular Basis of the Electron Bifurcation Mechanism in the [FeFe]-Hydrogenase Complex HydABC. *J. Am. Chem. Soc.* **145**, 5696–5709 (2023).
38. Buckel, W. & Thauer, R. K. Energy conservation via electron bifurcating ferredoxin reduction and proton/Na(+) translocating ferredoxin oxidation. *Biochim Biophys Acta* **1827**, 94–113 (2013).
39. Peters, J. W. *et al.* [FeFe]- and [NiFe]-hydrogenase diversity, mechanism, and maturation. *Biochim Biophys Acta* **1853**, 1350–1369 (2015).
40. Jenior, M. L. *et al.* Systems-ecology designed bacterial consortium protects from severe Clostridioides difficile infection. *Under Review* (2022).
41. Jenior, M. L., Glass, E. M. & Papin, J. A. *Reconstructor: A COBRApy Compatible Tool for Automated Genome-Scale Metabolic Network Reconstruction with Parsimonious Flux-Based Gap-Filling*. http://biorxiv.org/lookup/doi/10.1101/2022.09.17.508371 (2022) doi:10.1101/2022.09.17.508371.
42. Keaty, T. C. & Jensen, P. A. Gapsplit: efficient random sampling for non-convex constraint-based models. *Bioinformatics* **36**, 2623–2625 (2020).

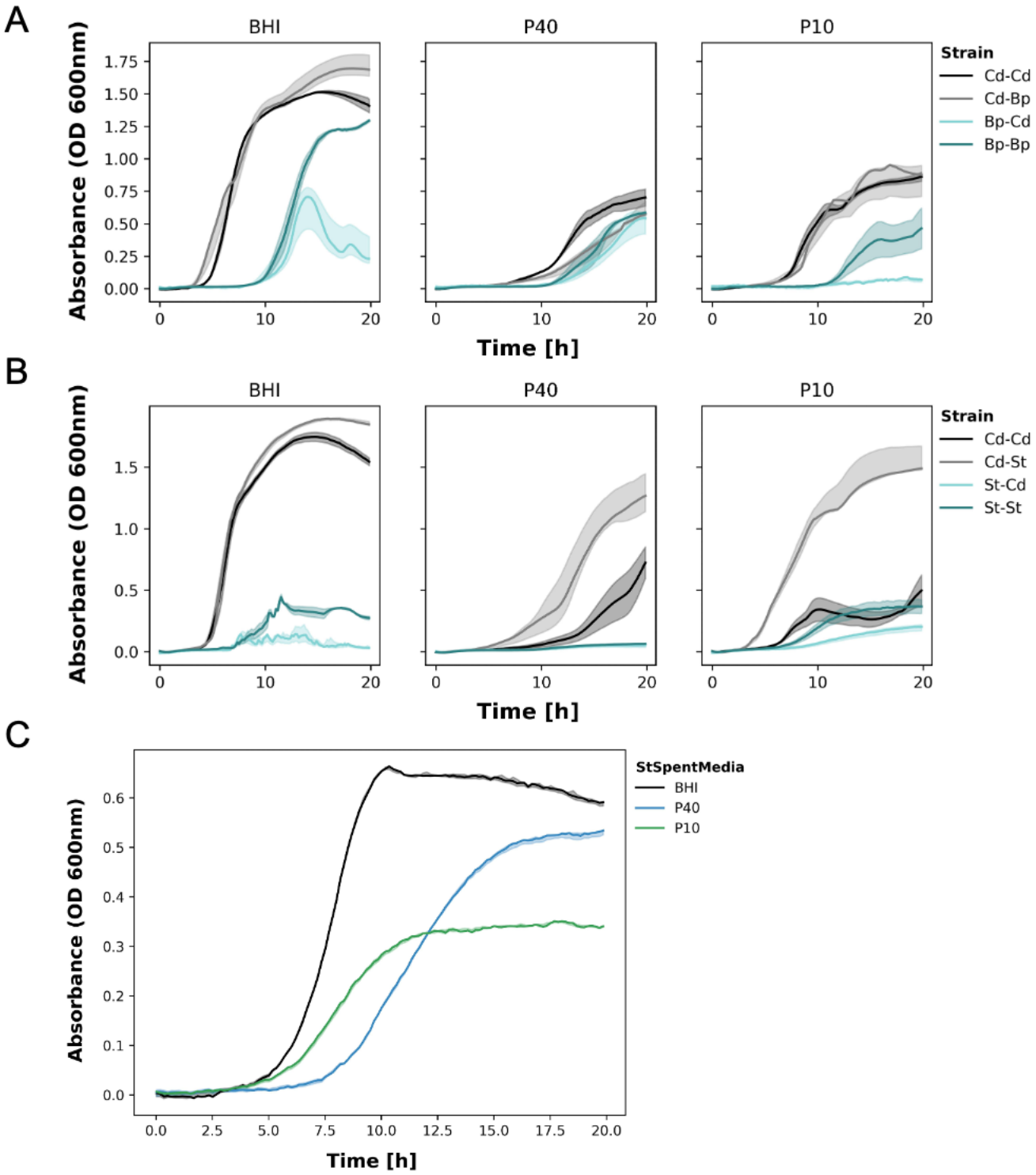
Supplemental Figures



Supplemental Figure 3.1. Metabolic contextualization of genetic information recapitulates differences in C. difficile isolates. (A) The workflow used to test if metabolic modeling can accurately capture metabolic differences between C. difficile strains. Whole genome sequences of three C. difficile strains with different fermentation profiles from Riedel, T. et al. 2017 were used as the input to reconstruct metabolic models. **(B)** PCA of flux distributions (n=50) for each model under each diet condition. **(C)** Normalized flux values for the 15 most important reactions from a Random Forest classification of strains. BHI: brain heart infusion; HPLC: high protein, low carbohydrate; LPHC: low protein, high carbohydrate.



Supplemental Figure 3.2. C. difficile strain clustering by diet. (A-B) Pair-wise Bray-Curtis dissimilarity index for flux samples from 105 C. difficile strains constrained to HPLC and LPHC in panels A and B, respectively.



Supplemental Figure 3.3. In vitro growth curves for C. difficile strains. (A-B) Median optical density (600 nm) and IQR of *C. difficile* growth with itself (black; n=3) or with its paired commensal (grey; n=3) and of commensal growth with itself (dark teal; n=3) or with *C. difficile* (light teal; n=3) in three media conditions (BHI, P40, and P10). *B. producta* and *S. thermophilus* are the commensals in A and B, respectively. **(C)** Median optical density (600 nm) and IQR for *C. difficile* grown in *S. thermophilus* spent media with pH adjusted to 7 (n=3).

Chapter 4: The gut microbiome is a cohort-dependent predictor of dementia

Authors: Deborah A. Powers[1], Glynis L. Kolling[2], Matthew L. Jenior[2], Laurie Archbald-Pannone[3,4], Jason A. Papin[1,2]

1. Biochemistry and Molecular Genetics, School of Medicine, University of Virginia, Charlottesville, Virginia, United States of America
2. Biomedical Engineering, School of Engineering, University of Virginia, Charlottesville, Virginia, United States of America
3. Department of Medicine, Division of General, Geriatric, Hospital & Palliative Medicine, University of Virginia., United States of America
4. Department of Medicine, Division of Infectious Diseases and International Health, University of Virginia., United States of America

Abstract
Alzheimer's disease and related dementias affect a significant proportion of elderly individuals worldwide,yet predicting which individuals will develop dementia is an active challenge. In an effort to improve predictions, biomarkers in blood, cerebrospinal fluid, and feces have been explored and used in conjunction with imaging techniques. The predictive capability of the intestinal microbiome is particularly intriguing because of its connection to the central nervous system via the gut-brain axis, which has been shown to be biologically relevant in other neuro-pathologies. However, a clear and consistent intestinal microbiome biomarker has yet to be identified. To explore the intestinal microbiome in dementia, we analyzed 16S rRNA and metagenomic sequence data from a new cohort of hospitalized patients collected from September – December 2020 (≥ 65 years old; dementia = 48; control = 322). We observed no significant differences in alpha- or beta-diversity metrics between the groups; however, we found that statistically significant, modest differences at the species level exist. To validate these findings, we compared our cohort with four independent, cross-sectional studies of Alzheimer's disease. Using a Random Forest classifier, we found that the predictive ability of the intestinal microbiome was cohort-specific. Additionally, differentially abundant genera were more unique across cohorts indicating a cohort-specific impact. Together these analyses suggest that while the intestinal microbiome may play a role in disease progression, its signals are complex and species compositional biomarkers may have limited utility as a dementia diagnostic.

Introduction

Dementia affects approximately 6.5 million Americans over the age of 65, costing an estimated 321 billion dollars annually[1,2]. It is defined as a syndrome that presents with deterioration of memory, language, thinking, and behavior[1,3]. Dementias are further categorized by the physiological cause of disease; Alzheimer's disease (AD) is the most common cause, comprising 60-80% of dementias[1]. The primary risk factors for dementia include age, genetics, and family history[1]. Dementia is one of the top 10 causes of death in the US, with dementia-related mortalities increasing 145% from 2000-2019[2]. Furthermore, the impact of dementia is only increasing: incidence in the US is projected to expand to 13.8 million by 2060[2].

Dementia is a progressive, neurodegenerative disease. Neurodegenerative disorders are caused by loss of neuron function in the central nervous system (CNS), resulting in degeneration or death of selective neuronal cells[4,5]. This neuronal damage can occur via oxidative damage, inflammation, or disrupted energy metabolism[4,5]. In AD, pathogenesis occurs through build-up of beta-amyloid plaques around neurons and neurofibrillary tangles of protein tau inside neurons[1,6]. Neurodegeneration is progressive, manifesting a continuum of symptoms from normal cognition to mild cognitive impairment to dementia, typically over a time span of 10-20 years[1,6]. Lifestyle and environmental factors including exercise and diet have been attributed to the rate of progression[1,7–12].

The gut microbiome (GM) appears to play a complex role in dementia. The GM is able to metabolize neurotransmitters and modulators such as short-chain fatty acids (SCFA), gamma-amino butyric acid (GABA), acetylcholine, dopamine, glutamate, and serotonin[3,6]. These neurotransmitters act through the gut-brain axis, a communication pathway between the enteric and central nervous systems (ENS, CNS)[3,6,13,14]. Furthermore, intestinal dysbiosis and inflammation are consistent symptoms across neurodegenerative disorders which may be due to general diffusion of neuroinflammation of the CNS via the gut-brain axis[6,14]. Mapping the associations between the GM composition and function and neurocognition has been a recent focus in dementia research[15–27] (Table 1). A subset of microbiome studies in dementia have found that *Akkermansia*[18,19], *Bacteroides*[15,24,27], *Dialister*[19,24], and *Roseburia*[15,19,23] species are differentially abundant in AD versus controls (Table 1). However, these findings are not consistent across all studies, likely due to a variety of factors such as study design, geographic locations, age, diet, disease stage, experimental protocols, and data processing. This inherent noise in microbiome data continues to cloud the relationship between the gut microbiome and dementia.

To investigate if there are taxonomic and metagenomic GM biomarkers or patterns consistently associated with dementia, we analyzed 16S rRNA gene sequencing and shotgun metagenomics from a clinical cohort of dementia patients from the University of Virginia (UVA). The 16S rRNA gene is a hypervariable region of the bacterial genome that is unique to each species; amplifying and sequencing this region enables bacterial identification from a complex milieu, such as stool[28]. Conversely, metagenomic sequencing reads all of the genomic DNA in a sample and provides more precise differentiation of species and strains[28,29]. Using these sequencing data, we found that there were no significant differences in gross GM composition or diversity between dementia and control groups; however, there were significant differentially abundant species between groups. To determine if these results are consistent with other studies, we summarized results from 13 microbiome studies of dementia cohorts around the world and then conducted a cross-cohort comparison using the UVA cohort and four independent cohorts. Overall, we found that differences between dementia and control groups within a cohort are not consistent across all cohorts and that there is a high degree of within-cohort specificity. These differences could be a function of regional differences in diet, environmental exposures, or any of a host of other variables that are frequently insufficiently tracked. Because of the lack of a robust microbiome signal in dementia, we posit that while the GM may still be important for health in dementia, it's utility for diagnosis remains limited.

Results

UVA 2020 cohort description and data collection

For this cohort, rectal swabs from 370 hospitalized UVA patients (age > 65) were collected from the UVA Clinical Microbiology Laboratory from September – December 2020 under IRB #22176. We conducted a manual chart review for all patients with rectal swabs to identify patients with dementia. Dementia was defined by a list of notations and/or dementia medications that were present in a patient's chart (Supplemental Data 4.1). We did not exclude patients based on antibiotic usage, with 78.1% of the cohort on antibiotics at the time of collection.

STUDY	SUBJECTS (N)	SETTING	AGE RANGE	DATA & ANALYSIS	↑ GENERA	↓ GENERA	DATA AVAILABILITY
HARAN 2019[1]	NC (51) AD (24) D (33)	Nursing home, MA, USA	72.8 - 93.2	Metagenomic Random Forest	Bacteroides† Odoribacter Barnsiella‡ Alistipes‡	Roseburia† Collinsella Eubacterium	On request of original authors
HOU 2021[2]	NC (40) AD (21)	Hospital*, Xinjin, China	64.4 - 78.8	16S V3-V4 QIIME	Escherichia/ Shigella Finegoldia Ruminococcaceae	Aerostipes Megamonas Enterococcus	Supplemental Material
LASKE 2022[3]	NC (100) AD (75)	Community*, Tübingen, Germany	60 - 77.2	Metagenomic Random Forest			PRJEB47976
LI 2019[4]	NC (30) MCI (30) AD (30)	Community, Shanghai, China	58.8 - 71.4	16S V3-V4 PiCRUSt	Lactobacillus Akkermansia† Dorea Bifidobacterium‡ Streptococcus Acinetobacter Blautia‡	Parabacteroides Alistipes‡ Alloprevotella Sutterella Barnesiella ‡ others	PRJNA489760
LING 2021[5]	NC (71) AD (100)	Community, Lishui, China	64.9 - 83.4	16S V3-V4 QIIME PiCRUSt	Bifidobacterium‡ Akkermansia†	Faecalibacterium† Roseburia† Coprococcus Dialister† others	PRJNA633959
LIU 2019[6]	NC (32) MCI (32) AD (33)	Community, Hangzhou, China	63.5 - 86.2	16S V3-V4 QIIME PiCRUSt	Proteobacteria*	Ruminococcus Blautia ‡	PRJNA496408
NAGPAL 2019[7]	NC (6) MCI (11)	Community, N. Carolina, USA	58.2 - 71	16S V4 QIIME			Availability not stated
SAJI 2020[8]	NC (82) D (25)	Community, Aichi, Japan	69-81	T-RFLP			On request of original authors
SHENG 2021[9]	NC (38) MCI (8) AD (6)	Hospital*, Beijing, China	61.7 - 81.1	16S V3-V4 ZOUT/UNOISE (ASV)		Faecalibacterium† Roseburia†	Availability not stated
VOGT 2017[10]	NC (25) AD (25)	Community, Wisconsin, USA	61.8 - 78.6	16S V4 MOTHUR	Bacteroides† Blautia‡ Alistipes‡ Gemella Biophila Phascolarctobacterium	Bifidobacteria‡ Dialister† Turicibacter Aldercreutzia	PRJEB51982
XI 2021[11]	NC (44) AD (21)	Community, Shanghai, China	66.3 - 86.1	16S V3-V4 QIIME	Faecalibacterium Agathobacter Alloprevotella Atopobium Parvimonas Solobacterium Pseudomonas	Tyzzerella Erysipelatoclostridium	PRJNA611839 Metadata does not ID groups (NC or AD)
YILDIRIM 2022[12]	NC (51) MCI (27) AD (47)	Community, Istanbul, Turkey	61.7 - 76.5	16S V3-V4 DADA2 (ASV)	Ruminococcaceae Prevotella_9/ Bacteroides Escherichia/ Shigella		PRJNA734525
ZHUANG 2018[13]	NC (43) AD (43)	Hospital*, Chongqing, China	60.5 - 79.0	16S V3-V4 QIIME PiCRUSt	Bacteroides† Lachnoclostridium	Subdoligranulum	PRJNA554111 Metadata does not ID groups (NC or AD)

Table 4.1 Literature survey. Summary of dementia microbiome cohorts including study size, setting, age range, data types and analysis approaches, genera reported as increased or decreased in the original study, and data availability. AD: Alzheimer's disease; D: dementia; MCI: mild cognitive impairment; NC: normal control. * Not explicitly stated. † Genera identified as differentially abundant in more than one study; ‡ genera identified as differentially abundant in more than one study, including in opposite directions.

Cohort demographics for sex, age, and race along with clinical metrics for dementia delirium, COVID status, antibiotic usage, and mortality during study collection are summarized in Table 4.2. Total DNA was isolated from all swabs resulting in 357 samples with sufficient DNA quality for 16S rRNA gene sequencing. Additionally, 20

samples representative of the cohort demographics were processed for shotgun metagenomic sequencing (Fig 4.1). This cohort has several unique features compared to other published cohorts that have explored the microbiome in AD. The samples were collected from hospitalized patients while most published studies are collected from community or nursing home settings (Table 4.1). Additionally, the dementia criterion is not limited to AD, including multiple dementia types. Most importantly, we included patients on antibiotic medications. Together these features represent common clinical settings that a diagnostic tool must be able to reliably perform in to be clinically useful.

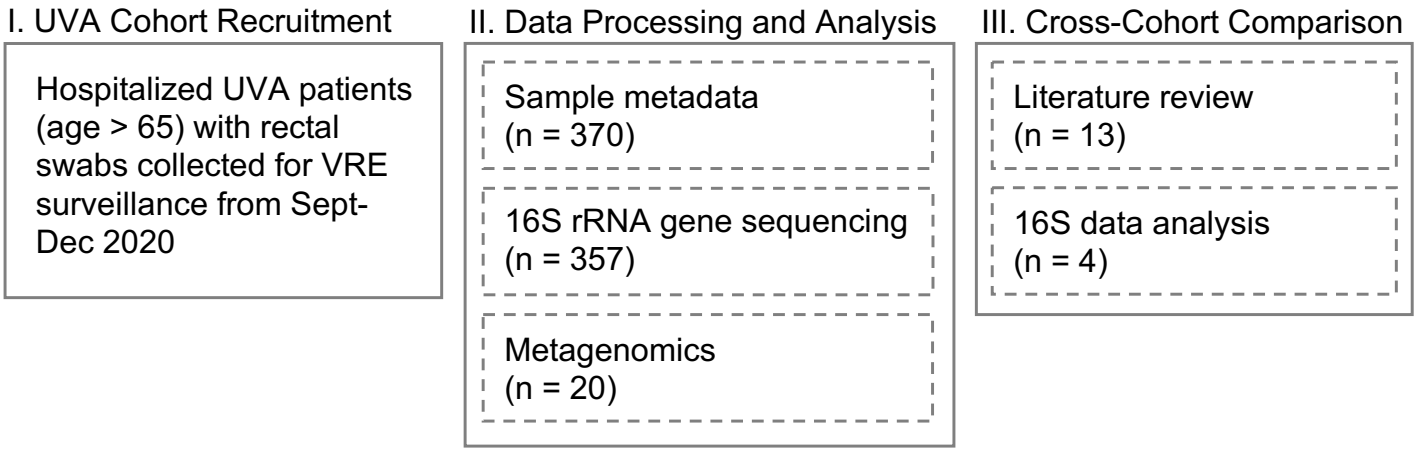


Figure 4.4. UVA 2020 dementia cohort workflow. Of the 370 recruited patients, rectal swabs from 357 patients were of sufficient quality for 16S rRNA gene sequencing. Out of those 357 patients, 20 representative patients were selected for metagenomics sequencing. The cross-cohort comparison reviewed 13 studies of the Alzheimer's microbiome; four of these studies were included in cross-cohort analysis of 16S rRNA gene sequencing. VRE: Vancomycin resistant Enterococcus.

	DEMENTIA (N=48)	CONTROL (N = 322)	P-VALUE
SEX (F/M)	26/22 (54.2%)	138/184 (42.9%)	NS
AGE (MEDIAN)	83.2	73.1	<0.00001
RACE			NS
ASIAN	0 (0%)	2 (0.6%)	
AFRICAN AMERICAN	2 (4.2%)	26 (8.1%)	
OTHER	1 (2.1%)	2 (0.6%)	
CAUCASIAN	26 (54.2%)	179 (55.6%)	
NR	19 (39.6%)	113 (35.1%)	
DELIRIUM	11 (22.9%)	39 (12.1%)	0.0693
COVID	1 (2.08%)	47 (8.07%)	NS
ANTIBIOTIC USAGE	35 (72.9%)	254 (78.9%)	NS
MORTALITY	17 (35.4%)	83 (25.8%)	NS

Table 4.2. Cohort demographics. Descriptive summary of cohort features. Significance of differences in sex, race, delirium, covid, antibiotic usage, and mortality between dementia and control groups was calculated using a chi-squared test. Significance of differences in age between dementia and control groups was calculated using the Wilcoxon rank sum test. NR = Not Reported, NS = Not Significant.

16S rRNA gene sequencing analysis shows minor differentiation between groups

Because of significant differences in the age and size of the dementia (n = 48) and control groups (n = 322) (Table 4.2), we randomly down sampled the control group while matching for age, sex, and race before

conducting our 16S analysis. Relative abundance of family-level taxa composition averaged by group as well as alpha- (Inverse Simpson) and beta-diversity do not show any high-level, significant differences between the dementia and control groups (Fig 4.2A-C). While it is possible this lack of differentiation between groups could be driven by the high antibiotic usage within the cohort (78.1% overall), we did not find any significant differences in Inverse Simpson or Shannon diversity metrics based on antibiotic usage (Fig S4.1A-B). Despite a lack of gross differences between groups, we identified significant differential taxa abundances between the groups using DESeq2 that included increased abundance of several *Bacteroides* species and decreased abundance of *Porphyromonas*, *Prevotella*, and *Corynebacterium* species in the dementia group relative to the control (Fig 4.2D). To investigate the metabolic function of the significant differentially abundant species, we correlated the 16S taxa profile of each sample with the inferred metabolic function of each sample using PiCRUSt2 (Fig 4.2E). Notably, degradation and detoxification pathways were present in species differentially abundant in the control group compared to the dementia group.

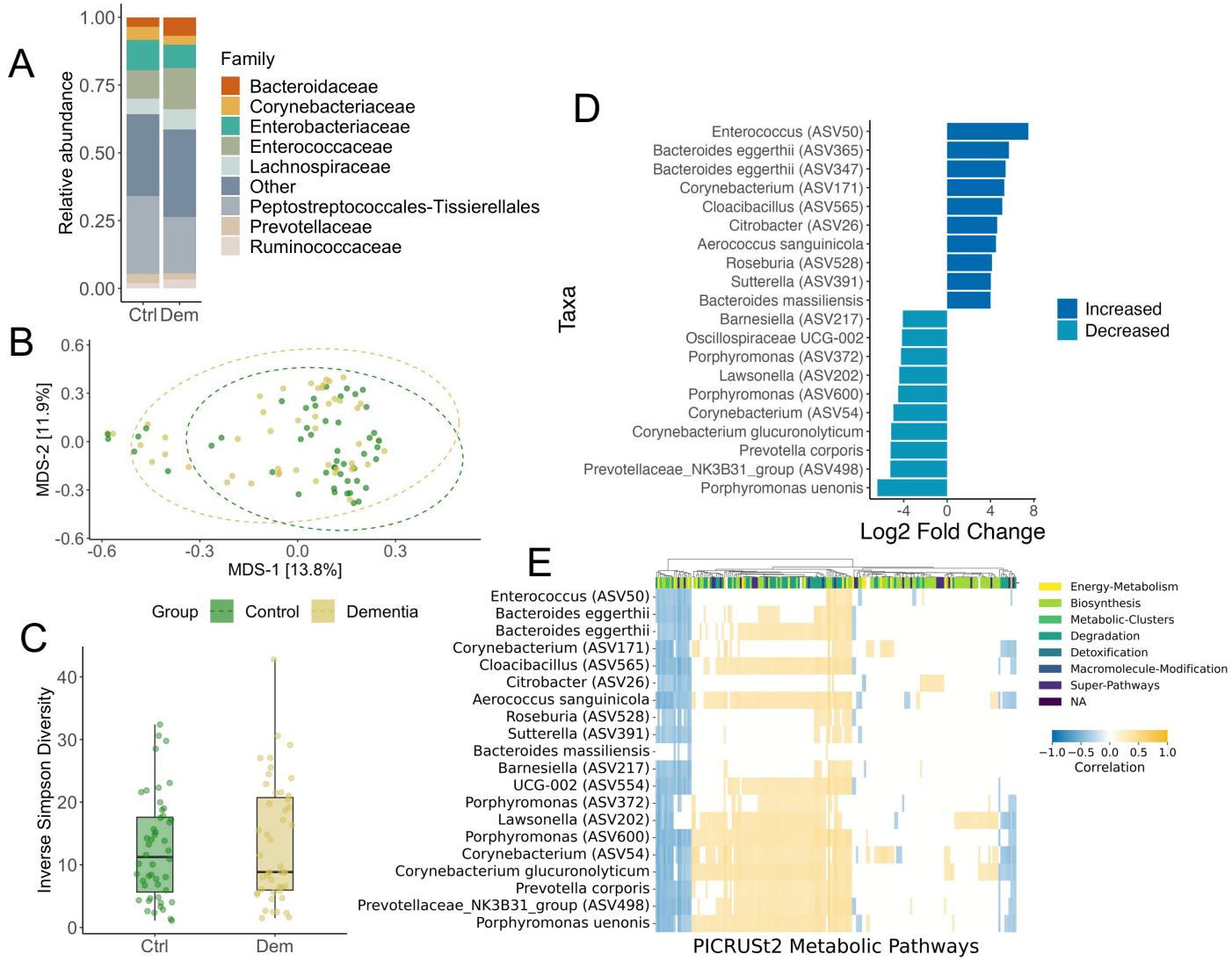


Figure 4.2. Differences between dementia and control groups in the UVA 2020 cohort are complex. (A) Taxonomic composition of dementia and control groups demonstrates similar abundances. (B) MDS with Bray-Curtis distance of beta-diversity of dementia and control groups; no significant difference between centroids by PERMANOVA ($\rho = 0.122$, $R^2 = 0.0139$). (C) Alpha-diversity measure by inverse Simpson of dementia and control groups; no significant difference between groups by Kolmogorov-Smironv test ($\rho = 0.75$). (D) Differentially increased (dark blue) and decreased (blue) abundance of taxa in dementia relative to control. (E) Pearson's correlation between the predicted metabolic function and the differentially abundant taxa.

Metagenomic sub-sampling supports lack of functional differentiation

Because there were minimal differences in gross GM compositions and few differentially abundant species, we considered whether microbial function was more important than microbial composition in differentiating between dementia and controls. To test the GM metabolic functional profile, 20 samples representative of the cohort

demographics were selected for shotgun metagenomic sequencing (Supplemental Data 4.1). The inferred taxonomic composition replicated similar compositional trends of the 16S rRNA sequencing from the same samples (Fig S4.1 E-F). However, there were no significant functional differences between the groups (Fig S4.1G). This result validates the PiCRUST2 inferred metabolic function from the 16S rRNA sequencing data which also found no differences between the dementia and control groups (Fig S4.1H).

Limited consensus in differentially abundant species across dementia microbiome cohorts

To discern whether the findings from the UVA 2020 cohort were consistent with microbiome data from other dementia cohorts, we surveyed findings from 13 unique studies of the intestinal microbiome in AD (Table 4.1). These studies encompassed a range of geographic locations, population environments, ages, dementia and AD definitions, sequencing regions, and analysis protocols. *Akkermansia*, *Alistipes*, *Bacteroides*, *Bifidobacterium*, and *Blautia* genera were reported as differentially abundant in dementia patients in more than one study while *Dialister*, *Faecalibacterium*, and *Roseburia* genera were differentially decreased in more than one study (Table 4.1). However, there were no genera that were consistently differential across all studies; the maximum consensus across cohorts for any genus reported as differential was 3 out of 13.

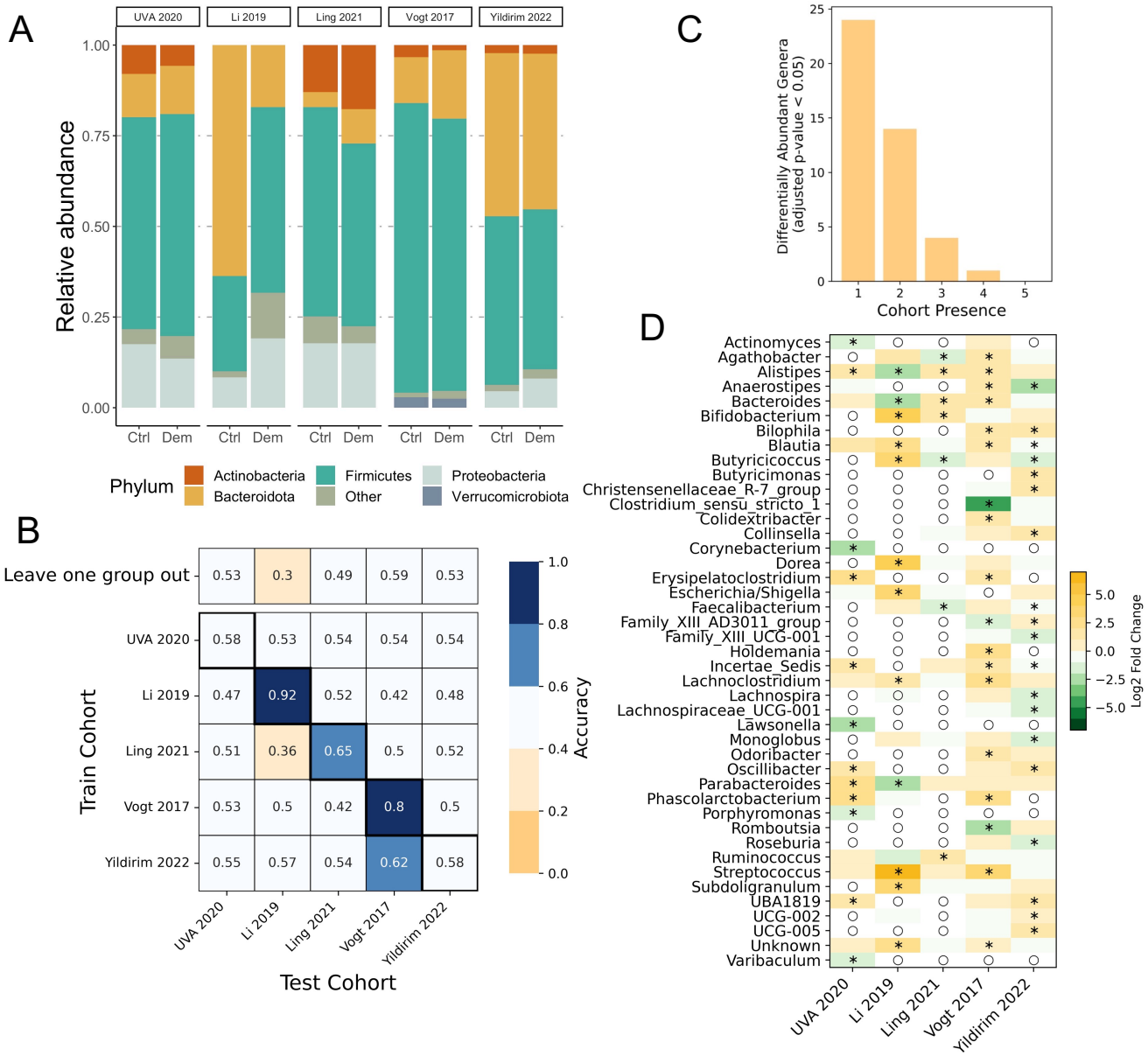


Figure 4.3. Cross-cohort 16S comparison fails to find a consistent, robust microbiome signal. (A) Taxonomic composition of dementia and control groups across cohorts. (B) Predictive accuracy of Random Forest model across cohorts. The bolded diagonal line highlights within-cohort testing. (C) Genera that were significantly differentially abundant in dementia binned by the number of cohorts in which they were found to be significantly differential. (D) Differential abundance of genera that were significant in at least one cohort as

calculated by DESeq2 with Benjamin-Hochberg correction. Asterisks indicate adjusted ρ-values < 0.05; circles indicate a genus not identified within the cohort.

Because differences in sequencing and differential abundance analysis methods can have a significant impact on the findings from microbiome data analysis[30], we decided to re-analyze the raw sequencing data from a subset of cohorts in our literature review to standardize the comparison. We selected data from Li 2019, Ling 2021, Vogt 2017, and Yildirim 2022 for our cross-cohort analysis because the sequencing data was publicly available, the metadata clearly identified sample dementia status, and the studies covered a range of geographic locations. There were no significant differences in phylum-level taxa composition, alpha, or beta diversity between dementia and control groups across cohorts (Fig 4.3A, Fig S4.2A-B) with the exception of Li 2019 which had differences in the *Bacteroidota* to *Firmicutes* ratio between the dementia and control groups and a significant difference in alpha diversity ($\rho = 0.015$).

To identify if there are consistently differential genera in dementia vs control microbiomes, we conducted a differential genera-level abundance analysis using DESeq2. The majority of genera identified as significantly differential ($\rho < 0.05$, Benjamin-Hochberg correction) were only detected as significant in one or two cohorts and there were none detected as significant across all cohorts (Fig 4.3C). Furthermore, for the genera that were significant in more than one cohort, there is often disagreement in the direction of the log2 fold changes. For example, *Alistipes* is detected as significant in UVA 2020, Li 2019, Ling 2021, and Vogt 2017; however, it is differentially decreased in Li 2019 and differentially abundant in the other cohorts (Fig 4.3D). The absence of any genera that are consistently differential across all cohorts further highlights the challenges with the microbiome as a marker for dementia diagnosis.

Poor predictive ability of dementia across cohorts

While we did not find specific species or genera that were consistently differential, it is possible that a machine learning model could be generated that accounts for small differences across the entire microbiome that could be predictive of dementia status. To generate such a model, we applied a Random Forest classifier to genera-level abundance data. Briefly, we prepared the data by filtering out low-abundance genera, applying a log10 transformation and then converting to standardized Z-scores. We then conducted three Random Forest analyses: within-cohort predictions, cohort-pairwise predictions, and leave-one-group-out (LOGO) predictions. For the within-cohort predictions, we applied a Random Forest classifier with a five-fold, stratified cross-validation repeated 20 times to each cohort and then calculated the median accuracy (Fig 4.3B, bolded diagonal). For the cohort-pairwise predictions, we trained a Random Forest classifier on a single cohort and then iteratively tested on the remaining cohorts. We repeated each train-test cohort pair 20 times and then calculated the median accuracy (Fig 4.3B). Finally, for the LOGO predictions, we trained a Random Forest classifier on all the cohorts except one; each LOGO permutation was repeated 20 times and the median accuracy calculated (Fig 4.3B). Overall, there is poor cross-cohort predictive ability, with the majority of accuracy scores centering around 0.5 (random chance). The within-cohort accuracy scores are much higher, in particular for Li 2019 and Vogt 2017. Both of these results point to the high-degree of cohort specificity and the lack of a robust signal that can consistently be used to predict dementia status across studies.

Discussion

The GM is a largely unmined, unexplored aspect of health and disease; it possesses great potential for diagnosis, staging, and treatment in a number of diseases. For example, *C. difficile* infection (CDI), which occurs when the GM has been disrupted through administration of antibiotics, can be completely resolved by restoring a healthy microbiome through fecal microbiota transplants (FMT)[31]. Similar bacteriotherapy interventions in intestinal bowel disease (IBD) are an active area of research because the GM in IBD presents with consistent patterns of reduced diversity, decreased abundances of specific taxa within the *Firmicutes* and *Bacteroides* phyla, as well as decreased synthesis of metabolites critical for gut health[32]. The GM has also been found to impact diseases not directly associated with intestinal health. Notably, in depression, the GM presents with specific compositional patterns and multiple studies have independently found reduced symptoms of depression following various probiotic interventions[33,34]. Hypothesizing analogous biological interactions via the gut-brain axis as well as novel diagnosis and treatment opportunities, the relationship between the GM and dementia has been studied extensively in multiple settings and geographical locations[15–27,35]. Unanimously, these studies have found associations between the GM and dementia; however, these studies have yet to uncover a consistent signal

across all cohorts. GM data is noisy and the recent best-practice recommendation in the field is to conduct cross-cohort analyses to increase statistical power[36–39]. Therefore, we conducted a cross-cohort analysis in which we collected and analyzed data from a UVA dementia cohort and then validated these findings using four independent dementia cohorts with the goal of identifying a signal within the noise.

The UVA cohort has small GM differences between the dementia and control groups. Taxa composition and alpha- and beta-diversity are not statistically different between groups (Fig 4.2A-C), indicating a similar ecological structure. We did find differentially abundant species between groups; when we correlated these species to their predicted associated metabolic function, species increased in dementia had reduced predicted functions in degradation and detoxification pathways (Fig 4.2D-E). This result may indicate that while there is overall a similar taxonomic structure, differentially abundant species are contributing to differences in function and metabolism. However, both PCA of the PiCRUSt2 inferred metabolic pathway abundances as well as the NMDS of functional metagenomic data do not show any functional clustering by group (Fig S4.1G-H).

The analysis of the UVA cohort does present specific challenges, such as the classification of dementia, antibiotic use, age, and number of patients between groups (Table 4.2). We addressed the differences in median age and group size by randomly down-sampling the control group while matching on age, sex, and race. We included patients on antibiotics because they represent a majority of elderly patients which should be considered when searching for a potential GM biomarker diagnostic. While it is possible that the wide antibiotic use throughout the cohort erases GM differences between the groups, we did not see significant differences in diversity between patients who were not on antibiotics versus patients who were on 1-4 antibiotic regimens; at 5 or more antibiotic treatments we do see significant drops in diversity (Fig S4.1A-B). An important caveat to this antibiotic analysis is that it is unknown if patients who were not on antibiotics at the time of collection were on antibiotics within the previous 3-6 months. Finally, the functional analysis of the cohort should be interpreted with caution as the metabolic function from PiCRUSt2 is an inferred function based on the taxonomic profile of each sample; the functional data from metagenomic sequencing is not inferred, but the small sample size (n = 20) has limited statistical power. However, PiCRUSt2 analysis of the 16S rRNA sequencing data from the cross-cohort comparison also showed no functional differences based on dementia status (Fig S4.2C), providing support to the interpretation that there are no consistent GM functional differences in dementia versus controls.

The cross-cohort comparison tests the robustness and generalizability of GM patterns in dementia, addressing limitations within the UVA cohort specifically but also single-cohort analyses in general. The literature survey of 13 independent dementia GM cohorts was geographically diverse and included multiple community settings (Table 4.1). Taxonomic structure was determined by 16S rRNA gene sequencing or metagenomic sequencing and certain genera were increased or decreased in more than one study; however, the maximum consensus across all studies was 3 out of 13 (Table 4.1). This poor consensus across studies indicates that either there is not a consistent GM signal in dementia or that differences in cohort preparation and analysis are obscuring the true signal. However, in the analysis of 16S rRNA gene sequencing data from five independent cohorts we failed to find a robust GM signal. Overall, there were not significant differences between the dementia and control groups by taxonomy and diversity (Fig 4.3A, Fig S4.2A-B). Additionally, most differentially abundant genera were found in only one or two cohorts, and none were found in all five cohorts (Fig 4.3C-D). Furthermore, Random Forest machine learning was only accurate for within-cohort training and testing and had poor prediction accuracy when trained on one cohort and then tested on another (Fig 4.3B). These results support the hypothesis that the lack of consensus between the studies in the literature survey is due to an inconsistent GM signal in dementia and not due to differences in sequencing analysis. Another possibility is that a larger metagenomics comparison could find functional similarities not seen in a species compositional comparison across geographical locations.

In conclusion, we conducted a cross-cohort analysis of dementia microbiome data to determine if the GM can be used as a reliable biomarker of dementia. The GM is critical to overall health and disruption of the GM in dementia can lead to a host of secondary symptoms[40]. In cases with a clear association between disease and the GM, the GM has successfully been used as a biomarker for disease or as a therapeutic target. A recent study used microbiome data in conjunction with current gold-standard imaging methods to predict preclinical status in AD but found only minor improvements in predictions[35]. Additionally, we have not found a consistent association between the GM and dementia. Instead, we find that the GM in dementia experiences non-specific dysbiosis,

which is more likely a secondary symptom of disease rather than a causative agent. Future microbiome research in dementia could focus on whether the GM can be used as a therapeutic target even if it cannot be used as a predictor. For example, intervention with probiotics or microbial communities could improve secondary symptoms of disease similar to depression. It would also be of interest to explore whether alleviation of these symptoms lead to overall improvement or slower progression of disease.

Methods

Cohort description: The study presented herein was conducted from September to December of 2020 at UVA Health (IRB-HSR# 22176) to pair microbiome and patient health data from vancomycin-resistant *Enterococcus* surveillance rectal swabs. Rectal swabs have been used as surrogates for fecal samples[41], and in the study presented here, enabled a survey of inpatients during a defined window of time. Specifically, all surveillance swabs were collected after processing by the clinical microbiology laboratory at UVA prior to discarding on a weekly basis. Once swabs were collected, the sample set was refined by retaining any swab from a patient ≥ 65 years old and a unique identifying number was assigned to each sample. In addition to age, we also collected sex, cognitive status (e.g., delirium, dementia), medications related to dementia, death within 10d of swab collection, and COVID19 status at sample collection (Supplementary Data File 1). Details of the cohort are presented in Table 2; significance of differences in sex, race, delirium, covid, antibiotic usage, and mortality between dementia and control groups was calculated using a chi-squared test; significance of differences in age between dementia and control groups was calculated using the Wilcoxon rank sum test. Samples were stored at -20C until DNA isolation was performed.

DNA isolation and sequencing: Total DNA was extracted from rectal swabs using the Qiagen QIAamp PowerFecal Pro DNA Kit by letting the swab sit in solution CD1 for 5-10 min, mixed to remove debris and then processed according to the manufacturer's instructions. Extracted DNA was stored at -20C until further processing. Total DNA was measured using the DeNovix Broad Range or High Sensitivity dsDNA kits on a DeNovix fluorometer. Libraries were prepared for 16S sequencing through amplification of the V4 region of the 16S rRNA gene according to established protocols[42] and sequenced by Genewiz. Metagenomic sequencing was performed on total DNA samples using Illumina NextSeq 2000 sequencing (2 x 151bp; SeqCenter, LLC., Pittsburgh, PA).

16S rRNA gene sequencing data processing: Processing and analysis of the 16S samples was performed using a combination of computational methods. Demultiplexing and adaptor and primer trimming of the raw reads were performed as part of the MiSeq platform standard protocol. The removal of adaptor and primer sequences as well as overall read quality was verified using FASTQC and MultiQC[43]. Further processing of the raw reads including filtering, trimming, merging, and sample inference was completed using DADA2[44] resulting in a final ASV counts table to be used for downstream analysis. For the UVA cohort, we filtered out control samples with delirium and then randomly downsampled the control group to match the dementia group using an optimization solver set to best match for the following constraints: sex, age, race, antibiotic use. In the cross-cohort analysis, samples classified as mild cognitive impairment (MCI) were removed before downstream analyses. We analyzed taxonomic composition, alpha diversity, and beta diversity using R packages Phyloseq[45] and microbiome[46]. Alpha diversity was calculated using inverse Simpson and significance between groups was calculated using the Kolmogorov-Smironv test with Bonferroni correction; beta diversity was calculated using multidimensional scaling (MDS) with Bray-Curtis distance and significance between group centroids was calculated using permutational multivariate analysis of variance (PERMANOVA). Differential abundance was calculated using DESeq2[47] with multiple tests correction by Benjamin-Hochberg. Metabolic function was inferred using PiCRUSt2[48] and visualized using principle component analysis (PCA). The correlation between taxa abundance and the PiCRUSt2 inferred metabolic pathway abundance was calculated using Pearson's correlation.

Metagenomics analysis: 20 samples from the cohort were chosen for metagenomics sequencing using best match for gender, age, antibiotics, and race metrics between the dementia and control groups while also maintaining an age and antibiotics distribution similar to the original cohort. Adaptors, primers, duplicate reads, and host genome contamination were removed using Trimmomatic[49], Bowtie2[50], and custom scripts. We used HUMAnN3[51] for functional annotation and Kraken2[52] for taxonomic profiling. Non-metric multidimensional scaling (NMDS) was used to visualize similarities between groups.

Random Forest cross-cohort analysis: For all five cohorts, we calculated the genera-level relative abundance and filtered out low-abundance genera (relative abundance < 0.001). Then we transformed the data using log10 transformation and converted to standardized Z-scores. Using this data, we performed three Random Forest analyses: within-cohort, pairwise, and leave-one-group-out (LOGO) testing. For within-cohort, we used stratified 5-fold cross-validation repeated 20 times to create 100 unique train-test permutations which were then fit to a Random Forest classifier. The median accuracy for each cohort was then calculated. For pairwise cohort testing, we trained a Random Forest classifier on one cohort and then tested on the remaining cohorts. For each pair, the classifier was fit 20 times using 20 randomized random-states and the median accuracy was then calculated. Finally, for the LOGO testing, the classifier was trained on all of the cohorts except for one which was reserved as the test set. For each LOGO permutation, the classifier was fit 20 times using 20 randomized random-states and the median accuracy was then calculated.

Acknowledgements
This work was funded by a University of Virginia Trans-University Microbiome Initiative (TUMI) award.

References

1. 2022 Alzheimer's disease facts and figures. *Alzheimers Dement* **18**, 700–789 (2022).
2. Rajan, K. B. *et al.* Population estimate of people with clinical Alzheimer's disease and mild cognitive impairment in the United States (2020-2060). *Alzheimers Dement* **17**, 1966–1975 (2021).
3. Shabbir, U., Arshad, M. S., Sameen, A. & Oh, D.-H. Crosstalk between Gut and Brain in Alzheimer's Disease: The Role of Gut Microbiota Modulation Strategies. *Nutrients* **13**, 690 (2021).
4. Erkkinen, M. G., Kim, M.-O. & Geschwind, M. D. Clinical Neurology and Epidemiology of the Major Neurodegenerative Diseases. *Cold Spring Harb Perspect Biol* **10**, a033118 (2018).
5. Jellinger, K. A. Basic mechanisms of neurodegeneration: a critical update. *J Cell Mol Med* **14**, 457–487 (2010).
6. Sochocka, M. *et al.* The Gut Microbiome Alterations and Inflammation-Driven Pathogenesis of Alzheimer's Disease-a Critical Review. *Mol Neurobiol* **56**, 1841–1851 (2019).
7. De la Rosa, A. *et al.* Physical exercise in the prevention and treatment of Alzheimer's disease. *J Sport Health Sci* **9**, 394–404 (2020).
8. Huang, X. *et al.* Comparative efficacy of various exercise interventions on cognitive function in patients with mild cognitive impairment or dementia: A systematic review and network meta-analysis. *J Sport Health Sci* **11**, 212–223 (2022).
9. Wajman, J. R., Mansur, L. L. & Yassuda, M. S. Lifestyle Patterns as a Modifiable Risk Factor for Late-life Cognitive Decline: A Narrative Review Regarding Dementia Prevention. *Curr Aging Sci* **11**, 90–99 (2018).
10. Baranowski, B. J., Marko, D. M., Fenech, R. K., Yang, A. J. T. & MacPherson, R. E. K. Healthy brain, healthy life: a review of diet and exercise interventions to promote brain health and reduce Alzheimer's disease risk. *Appl Physiol Nutr Metab* **45**, 1055–1065 (2020).
11. Dhana, K., Evans, D. A., Rajan, K. B., Bennett, D. A. & Morris, M. C. Healthy lifestyle and the risk of Alzheimer dementia: Findings from 2 longitudinal studies. *Neurology* **95**, e374–e383 (2020).
12. Shannon, O. M. *et al.* Mediterranean diet adherence is associated with lower dementia risk, independent of genetic predisposition: findings from the UK Biobank prospective cohort study. *BMC Med* **21**, 81 (2023).
13. Liu, S., Gao, J., Zhu, M., Liu, K. & Zhang, H.-L. Gut Microbiota and Dysbiosis in Alzheimer's Disease: Implications for Pathogenesis and Treatment. *Mol Neurobiol* **57**, 5026–5043 (2020).
14. Cryan, J. F., O'Riordan, K. J., Sandhu, K., Peterson, V. & Dinan, T. G. The gut microbiome in neurological disorders. *Lancet Neurol* **19**, 179–194 (2020).
15. Haran, J. P. *et al.* Alzheimer's Disease Microbiome Is Associated with Dysregulation of the Anti-Inflammatory P-Glycoprotein Pathway. *mBio* **10**, e00632-19 (2019).
16. Hou, M., Xu, G., Ran, M., Luo, W. & Wang, H. APOE-ε4 Carrier Status and Gut Microbiota Dysbiosis in Patients With Alzheimer Disease. *Front Neurosci* **15**, 619051 (2021).
17. Laske, C. *et al.* Signature of Alzheimer's Disease in Intestinal Microbiome: Results From the AlzBiom Study. *Front. Neurosci.* **16**, 792996 (2022).
18. Li, B. *et al.* Mild cognitive impairment has similar alterations as Alzheimer's disease in gut microbiota. *Alzheimers Dement* **15**, 1357–1366 (2019).

19. Ling, Z. *et al.* Structural and Functional Dysbiosis of Fecal Microbiota in Chinese Patients With Alzheimer's Disease. *Front Cell Dev Biol* **8**, 634069 (2020).
20. Liu, P. *et al.* Altered microbiomes distinguish Alzheimer's disease from amnestic mild cognitive impairment and health in a Chinese cohort. *Brain Behav Immun* **80**, 633–643 (2019).
21. Nagpal, R., Neth, B. J., Wang, S., Craft, S. & Yadav, H. Modified Mediterranean-ketogenic diet modulates gut microbiome and short-chain fatty acids in association with Alzheimer's disease markers in subjects with mild cognitive impairment. *EBioMedicine* **47**, 529–542 (2019).
22. Saji, N. *et al.* Relationship between dementia and gut microbiome-associated metabolites: a cross-sectional study in Japan. *Sci Rep* **10**, 8088 (2020).
23. Sheng, C. *et al.* Altered Gut Microbiota in Adults with Subjective Cognitive Decline: The SILCODE Study. *J Alzheimers Dis* **82**, 513–526 (2021).
24. Vogt, N. M. *et al.* Gut microbiome alterations in Alzheimer's disease. *Sci Rep* **7**, 13537 (2017).
25. Xi, J. *et al.* Disturbed microbial ecology in Alzheimer's disease: evidence from the gut microbiota and fecal metabolome. *BMC Microbiol* **21**, 226 (2021).
26. Yıldırım, S. *et al.* Stratification of the Gut Microbiota Composition Landscape across the Alzheimer's Disease Continuum in a Turkish Cohort. *mSystems* **7**, e0000422 (2022).
27. Zhuang, Z.-Q. *et al.* Gut Microbiota is Altered in Patients with Alzheimer's Disease. *J Alzheimers Dis* **63**, 1337–1346 (2018).
28. Wensel, C. R., Pluznick, J. L., Salzberg, S. L. & Sears, C. L. Next-generation sequencing: insights to advance clinical investigations of the microbiome. *J Clin Invest* **132**, e154944 (2022).
29. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol* **35**, 833–844 (2017).
30. Nearing, J. T. *et al.* Microbiome differential abundance methods produce different results across 38 datasets. *Nat Commun* **13**, 342 (2022).
31. Bakken, J. S. Fecal bacteriotherapy for recurrent Clostridium difficile infection. *Anaerobe* **15**, 285–289 (2009).
32. Zuo, T. & Ng, S. C. The Gut Microbiota in the Pathogenesis and Therapeutics of Inflammatory Bowel Disease. *Front Microbiol* **9**, 2247 (2018).
33. Liu, L. *et al.* Gut microbiota and its metabolites in depression: from pathogenesis to treatment. *eBioMedicine* **90**, 104527 (2023).
34. Valles-Colomer, M. *et al.* The neuroactive potential of the human gut microbiota in quality of life and depression. *Nat Microbiol* **4**, 623–632 (2019).
35. Ferreiro, A. L. *et al.* Gut microbiome composition may be an indicator of preclinical Alzheimer's disease. *Sci. Transl. Med.* **15**, eabo2984 (2023).
36. Muller, E., Algavi, Y. M. & Borenstein, E. The gut microbiome-metabolome dataset collection: a curated resource for integrative meta-analysis. *NPJ Biofilms Microbiomes* **8**, 79 (2022).
37. Thomas, A. M. *et al.* Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat Med* **25**, 667–678 (2019).
38. Lee, K. A. *et al.* Cross-cohort gut microbiome associations with immune checkpoint inhibitor response in advanced melanoma. *Nat Med* **28**, 535–544 (2022).
39. Muller, E., Algavi, Y. M. & Borenstein, E. A meta-analysis study of the robustness and universality of gut microbiome-metabolome associations. *Microbiome* **9**, 203 (2021).
40. Varesi, A. *et al.* The Potential Role of Gut Microbiota in Alzheimer's Disease: From Diagnosis to Treatment. *Nutrients* **14**, 668 (2022).
41. Bassis, C. M. *et al.* Comparison of stool versus rectal swab samples and storage conditions on bacterial community profiles. *BMC Microbiol* **17**, 78 (2017).
42. Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K. & Schloss, P. D. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl Environ Microbiol* **79**, 5112–5120 (2013).
43. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
44. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* **13**, 581–583 (2016).
45. McMurdie, P. J. & Holmes, S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE* **8**, e61217 (2013).

46. Lahti, Leo, S., Sudarshan. Tools for microbiome analysis in R. Version. (2023).
47. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
48. Douglas, G. M. *et al.* PICRUSt2 for prediction of metagenome functions. *Nat Biotechnol* **38**, 685–688 (2020).
49. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
50. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).
51. Beghini, F. *et al.* Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife* **10**, e65088 (2021).
52. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol* **20**, 257 (2019).

Supplemental Data 4.1. Deidentified clinical metadata for UVA cohort.

Supplemental Figures

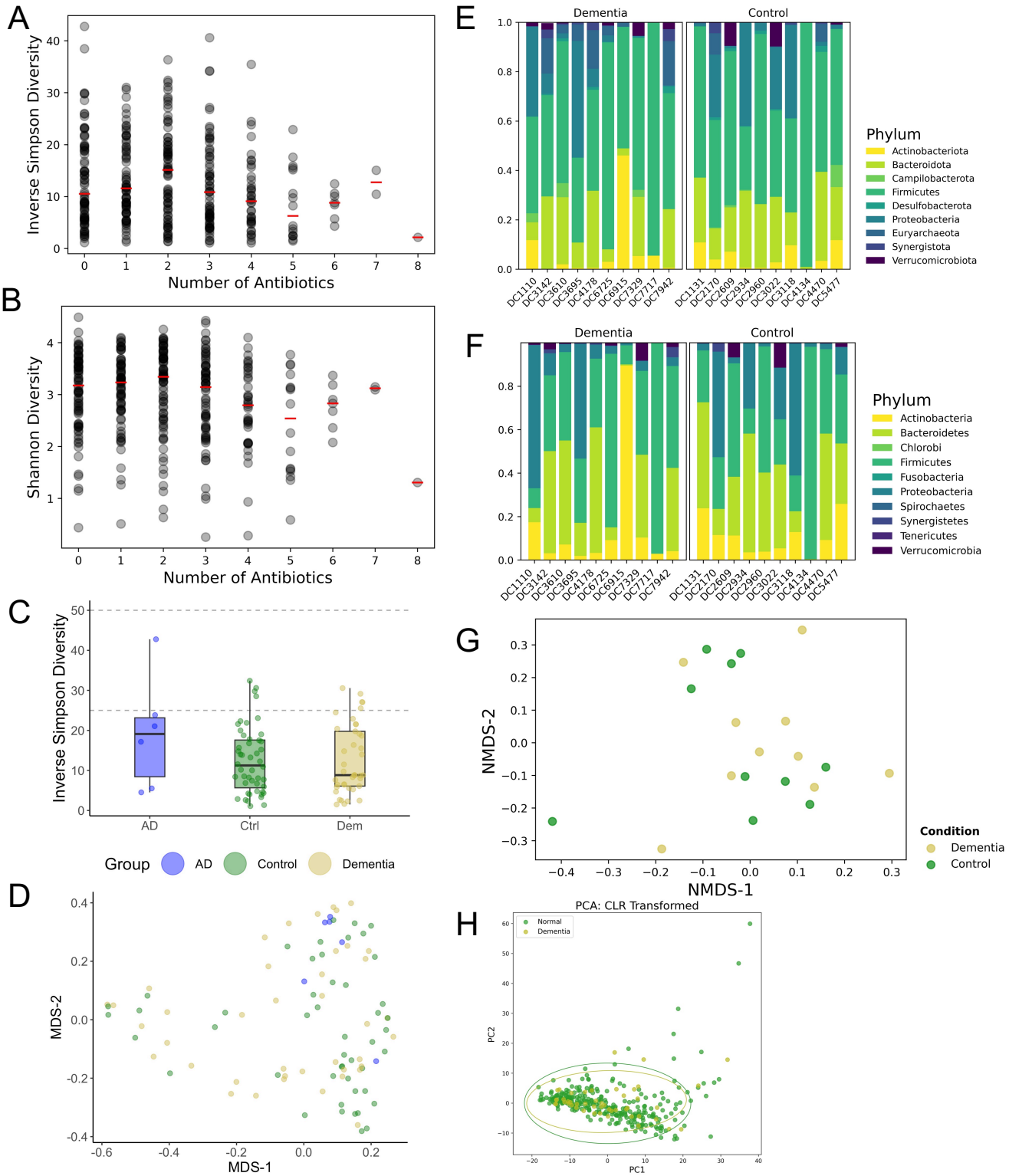


Figure S4.1. Supporting information for UVA 2020 cohort. (A-B) 16S alpha-diversity by the number of antibiotics at the time of collection calculated by Inverse Simpson and Shannon diversity metrics, respectively. (C) Inverse Simpson alpha-diversity of AD, dementia, and control 16S samples. (D) MDS with Bray-Curtis distance of beta-diversity with samples colored by group. (E-F) Taxa composition for the 20 samples with both 16S and metagenomics data, respectively. (G) NMDS of functional metagenomic data using Bray-Curtis distance. (H) PCA of PiCRUSt2 predicted functional pathway abundance.

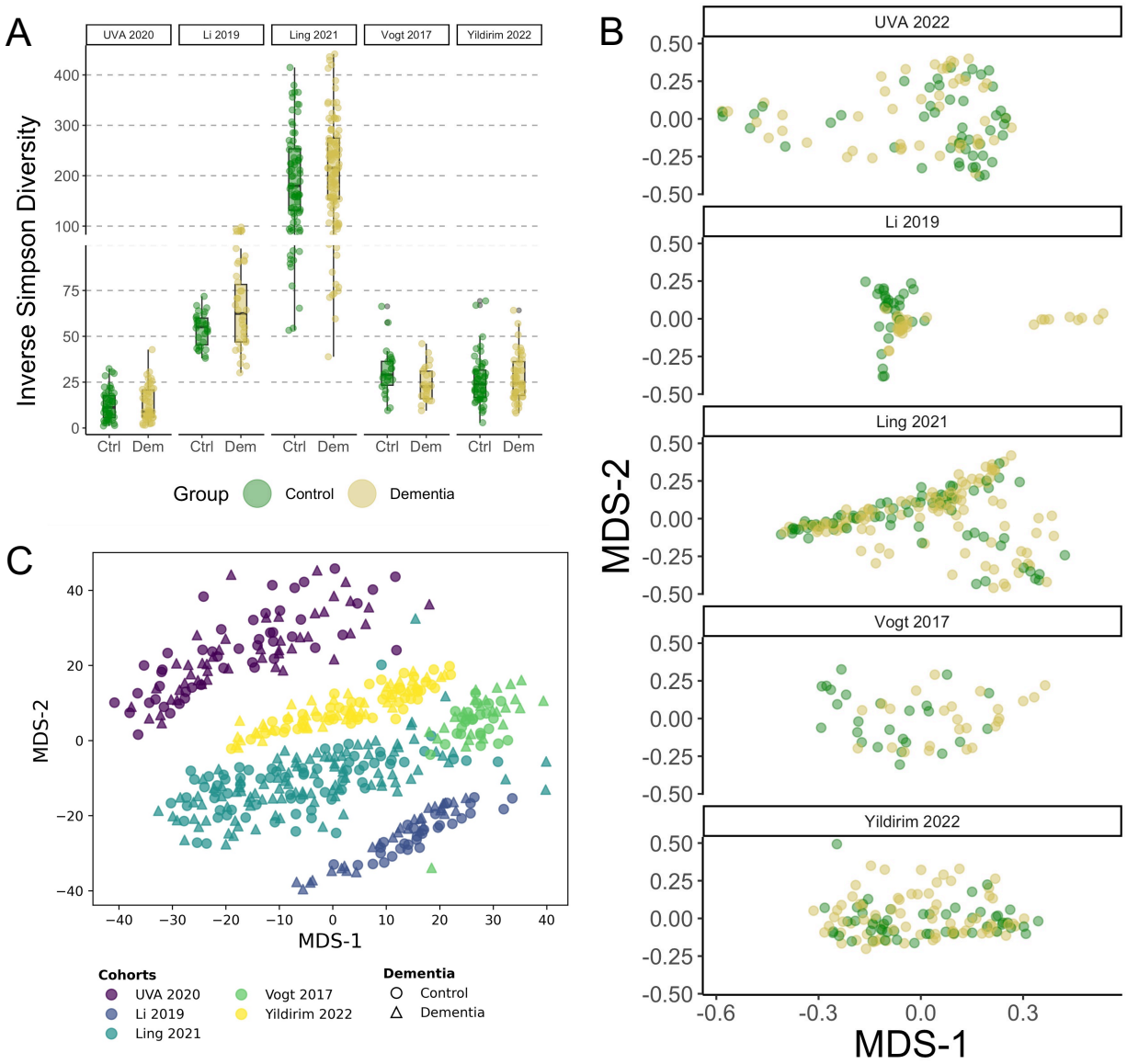


Figure S4.2. 16S rRNA sequencing data does not have a consistent or robust signal across cohorts. (A) Alpha-diversity measured by inverse Simpson for all cohorts. There is no significant difference in diversity between the dementia and control groups with the exception of Li 2019 which had a p-value of 0.015. (B) MDS with Bray-Curtis distance of beta-diversity for all cohorts. (C) MDS of 16S-inferred pathway abundance using PiCRUSt2.

Chapter 5: Discussion and Future Directions

Network analysis of toxin production in *C. difficile*

Summary

In chapter 2, we sought to systematically define the metabolic drivers of toxin production in *C. difficile* using two strains (630 and R20291) with publicly available transcriptomic data from a total of 16 growth conditions. Arginine and isoleucine play particularly conflicting roles in toxin production in *C. difficile*. For example, arginine has been shown to both increase and decrease toxin production depending on the experimental study[1,2]. Our *in-silico* analysis supports a model of increased arginine and ornithine metabolism in a low toxin state. Additionally, we identified a possible up-stream regulatory relationship of fatty acids on arginine which may be an uncontrolled variable that explains experimental differences in previous literature. Isoleucine has a similarly confusing role on toxin production[3,4]. The modified Metabolic Transformation Algorithm (mMTA) analysis suggests that *C. difficile* may preferentially metabolize isoleucine as an energy source via Stickland fermentation during exponential phase, resulting in decreased bioavailable isoleucine in stationary phase. Decreased isoleucine in the stationary phase would then result in CodY deactivation and thereby increased toxin production. In addition to these specific amino acid interactions with toxin production, the mMTA analysis also identified 20 reactions that when knocked-out switched *C. difficile* from a high-toxin metabolic state to a low-toxin state. In this aim we also developed a novel application for the MTA, resulting in our modified algorithm, mMTA. By applying a systems analysis approach to this aim, we were able to identify mechanistic hypotheses that resolve discrepancies and gaps in the experimental literature on toxin production in *C. difficile*.

Next steps

Direct *in vitro* toxin measurement was a significant limitation in this aim. We measured toxin phenotypes using qRT-PCR, cytotoxicity assays, and ELISA assays. We also consulted and collaborated with labs with expertise in *C. difficile* toxin quantification. However, despite extensive efforts, neither we or our collaborators were able to reproducibly or reliably measure toxin phenotypes using these methods. Therefore, we propose further development of toxin quantification protocols at the gene and protein level. At the gene level, qRT-PCR provides a focused snapshot of relative and absolute gene expression. In our initial experiments, we iterated though multiple housekeeping genes from literature, prior work within our own lab, and consultation with Dr. Tamayo's lab at UNC; none of these provided consistent, interpretable results. To address this problem, we will use the geNorm algorithm to identify the most stable reference genes in *C. difficile* and then test this panel of housekeeping genes with *C. difficile* grown in experimental conditions (e.g. BHIS, defined minimal media). At the protein level, primary monoclonal antibodies to TcdA and TcdB have been developed for use in Western blots. We will validate these antibodies to qualitatively assess toxin protein presence or absence in *C. difficile* in our experimental conditions of interest. By improving the reproducibility of experimental tools and protocols for toxin measurement, we will expand the ability to validate the relationships between a multitude of variables and toxin production.

Our *in silico* mMTA analysis identified several reactions that could function as metabolic switches for transforming from a high protein to low protein state. We propose knocking-out these reactions *in vitro* to see if toxin production is decreased. First, we will filter the top reaction hits from mMTA by the gene-protein-reaction rules. Specifically, we will search for reactions whose activity is controlled through either a single gene or a gene combination with a Boolean "AND" expression. Historically, generating gene knockouts in *C. difficile* has been difficult due to low transformation efficiency and restriction-modification systems. However, recently, CRISPR-Cas9 systems have been successfully used to create gene knockouts in *C. difficile*. Therefore, after identifying the target genes from GPR rules, we will use this list to create gene knockouts using CRISPR-Cas9. After creating and validating our knockouts, we will compare toxin production between the wild-type and mutants in both rich and minimal media conditions.

Translational application

Metabolic drivers of toxin production in *C. difficile* can be leveraged to treat CDI. In CDI, toxin damage to the colon is a severe symptom of disease, both inducing significant pain for patients and establishing nutritional niches for *C. difficile* to perpetuate in the colon. Blocking toxin production is an attractive therapeutic strategy because 1) it is specific to *C. difficile* thereby preventing further microbiome loss through broad spectrum

antibiotics and 2) it alleviates the physical damage and toxic effects on the colon thereby improving patient outcomes. By conducting a systematic network analysis of toxin states, we have provided a framework for understanding opposing reports on the connections between metabolism and toxin production. Additionally, mMTA provides a starting point for exploring candidate metabolic genes and reactions whose knockout could attenuate the toxic impacts of *C. difficile.* Together, these findings establish a necessary foundation for continued research.

Nutritional, commensal, and genetic drivers of metabolic phenotypes in *C. difficile*

Summary

In chapter 3, we investigated the metabolic interactions between *C. difficile*, the gut microbiome, and diet using genome-scale metabolic models of 105 *C. difficile* strains and 8 gut commensals. The composition and function of the gut microbiome are responsive to a variety of variables such as diet, exercise, genetics, and antibiotics. For example, broad-spectrum antibiotics can clear the normal microbiome creating an open niche for opportunistic clonal expansion of *C. difficile*. In cases of recurrent CDI, a fecal microbiota transplant (FMT) can resolve disease by restoring a healthy microbiome. However, FMTs are not without risk, therefore extensive research efforts are being made in designing synthetic microbial communities that could replace FMTs. Expanding our knowledge of the underlying microbial interactions and variables in CDI resolution is a critical step for engineering successful communities. In our *in-silico* analysis, we found that strain differences in *C. difficile* give rise to different metabolic phenotypes under the same diet conditions. Specifically, a subset of strains had increased flux through energy-generating reactions across all diet conditions. However, in our Metabolic CrossTalk (MetCT) analysis, strain differences play a negligible role in metabolic interactions between commensals and *C. difficile*. In these instances, competition and mutualism appear to dominate interactions more than diet or strain differences in *C. difficile*. In completing this aim, we created a new collection of metabolic models for 105 *C. difficile* strains and applied novel ecological interaction simulation methods.

Next steps

The first portion of this aim found significant differences in energy metabolism across strains in the same diet condition. To experimentally validate these metabolic differences, we will perform metabolomics on *C. difficile* strains from different clusters. We have selected two strains (Cd10 and Cd4) which are from clusters 1 and 2, respectively. These strains have been grown *in vitro* in defined diet media that replicates the HPLC and LPHC media used *in silico*. The spent media from these cultures has been collected and filter-sterilized in preparation for metabolomics analysis at the Duke Metabolomics Core Facility. The samples will be tested with a panel of 500 metabolites. Based on modeling simulations, we expect to see differences in the relative abundance of amino acids and their Stickland fermentation by-products.

From the MetCT simulations and validating *in vitro* experiments, the commensal pairs were more important in predicting *C. difficile* growth than diet or *C. difficile* strain differences. We conducted a limited number of validating experiments for these results, however multiple facets of validation remain. Therefore, we propose two sets of validating experiments. In the first set of validating experiments, we will test the robustness of potential therapeutic microbial communities under different diet conditions. Previous work in the lab has identified 4-member community that can resolve CDI in mice. We will grow this community under four diet conditions: HPLC, LPHC, high fat/high protein (HFHP), and high fat/high carbohydrate (HFHC). We will assess community composition by measuring the relative species abundance with 16S rRNA sequencing and we will assess the community function by measuring *C. difficile* growth in the community spent media. In the second set of validating experiments, we will test whether ecological interactions between *C. difficile* and commensals are robust to *C. difficile* strain differences. We will use a panel of 20-30 clinical *C. difficile* isolates collected from the UVA hospital. Using the procedure above, we will collect community spent media and then measure growth of the *C. difficile* strains in the spent media. These two complementary experiments will create a valuable dataset for formalizing the ecological relationships between diets, *C. difficile* strains, and commensals.

In this project, we developed two defined diets, HPLC and LPHC, with *in silico* and *in vitro* formulas for both. However, these two diets represent a minute fraction of possible diets. To expand on nutritional research, we propose developing new, high-throughput experimental assays for common diets across the world. In contrast to a Biolog phenotype microarray plate which measures the effect of a single metabolite, this diet assay will contain a panel of defined, complex media based on the three primary macronutrients: proteins, carbohydrates,

and fats. Using a high, medium, and low level for each macronutrient will yield 27 possible formulations. These media formulas could be further refined by altering mineral or other micronutrient amounts to reflect specific diets commonly associated with malnutrition or other metabolic disorders. Furthermore, each *in vitro* diet formula will be paired with an *in-silico* version to facilitate complementary use of computational and experimental approaches. The development of this experimental resource will advance nutritional research not only in relationship to the gut microbiome and *C. difficile*, but also in metabolic disorders such as diabetes, obesity, or malnutrition.

Finally, this aim highlights an opportunity in the metabolic modeling field to expand microbial community tools for investigating ecological dynamics. Using computational approaches to identify reduced microbial community structures that can carry out a desired metabolic function increases the chances for success *in vitro*. Metabolic modeling can also provide mechanistic insights into the specific metabolic cross-feeding pattern of a community that can be leveraged in developing solutions. Therefore, we propose updating the MetCT algorithm to accommodate larger microbial communities and to improve the ecological functions. These updates would allow investigation into overall community function as well as individual interactions within the community. Furthermore, by enabling the analysis of larger communities, MetCT could be applied to an array of microbiome-associated disorders, such as depression, inflammatory bowel disease, Crohn's disease, dementia, diabetes, and others. Moreover, this tool could also be used to study microbial solutions outside of human health, such as reversing coral bleaching, treating wastewater or algal blooms, degrading plastic, or producing bio-compounds of interest.

Translational application

Microbial communities can be powerful biological solutions. For example, in the case of the current FDA-approved microbial therapeutic for CDI, the synthetic community prevents recurrent CDI by restoring a healthy microbiome[5–7]. Microbial therapeutic treatment of CDI has several advantages over traditional antibiotics: 1) it breaks the cycle antibiotic-induced dysbiosis of the gut microbiome and 2) it does not contribute to the development of antibiotic resistance in pathogens. In general, microbial antibiotic resistance is a growing problem that far outpaces the development of new antibiotics. To combat these challenges, new bacterial infection treatments will need to be effective, specific, and robust. Microbial therapeutics provide one alternate solution; however, they have been limited by the lack of mechanistic research into the community interactions that lead to resolution of infection. By reducing the system to three variables (diet, *C. difficile* strain, and commensal), we were able to distinguish between metabolic phenotypes that stem from genetic differences and metabolic phenotypes that stem from commensal interactions. Isolating the variables that induce these metabolic phenotypes as well as the hierarchical relationships of those variables is critical for designing effective microbial solutions.

Cohort-specific microbiome predictors of dementia

Summary

In chapter 4, we investigated compositional and functional patterns in the GM associated with dementia using 16S rRNA gene sequencing and shotgun metagenomics from a clinical cohort of UVA patients and four publicly available datasets. The association between the GM and dementia has been studied extensively in recent years to discover both its role in disease and predictive GM biomarkers of disease. However, consistent intestinal microbiome patterns across studies have not yet been identified. In our cross-cohort analysis using both new and publicly available data, we found that the GM successfully predicted dementia within most cohorts. However, the predictive ability of the biomarkers was limited to the cohort in which they were identified. Furthermore, the differentially abundant species in dementia and control groups were not consistent across all cohorts. This indicates that uncontrolled variables such as diet, geography, or environmental exposures may hinder the identification and development of a robust GM biomarker for dementia. In completing this project, we also collected, processed, and analyzed rectal swabs from 357 hospitalized UVA patients providing a valuable new dataset for future research.

Next steps

The cross-cohort analysis in this chapter was limited to compositional comparisons and predictions because most publicly available data for the GM in dementia are 16S rRNA sequencing data. However, as discussed previously, taxonomic composition is the most variable feature of the GM. Investigating the GM function across cohorts could identify a more universal metabolic function or ecological pattern underlying the noise of taxonomic

composition. GM function is primarily measured via metagenomics, but there is a paucity of GM metagenomic data in dementia making a statistically powered cross-cohort analysis challenging. Therefore, we propose modeling community metabolic phenotypes across cohorts using metabolic models. First, we will create a reduced microbial community (n=10-20) for each group (dementia and normal controls) in each cohort to a representative set of species. Next, we will create GENREs for each species using Reconstructor and then analyze the ecology and function of each community using MetCT. Comparing the community metabolic functions across cohorts will provide insight into the underlying ecological patterns of dysbiosis in dementia

Translational application
While the nature of the association between the GM and dementia remains elusive, evidence continues to build that there *is* an association. In this chapter we find that the compositional features of the GM are inconsistent predictors of disease due to shifting taxonomic composition of the GM. The negative space of taxonomic noise accentuates the potential of in-depth research into the functional and ecological patterns of the GM in dementia as the path towards understanding the correlative and causative relationships. Additionally, this finding focuses future research questions towards investigating the GM in dementia as a therapeutic target for improving or slowing the progression of symptoms rather than as a tool for diagnosis.

Conclusion

Frequently, discoveries regarding the disparate variables that effect metabolic phenotypes in health and disease arrive in piecemeal: *in vitro* experiments, animal studies, and clinical studies. This research is hard-won and foundational to furthering our understanding of the relationships between bacterial metabolism and disease. However, between these fragmented studies are gaps in knowledge and gaps in context. By applying a systems biology approach to these gaps, we can consider the data and variables from multiple studies at once. This type of integrative analysis can provide context to and resolve discrepancies between different studies. We have used a systems approach to understand the diverse metabolic phenotypes in *C. difficile* associated with toxin production, diet, genetics, and microbiome interactions as well as the characteristics of the GM in dementia. These analyses have generated new hypotheses for metabolic phenotypes in *C. difficile*, identified variables that affect *C. difficile* metabolism in a weighted order, and provided context for interpreting GM characteristics in dementia. These wholistic analyses complement the current body of work and expand our knowledge.

References

1. Lei, X.-H. & Bochner, B. R. Using Phenotype MicroArrays to Determine Culture Conditions That Induce or Repress Toxin Production by Clostridium difficile and Other Microorganisms. *PLoS ONE* **8**, e56545 (2013).
2. Karasawa, T., Maegawa, T., Nojiri, T., Yamakawa, K. & Nakamura, S. Effect of Arginine on Toxin Production by *Clostridium difficile* in Defined Medium. *Microbiology and Immunology* **41**, 581–585 (1997).
3. Dineen, S. S., McBride, S. M. & Sonenshein, A. L. Integration of metabolism and virulence by Clostridium difficile CodY. *J Bacteriol* **192**, 5350–5362 (2010).
4. Ikeda, D. *et al.* Effect of isoleucine on toxin production by Clostridium difficile in a defined medium. *Zentralbl Bakteriol* **287**, 375–386 (1998).
5. Boyle, B. L. & Khanna, S. Fecal microbiota live - jslm (Rebyota™/RBL) for management of recurrent Clostridioides difficile infection. *Future Microbiol* 1–9 (2024) doi:10.1080/17460913.2024.2364583.
6. Berry, P. & Khanna, S. Fecal microbiota spores, live-brpk (VOWST™/VOS) for prevention of recurrent Clostridioides difficile infection. *Future Microbiol* 1–10 (2024) doi:10.1080/17460913.2024.2403892.
7. Louie, T. *et al.* VE303, a Defined Bacterial Consortium, for Prevention of Recurrent *Clostridioides difficile* Infection: A Randomized Clinical Trial. *JAMA* **329**, 1356 (2023).

Acknowledgements

Science is a team sport and I have had a phenomenal team, both professionally and personally. Many thanks to my committee, Dr. Jeff Smith, Dr. Charles Farber, Dr. Barb Mann, and Dr. Ani Manichaikul. You have helped me refine my scientific ideas and create a body of work that I am proud to present.

None of this work would have been possible without my advisor, Dr. Jason Papin. Jason, thank you for embodying the type of scientist, professional, and person I strive to be; your mentorship has been formative and invaluable. Thanks to Dr. Glynis Kolling for providing endless experimental advice and guidance; you are the best wet lab buddy. My fellow grad students are not only excellent scientific sounding boards but also wonderful friends: thanks to Joe Ficarrotta, Sami Felton, Emma Glass, Connor Moore, and Will Shao. Together, you all have made work a joy.

Thank you to my parents for teaching me how to read and for reading to me: it has been the best gift I have received. I love you both. Thanks to my sisters Natalie, Martha, Sarah, Anna, and Miriam and my “cousin” Rebecca for visiting me, sending me pictures of the kids, and being the quirky, passionate people you are: I learn so much from you all. Thank you for being my first and best friends; I can’t wait for us to be treasures by the sea.

Thank you, Lillian: you made grad school explode with color. You are a kindred spirit and my sister-in-heart. I feel so lucky that we happened to be in the same lab; you make everything more.

Thank you, Dawson and Maggie, for being a slow, sunny afternoon: everything feels brighter and more relaxed when you’re around.

Thank you, Noah and Emily: Saturday night dinner is home and harbor. Our weekly dinners nourish my heart as much as my body.

Thank you, Maira for being a calming presence at home and the best roommate. You are so wise and one of the most empathetic persons I know.

Thank you, Shaylyn, for being a treasured friend and crafting wizard. I appreciate you making me take it out, redo it, cast it on, write it down, and all the other crafting and life things you encourage me to do to make life easier for future Deb.

Dinner with friends has been a gold thread running through my years in Charlottesville: watermelon group therapy with Michelle and Natalia; cheffing with Dawson and Lillian (FYP); dinner party with Lindsey, Katie, Mel, and Lillian; bach brunch with the house; crafting dinners with Shaylyn and Maira; kids night with Joe, Sami, Emma, Connor, Will, and Rusty; summer sips with Glynis, Sami, and Emma; and Saturday night dinners with Noah and Emily. You all have made every failure easier and every success sweeter.